Cognitive Agents in Interaction: a formal approach

Emiliano Lorini

Habilitation à Diriger des Recherches (HDR)

Cognitive Agents in Interaction: a formal approach

Emiliano Lorini

Emiliano.Lorini@irit.fr
www.irit.fr/~Emiliano.Lorini/

Under the supervision of Andreas Herzig

Paul Sabatier University, Toulouse, France

Contents

Chapter 1

Introduction

Agents in the societies can be either human agents or artificial agents. The focus of my present and past research is both on: (i) the present society in which human agents interact with the support of ICT through social networks and media, and (ii) future society with mixed interactions between human agents and artificial systems such as autonomous agents and robots. Indeed, new technologies will come for future society in which such artificial systems will play a major role, so that humans will necessarily interact with them in their daily lifes. This includes autonomous cars and other vehicles, robotic assistants for rehabilitation and for the elderly, robotic companions for learning support.

The general aim of my research is to provide formal models of social interaction between cognitive agents. I have mainly worked in the area of artificial intelligence (AI) with strong interaction with other disciplines such as economics, philosophy and cognitive sciences.

There are two main general observations underlying my research project. The first is that interaction plays a fundamental role in existing information and communication technologies (ICT) and applications (e.g., Facebook, Ebay, peer-to-peer systems) and will become even more fundamental in future ICT. The second is that the cognitive aspect is crucial for the design of intelligent systems that are expected to interact with human agents (e.g., embodied conversational agents, robotic assistants, etc.). In these situations the system must be endowed with a psychologically plausible model of human reasoning and human cognition in order to be able to understand the human agent's needs and to predict her behaviour.

Formal methods have been widely used in AI for modelling intelligent systems as well as different aspects of social interaction between artificial and/or human agents such as robots in a team interacting between them (e.g., Robocup) or virtual agents interacting with humans (e.g., tutoring agents). In my research work I have mainly used logic and game theory as formal tools for building models of interaction between cognitive agents. On the methodological side, I have been interested both in the expressivity aspect and in the computational aspects of formal models of interaction. On the one hand, my aim is to develop formal languages that are sufficiently expressive to represent interesting aspects of social interaction such as the trust and power relationships

between the agents in a group, the epistemic aspect, as well as the emotions involved in social interaction. On the other hand, I am interested in studying the mathematical and computational properties of the formal languages I develop including axiomatizability, completeness, decidability and complexity.

My past research has been organized along the three different axes represented in Figure 1.1:

- Actions,
- Mind
- Collectives

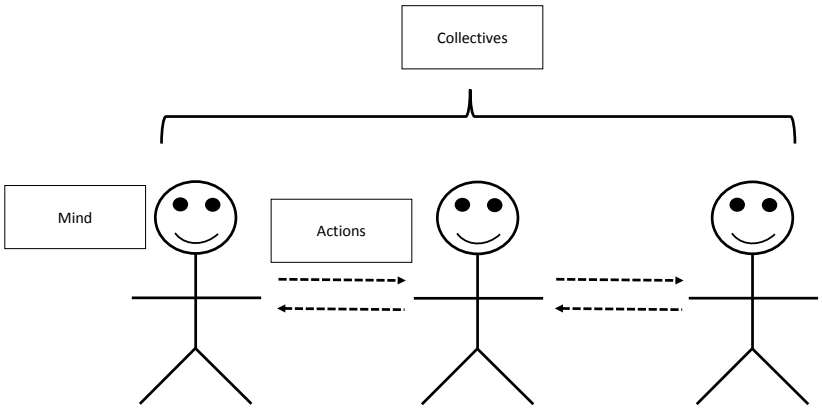


Figure 1.1: Three axes of my research

Let me briefly discuss the content of each axis.

Actions A general theory of social interaction between cognitive agents requires a clear understanding of the concepts of individual and joint action, in particular:

- how the actions of different agents can positively and negatively interfere between each other, in the sense that what an agent can achieve by her acting might depend on what the others decide to do;
- how an agent's responsibility for a given outcome depends either on what the agent causes by her acting (responsibility for action) or on what the actual outcome would have been, if the agent had made a different choice (responsibility for omission);
- how the actions of different agents can be executed in different ways, e.g., in parallel or sequentially;

- how the achievement of an agent's goal may depend on what another agent will decide to do.

This is exactly the issue of the axis "Actions" of my research projet.

In the recent years, I have been working on the constructions of different logics and formal systems based on a game-theoretic semantics elucidating the subtle aspects of the structure of interaction. I have mainly worked in the context of propositional dynamic logic (PDL) [68] and STIT logic (the logic of "Seeing To It That") [16, 78] in connection with some existing logics of strategic reasoning such as Coalition Logic (CL) [108], Coalition Logic of Propositional Control (CL-PC) [157], Alternating-time Temporal Logic (ATL) [3] and Dynamic Logic of Propositional Assignments (DL-PA) [9, 146]. The reason why I am interested in formal systems and logics of action based on a game-theoretic semantics is that game theory offers a powerful and relatively simple framework for describing the structure of interaction both in a static setting (games in normal form) and in a dynamic setting (games in extensive form).

I have also proposed several applications of these logics to the theory of power, to the theory of social influence, to the theory of responsibility, to the theory of institutional action and to the theory of delegation and social dependence.

More details about my research in this axis are given in Chapter 2.

Mind Cognitive agents are, by definition, endowed with a variety of mental attitudes such as beliefs, desires, preferences and intentions that provide input for practical reasoning and decision-making, trigger action execution, and generate emotional responses. The axis of my research "Mind", which is described in detail in Chapter 3, has been devoted to develop formal models of interaction between cognitive agents that:

- clarify the relationship between intention and action and the role of intention in practical reasoning;
- explain how moral attitudes such as standards, ideals and moral values influence decision-making and how preferences are formed;
- elucidate how mental attitudes including beliefs, desires and intentions trigger emotional responses, and how emotions retroactively influence decision-making and mental attitudes by triggering belief revision, desire change and intention reconsideration;
- clarify the role of trust in belief dynamics;
- predict how the agents will behave and what they will choose depending on their preferences, their beliefs about the others' future choices and the choice criteria they adopt (e.g., expected-utility maximization, minimization of the highest possible loss and maximization of the highest possible gain, satisficing criterion).

On the methodological side this part of my research has taken advantage of two modeling tools widely used in artificial intelligence (AI), multi-agent systems (MAS)

and economics: logic and game theory.[1] Specifically, the two main modelling tools used in this part of my research have been: (i) *logics of mental states and their evolution*, including epistemic logic and dynamic epistemic logic [48, 158], BDI (belief, desire, intention) logics [75, 115, 131, 167], logics of preference and preference change [153], theories and logics of belief revision [137, 11], and (ii) *epistemic game theory* [109], i.e., the branch of game theory [107] that studies the epistemic and rationality conditions behind solution concepts (e.g., the epistemic and rationality conditions that are necessary and/or sufficient for players in a game to play a Nash equilibrium).

Collectives Human and artificial societies are formed not only by individuals but also by collective entities such as groups, teams, corporations, organizations, institutions, etc. The part of my research on "Collectives" has been devoted to understand what such collective entities are, how they relate with individual agents, how they are formed and how they evolve over time. I have been working on two central issues in this area, namely collective decision-making and collective attitudes.

As for collective decision-making, I have developed formal theories of pro-social behavior and social preferences, i.e., the kind of behavior that is driven not only by the to goal of maximizing one's welfare, but also by the goal of maximizing the welfare of others. I have been working on the logical formalization of different kinds of pro-social motivations including fairness and reciprocity. Also, I have investigated the concept of team-directed reasoning, i.e., the mode of reasoning that people use when they take themselves to be acting as members of a group or a team. In contrast with existing models of team reasoning, I have proposed a formal model of social ties that well explains how much an agent is willing to act in favour of the collective benefit for the group, depending on how much she is tied to the other members of the group. I have used different methodologies in this part of my research project ranging from game theory and modal logic to empirical validation through experiments.

Collectives attitudes such as joint intention, group belief, group goal, collective acceptance and joint commitment have been widely explored in the area of collective intentionality, the domain of social philosophy that studies how agents function and act at the group level and how institutional facts relate with physical (brute) facts (cf. [96, 147] for a general introduction of the research in this area). My research in the area of collective attitudes has been devoted to develop a number of logical theories that explain: (i) how collective attitudes such as collective acceptance or common belief are formed either through aggregation of individual attitudes or through a process of joint perception, (ii) how institutional facts are grounded on collective attitudes and, in particular, how the existence and modification of institutional facts depend on the collective acceptance of these facts by the agent in the society and on the evolution of this collective acceptance.

More details about my research in this axis are given in Chapter 4.

Method of presentation I have tried to be as much concise as possible and to offer a unified view of my research. Each chapter starts with a general introduction on the research area and on the conceptual background underlying my work in this area,

[1] For an overview of the use of logic and game theory in MAS see [132].

followed by a short summary of my contributions to it. Sometimes I zoom in on the details of my work in order to offer a better understanding and a more concrete idea of the concepts and of the formalization tools I have been using. Chapter 5 discusses some perspectives for my future research.

The list of my publications is given in Annex A at the end of the document and is kept separated from the bibliography. In the text, I use the special tag [**my papers:**] to refer to my papers as listed there. I will only refer to a strict subset of all my publications, as some of them are general overview articles that do necessarily fit the specific topics discussed in each chapter.

In order to help the reader to have a more in-depth view of my work, the present document is complemented by the following selection of articles which are somehow representative of my research in each axis:

- For Axis 1 "Actions":
  - Lorini, E. (2013). Temporal STIT logic and its application to normative reasoning. *Journal of Applied Non-Classical Logics*, 23(4), pp. 372-399.
  - Lorini, E., Sartor, G. (forthcoming). A STIT logic for reasoning about social influence. *Studia Logica*.
- For Axis 2 "Mind":
  - Lorini, E., Schwarzentruber, F. (2011). A logic for reasoning about counterfactual emotions. *Artificial Intelligence*, 175(3-4), pp. 814-847.
  - Lorini, E. (2016). A minimal logic for interactive epistemology. *Synthese*, 193(3), pp. 725-755.
- For Axis 3 "Collectives":
  - Lorini, E., Longin, D., Gaudou, B., Herzig, A. (2009). The logic of acceptance: grounding institutions on agents' attitudes. *Journal of Logic and Computation*, 19(6), pp. 901-940.
  - Attanasi, G., Hopfensitz, A., Lorini, E., Moisan, F. (forthcoming). Social connectedness improves co-ordination on individually costly, efficient outcomes. *European Economic Review*.

The previous articles are included at the end of the document and are also briefly discussed in the next three chapters.

Let me finally remark that when I talk of 'we', this means 'the authors, including me, of the research or the publication discussed here'. This 'we' is a constantly shifting quality and quantity. It can also be just me.

Chapter 2

Actions

Interacting agents perform actions that change the environment. In a multi-agent context agents act together either in parallel or sequentially in a such a way that the consequence of the action of a certain agent depends on what the other agents decide to do. Consequently, in order to describe social interaction in a formal way, it is necessary to have a representation language that allows to describe, at the same time, the causal relations between the actions and their effects, the agents' action repertoires and capabilities as well as the effects of the joint actions of the agents in the system. Actions of agents also occur in time and have a duration. Thus, a comprehensive logical theory of interaction requires a clear understanding of the relationship between actions and time.

One of the logics that has been largely used to represent the concepts of individual and joint actions is Propositional Dynamic Logic (PDL) [68]. PDL has been introduced in theoretical computer science about thirty years ago in order to represent the concept of program and the basic operations on programs (e.g., sequential composition, nondeterministic choice, iteration, test). Its semantics is based on the concept of labelled transition system, that is to say, a graph whose vertices represent possible states of the system and whose edges are labelled with actions of agents. These edges represent transitions between states that are determined by the joint execution of a set of actions by the agents in the system. In PDL actions are abstract entities in the sense that their semantics is just specified in terms of state transitions.

A concrete variant of PDL, called DL-PA (Dynamic Logic of Propositional Assignments) [9, 146], has been proposed in the field of theoretical computer science. Differently from PDL, in DL-PA atomic programs are concrete. Specifically, they are assignments of propositional variables (i.e., the action of a certain agent consists in setting to either $\top$ or $\bot$ the value of a given propositional variable p).

This notion of action, viewed as a propositional assignment, is shared with other formal systems proposed in the recent years in AI and in the area of multi-agent systems such as boolean games and the Coalition Logic of Propositional Control (CL-PC). CL-PC was introduced by [157] as a formal language for reasoning about capabilities of agents and coalitions in multiagent environments. In this logic the notion of capability is modeled by means of the concept of *control*. In particular, it is assumed that each agent i is associated with a specific finite subset Atm_i of the finite set of all atomic

propositions Atm. Atm_i is the set of propositions *controlled* by the agent i. That is, the agent i has the ability to assign a (truth) value to each proposition Atm_i but cannot affect the truth values of the propositions in $Atm \setminus Atm_i$. It is also assumed that control over propositions is exclusive, that is, two agents cannot control the same proposition (i.e., if $i \neq j$ then $Atm_i \cap Atm_j = \emptyset$). Moreover, it is assumed that control over propositions is complete, that is, every proposition is controlled by at least one agent (i.e., for every $p \in Atm$ there exists an agent i such that $p \in Atm_i$).

Boolean games [69, 21] share with CL-PC the idea that an agent's action consists in affecting the truth values of the variables she controls. Boolean games are games in which each player wants to achieve a certain goal represented by a propositional formula: they correspond to the specific subclass of normal form games in which agents have binary preferences (i.e., payoffs can be either 0 or 1). They have been proved to provide a useful and natural abstraction for reasoning about social interaction in multi-agent systems.

An alternative approach to the logical formalization of actions in multi-agent domains is the logic STIT (the logic of "seeing to it that") introduced for the first time in the philosophical area [16, 78] and become popular in computer science in the recent years. This logic is well-suited to represent the concept of causality (whether an agent brings about a certain state of affairs as a result of her current choice) as well as social concepts such as the concepts of responsibility, guilt, delegation and social influence that are of primary importance in modelling social relations between human and artificial agents. Two variants of STIT have been studied in the literature which differ at the syntactic level: an atemporal version and a temporal version. The temporal version of STIT is a combination of temporal operators of temporal logic for expressing temporal properties of facts (e.g., whether a given fact φ will be true in the future) and operators of agency that allow to express the consequences of the choice of an agent or group of agents. The language of atemporal STIT is nothing but the language of temporal STIT restricted to the agency operators which does not include the temporal operators. Differently from PDL, whose mathematical and computational properties are well-known, the mathematical and computational properties of STIT are less far studied and understood. Moreover, STIT has been shown to more complex than PDL. For instance, while the satisfiability problem is EXPTIME-complete for PDL and PSPACE-complete for its concrete variant DL-PA, the satisfiability problem for the variant of STIT with agents and groups of agents (group STIT) is undecidable [74].

Other logical systems have been proposed in the recent years which move from the concept of action to the game-theoretic concept of strategy. Informally speaking, a strategy for a certain agent specifies, for every state of the system characterized by a tree or by a transition system, what the agent is expected to do at this state of the system. The most representative example of strategy logics is Alternating-time Temporal Logic (ATL) [3] which can be seen as the strategic variant of Coalition Logic (CL) by [108] and which allows to formally represent the consequences of the strategy of a certain player or coalition of players.[1]

[1]Differently from STIT, CL can only represent the consequences of the choice of a certain player or coalition players, a choice being the restriction of a strategy to the current state of the system.

My research work in the area of logics of action has been organized along the following axes:

- study of the computational and mathematical properties of STIT both in its atemporal variant and its atemporal variant and systematic comparison of existing formal semantics for STIT [**my papers: A11,A17**];
- study of the formal relationships between different logics of action including the relationships between atemporal STIT, CL-PC and DL-PA [**my papers: A7,B6,C19**];
- development of a general logical theory, called Ockhamist Propositional Dynamic Logic (OPDL) unifying existing logics of action and time including STIT, Coalition Logic (CL), Propositional Dynamic Logic PDL and Full Computation Tree Logic CTL^* [**my papers: A19,A22,C17**];
- development of a variant of ATL with explicit actions in the style of PDL [**my papers: C14**];
- application of STIT, in its atemporal and temporal variants, to the formalization of social concepts including delegation, responsibility, commitment, power and social influence [**my papers: A2,A8,A14,A17,C4,C10,C35,C36,C47,C54,C56**];
- application of DL-PA to the formalization of social and normative concepts such as the concept of segregation, the distinction between practical ability and legal permissibility and the distinction between physical action and institutional action [**my papers: C22,C24,C28,C29**];
- application of boolean games to the formalization of the concept of social power [**my papers: C11**].

In the next three sections, I will focus on the previous first, third and fifth items. In particular, I will present my works on: (i) the axiomatization of a temporal variant of STIT called TSTIT, (ii) Ockhamist Propositional Dynamic Logic (OPDL), and (iii) the application of the logic STIT to the formalization of the concept of social influence.

What unifies TSTIT and OPDL is the their common Ockhamist conception of time, that is, the idea that in a branching-time structure describing all possible evolutions of the system, one can always identify a path or history describing the *actual* evolution of the system.

The reason why I present the application of STIT to social influence is that it naturally follows the presentation of the syntax and semantics of TSTIT.

2.1 Temporal STIT

STIT logic (the logic of *seeing to it that*) by Belnap et al. [16] is one of the most prominent formal accounts of agency. It is the logic of sentences of the form "the agent i sees to it that φ is true". Different semantics for STIT have been proposed in the literature (see, e.g., [16, 30, 166, 125]). The original semantics of STIT by Belnap

et al. [16] is defined in terms of **BT+AC** structures: branching-time structures (**BT**) augmented by agent choice functions (**AC**). A **BT** structure is made of a set of moments and a tree-like ordering over them. An **AC** for an agent i is a function mapping each moment m into a partition of the set of histories passing through that moment, a history h being a maximal set of linearly ordered moments and the equivalence classes of the partition being the possible choices for agent i at moment m.

In [**my papers: A11**], we have proposed a new Kripke-style semantics for STIT. On the conceptual side, the main difference between this Kripke semantics for STIT and Belnap et al.'s **BT+AC** semantics is that the former takes the concept of *world* as a primitive instead of the concept of *moment* and defines: (i) a *moment* as an equivalence class induced by a certain equivalence relation over the set of worlds, (ii) a *history* as a linearly ordered set of worlds induced by a certain partial order over the set of worlds, and (iii) an agent i's set of *choices* at a moment as a partition of that moment. The main advantage of the Kripke semantics for STIT over Belnap et al.'s original semantics in terms of **BT+AC** structures is that the former is a standard multi-relational semantics commonly used in the area of modal logic [20], whereas the latter is non-standard.

It is worth noting that, at the semantic level, temporal STIT can be conceived as logic of action interpreted over infinitely repeated games. This highlights the connection between STIT and game theory.

The Kripke semantics of STIT is illustrated in Figure 2.1, where each moment m_1, m_2 and m_3 consists of a set of worlds represented by points. For example, moment m_1 consists of the set of worlds $\{w_1, w_2, w_3, w_4\}$. Moreover, for every moment there exists a set of histories passing through it, where a history is defined as a linearly ordered set of worlds. For example, the set of histories passing through moment m_1 is $\{h_1, h_2, h_3, h_4\}$. Finally, for every moment, there exists a partition which characterizes the set of available choices of agent 1 in this moment. For example, at moment m_1, agent 1 has two choices available, namely $\{w_1, w_2\}$ and $\{w_3, w_4\}$. Note that an agent's set of choices at a certain moment can also be seen as a partition of the set of histories passing through this moment. For example, we can identify the choices available to agent 1's at m_1 with the two sets of histories $\{h_1, h_2\}$ and $\{h_3, h_4\}$. Following [78], the Kripke semantics for STIT also account for collective choices of groups of agents. Specifically, the choice of a group coincides with the intersection of the choices of the agents in the group. For instance, in Figure 2.2, the individual choices of agents 1 and 2 are, respectively, $\{w_1, w_2, w_5, w_6\}$ and $\{w_1, w_2, w_3, w_4\}$, while the collective choice of group $\{1, 2\}$ is $\{w_1, w_2\}$.

Clearly, for every moment m in a Kripke semantics for STIT, one can identify the set of histories passing through it by considering all histories that contain at least one world in the moment m. Moreover, an agent i's set of choices available at m can also be seen as as a partition of the set of histories passing through m. At first glance, an important difference between Belnap et al.'s semantics and Kripke semantics for STIT seems to be that in the former the truth of a formula is relative to a moment-history pair m/h, also called *index*, whereas in the latter it is relative to a world w. However, this difference is only apparent, because in the Kripke semantics for STIT there is a one-to-one correspondence between worlds and indexes, in the sense that: (i) for every index m/h there exists a unique world w at the intersection between m and h, (ii) and for every world w there exists a unique index m/h such that the intersection between

m and h includes w.

In the Kripke semantics for STIT the concept of world should be understood as a 'time point' and the equivalence class defining a moment as a set of alternative concomitant 'time points'. In this sense, the concept of moment captures a first aspect of indeterminism, as it represents the alternative ways the *present* could be. A second aspect of indeterminism is given by the fact that moments are related in a (tree-like) branching time structure. In this sense, the *future* could evolve in different ways from a given moment. In the Kripke semantics for STIT these two aspects of indeterminism are related, as illustrated in Figure 2.1. Indeed, if two distinct moments m_2 and m_3 are in the future of moment m_1, then there are two distinct worlds in m_1 (w_1 and w_3) such that a successor of the former (w_5) is included in m_2 and a successor of the latter (w_7) is included in m_3.

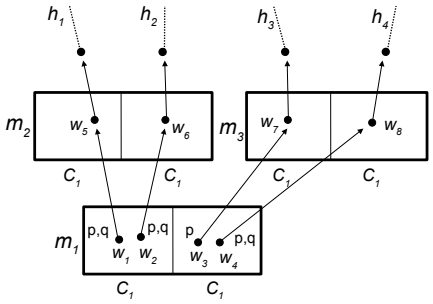


Figure 2.1: Illustration of Kripke semantics of STIT

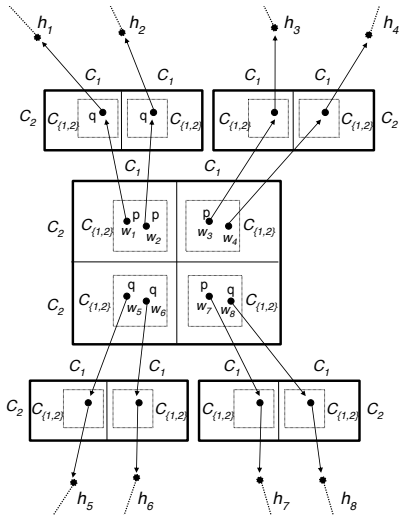


Figure 2.2: Kripke semantics of STIT with groups

In [**my papers: A11**], we have focused on the following language of temporal

STIT (TSTIT), i.e., the variant of STIT with tense operators:

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \psi \mid [i\ \mathsf{stit}]\varphi \mid [Agt\ \mathsf{stit}]\varphi \mid \Box\varphi \mid \mathsf{G}\varphi \mid \mathsf{H}\varphi$$

where p ranges over an infinite set of atomic propositions Atm and i ranges over a finite set of agents Agt.

The previous TSTIT language allows us to talk about time. Specifically, it includes the future tense operator G and the past tense operator H, where $\mathsf{G}\varphi$ and $\mathsf{H}\varphi$, respectively, stand for "φ will always be true in the future" and "φ has always been true in the past". For example, the formula $\mathsf{G}\neg p$ is true at world w_1 in Figure 2.1. Indeed, it is the case that p is false at all future worlds of w_1. Moreover, the formula $\mathsf{H}p$ is true at world w_5 since it is the case that p is true at all past worlds of w_5. In [**my papers: A2,C10**], we have studied a variant of temporal STIT with discrete time which includes the 'next' operator X (where $\mathsf{X}\varphi$ stands for "φ is going to be true in the next world") and the 'yesterday' operator Y (where $\mathsf{Y}\varphi$ stands for "φ was true in the previous world").

The previous TSTIT language also includes the so-called 'historical necessity' operator $\Box$ which allows us to represent those facts that are necessarily true, in the sense of being true at every point of a given moment or, equivalently, at every history passing through a given moment. For example, the formula $\Box p$ is true at world w_1 in Figure 2.1 since p is true at every point of moment m_1 including world w_1. As expected, the 'historical possibility' operator $\Diamond$ is defined to be the dual of $\Box$.

STIT logic provides for different concepts of agency, all characterized by the fact that an agent acts only if she sees to it that a certain state of affairs is the case. In the previous TSTIT language, the so-called Chellas STIT operator $[i\ \mathsf{stit}]$, named after its proponent [35], is taken as a primitive. According to the STIT semantics, an agent i Chellas-sees-to-it that φ, denoted by formula $[i\ \mathsf{stit}]\varphi$, at a certain world w if and only if, for every world v, if w and v belong to the same choice of agent i then φ is true at world v. For example, in Figure 2.1, agent 1 Chellas-sees-to-it that p at world w_1 because p is true both at world w_1 and at at world w_2. The previous TSTIT language also includes the Chellas STIT operator $[Agt\ \mathsf{stit}]$ for the group of all agents, called the 'grand coalition'. For example, suppose $Agt = \{1, 2\}$. Then, in Figure 2.2, group $\{1, 2\}$ Chellas-sees-to-it that p at world w_1 because $\{w_1, w_2\}$ corresponds to the collective choice of group $\{1, 2\}$ at w_1, and p is true both at world w_1 and at at world w_2.

A more sophisticated operator of agency is the deliberative STIT [79] which is defined as follows by means of the Chellas STIT operator and the historical necessity operator $\Box$:

$$[i\ \mathsf{dstit}]\varphi \stackrel{\mathtt{def}}{=} [i\ \mathsf{stit}]\varphi \wedge \neg\Box\varphi$$

In other words, deliberative STIT satisfies the same positive condition as Chellas STIT *plus* a negative condition: an agent i deliberately-sees-to-it that φ, denoted by formula $[i\ \mathsf{dstit}]\varphi$, at a certain world w if and only if: (i) agent i Chellas-sees-to-it that φ at w, that is to say, agent i's current choice at w ensures φ, and (ii) at w agent i could make a choice that does not necessarily ensure φ. Notice that the latter is equivalent to say that there exists a world v such that w and v belong to the same moment and φ is false at v. For example, in Figure 2.1, agent 1 deliberately sees to it that q at world w_1 because q

is true both at world w_1 and at world w_2, while being false at world w_3. In other terms, while the truth of $[i\ \mathsf{stit}]\varphi$ only requires that i's choice ensures that φ, the truth of $[i\ \mathsf{dstit}]\varphi$ also requires that i had the opportunity of making an alternative choice that would not guarantee that φ would be the case. Deliberative STIT, we would argue, captures a fundamental aspect of the concept of action, namely, the idea that for a state of affairs to be the consequence of an action (or for an action to be the cause of a state of affairs), it is not sufficient that the action is a sufficient condition for that state of affairs to hold, it is also required that, without the action, the state of affairs possibly would not hold. In this sense, while $[i\ \mathsf{stit}]\varphi$ at w is consistent with (and is indeed entailed by) the necessity of φ at w, $[i\ \mathsf{dstit}]\varphi$ at w is incompatible with the necessity of φ at w, since it requires that at w also $\neg\varphi$ was an open possibility. Consequently, the deliberative STIT is more appropriate than the Chellas STIT to describe the consequences of an agent's action, as *incompatibility with necessity* is a requirement for any reasonable concept of action.[2]

In [**my papers: A11**], we provided a sound and complete axiomatization for the previous TSTIT language with respect to the Kripke semantics illustrated above. It is summarized in Figure 2.3.

This includes all tautologies of classical propositional calculus **(PC)** as well as modus ponens **(MP)**. Moreover, we have all principles of the normal modal logic S5 for every operator $[i\ \mathsf{stit}]$, for the operator $[Agt\ \mathsf{stit}]$ and for the operator $\Box$, all principles of the normal modal logic KD4 for the future tense operator G and all principles of the normal modal logic K for the past tense operator H. That is, we have Axiom K for each operator: $(\blacksquare\varphi \wedge \blacksquare(\varphi \rightarrow \psi)) \rightarrow \blacksquare\psi$ with $\blacksquare \in \{\Box, \mathsf{G}, \mathsf{H}, [Agt\ \mathsf{stit}]\} \cup \{[i\ \mathsf{stit}] \mid i \in Agt\}$. We have Axiom D for the future tense modality G: $\neg(\mathsf{G}\varphi \wedge \mathsf{G}\neg\varphi)$. We have Axiom 4 for $\Box$, G, $[Agt\ \mathsf{stit}]$ and for every $[i\ \mathsf{stit}]$: $\blacksquare\varphi \rightarrow \blacksquare\blacksquare\varphi$ with $\blacksquare \in \{\Box, [Agt\ \mathsf{stit}], \mathsf{G}\} \cup \{[i\ \mathsf{stit}] \mid i \in Agt\}$. Furthermore, we have Axiom T for $\Box$, $[Agt\ \mathsf{stit}]$ and for every $[i\ \mathsf{stit}]$: $\blacksquare\varphi \rightarrow \varphi$ with $\blacksquare \in \{\Box, [Agt\ \mathsf{stit}]\} \cup \{[i\ \mathsf{stit}] \mid i \in Agt\}$. We have Axiom B for $\Box$, $[Agt\ \mathsf{stit}]$ and for every $[i\ \mathsf{stit}]$: $\varphi \rightarrow \blacksquare\neg\blacksquare\neg\varphi$ with $\blacksquare \in \{\Box, [Agt\ \mathsf{stit}]\} \cup \{[i\ \mathsf{stit}] \mid i \in Agt\}$. Finally we have the rule of necessitation for each modal operator: $\frac{\varphi}{\blacksquare\varphi}$ with $\blacksquare \in \{\Box, [Agt\ \mathsf{stit}], \mathsf{G}, \mathsf{H}\} \cup \{[i\ \mathsf{stit}] \mid i \in Agt\}$.

$(\Box \rightarrow i)$ and **(AIA)** are the two central principles in Xu's axiomatization of the Chellas's STIT operators $[i\ \mathsf{stit}]$ [168]. According to Axiom **$(\Box \rightarrow i)$**, if φ is true regardless of what every agent does, then every agent sees to it that φ. In other words, an agent brings about those facts that are inevitable.[3] According to Axiom **$(i \rightarrow Agt)$**, all agents bring about together what each of them brings about individually.

We have principles for the tense operators and for the relationship between time and action. **(Connected$_{\mathsf{G}}$)** and **(Connected$_{\mathsf{H}}$)** are the basic axioms for the linearity of the future and for the linearity of the past [59]. **(Conv$_{\mathsf{G,H}}$)** and **(Conv$_{\mathsf{H,G}}$)** are the basic interaction axioms between future and past of minimal tense logic according to which

[2]The classical argument against the use of Chellas STIT for modeling action is that, according to Chellas STIT, an agent brings about all tautologies and that it is counterintuitive to say that a tautology is a consequence of an agent's action.

[3]Xu considers a family of axiom schemas **(AIA$_k$)** for independence of agents of the form $(\Diamond[1\ \mathsf{stit}]\varphi_1 \wedge \ldots \wedge \Diamond[k\ \mathsf{stit}]\varphi_k) \rightarrow \Diamond([1\ \mathsf{stit}]\varphi_1 \wedge \ldots \wedge [k\ \mathsf{stit}]\varphi_k)$ that is parameterized by the integer k. As pointed out by [16], **(AIA$_{k+1}$)** implies **(AIA$_k$)**. Therefore, as Agt is finite, in OPDL the family of axiom schemas can be replaced by the single axiom **(AIA)**.

PC	All tautologies of classical propositional calculus
S5(i)	All S5-principles for the operators $[i\ \mathsf{stit}]$
S5($\Box$)	All S5-principles for the operator $\Box$
S5(Agt)	All S5-principles for the operator $[Agt\ \mathsf{stit}]$
KD4(G)	All KD4-principles for the operator G
K(H)	All K-principles for the operator H
($\Box \to i$)	$\Box\varphi \to [i\ \mathsf{stit}]\varphi$
($i \to Agt$)	$([1\ \mathsf{stit}]\varphi_1 \wedge \ldots \wedge [n\ \mathsf{stit}]\varphi_n) \to [Agt\ \mathsf{stit}](\varphi_1 \wedge \ldots \wedge \varphi_n)$
(AIA)	$(\Diamond[1\ \mathsf{stit}]\varphi_1 \wedge \ldots \wedge \Diamond[n\ \mathsf{stit}]\varphi_n) \to \Diamond([1\ \mathsf{stit}]\varphi_1 \wedge \ldots \wedge [n\ \mathsf{stit}]\varphi_n)$
(Conv$_{\mathsf{G,H}}$)	$\varphi \to \mathsf{GP}\varphi$
(Conv$_{\mathsf{H,G}}$)	$\varphi \to \mathsf{HF}\varphi$
(Connected$_{\mathsf{G}}$)	$\mathsf{PF}\varphi \to (\mathsf{P}\varphi \vee \varphi \vee \mathsf{F}\varphi)$
(Connected$_{\mathsf{H}}$)	$\mathsf{FP}\varphi \to (\mathsf{P}\varphi \vee \varphi \vee \mathsf{F}\varphi)$
(NCUH)	$[Agt\ \mathsf{stit}]\mathsf{G}\varphi \to \mathsf{G}\Box\varphi$
(MP)	$\dfrac{\varphi,\ \varphi \to \psi}{\psi}$
(IRR)	$\dfrac{(\Box\neg p \wedge \Box(\mathsf{G}p \wedge \mathsf{H}p)) \to \varphi}{\varphi}$, provided p does not occur in φ

Figure 2.3: Axiomatization of TSTIT

"what is, will always have been" and "what is, has always been going to be".

Axiom **(NCUH)** corresponds to so-called property of 'no choice between undivided histories' which is implicit in the Kripke semantics for STIT illustrated above: if in some future world φ will be possible then the actual collective choice of all agents will possibly result in a state in which φ is true.

(IRR) is a variant of the well-known Gabbay's irreflexivity rule that has been widely used in the past for proving completeness results for different kinds of temporal logic in which time is supposed to be irreflexive (see, e.g., [54, 170, 120, 162]). The idea is that the irreflexivity for time, although not definable in terms of an axiom, can be characterized in an alternative sense by means of the rule **(IRR)**. This rule is perhaps more comprehensible if we consider its contrapositive: if p does not occur in φ and φ is TSTIT consistent, then $\Box\neg p \wedge \Box(\mathsf{G}p \wedge \mathsf{H}p) \wedge \varphi$ is TSTIT consistent.

2.2 Ockhamist Propositional Dynamic Logic OPDL

The distinction between the 'Ockhamist' semantics and the 'Peircean' semantics for branching-time temporal logic was proposed by Prior in his seminal work on the logic of time [112] (see also [145]). According to the 'Peircean' view the truth of a temporal formula should be evaluated with respect either to some history or all histories starting in a given state. According to the 'Ockhamist' view, which is assumed in the context of the STIT semantics illustrated in the previous section, a formula should be evaluated with respect to an *actual* course of events of a certain branching-time structure. In particular, according to the 'Ockhamist' view, the truth of a temporal formula should be evaluated with respect to a particular *actual* history starting in a given state or moment.

While the branching-time temporal logic CTL* [119] is compatible with the Ockhamist conception of time, the semantics for PDL in terms of labelled transition systems is closer to the Peircean view than to the Ockhamist view since it does not consider a notion of actual history or actual path in a transition system.

In [**my papers: C17**], we introduced a new logic, called Ockhamist Propositional Dynamic Logic (OPDL), as a variant of PDL based on the Ockhamist view of time. Specifically, OPDL is a variant of PDL in which the truth of a formula is evaluated with respect to a given actual history. In previous work [**my papers: A19,A22**], we have studied less expressive fragments of OPDL that only deal with one-step actions and the next time point.

The Ockhamist semantics of OPDL is illustrated in Figure 2.4. As in the STIT semantics defined above, we have moments defined as equivalence classes over worlds. For instance, the initial moment is defined by the set of worlds $\{w_1, w_2, w_3, w_4\}$. This moment is followed by the two moments $\{w_5, w_6\}$ and $\{w_7, w_8\}$, and so on. There is also a temporal ordering over the worlds. Specifically, for every world we can identify a successor. For example, the successor of world w_1 is world w_5, the successor of world w_5 is world w_9, and so on. Similarly to the STIT semantics illustrated above, a sequence of worlds ordered by this successor relation identifies a history. In the OPDL semantics, transitions between worlds are labelled with non-empty sets of atomic actions. For example, in Figure 2.4, the actions a and c are responsible for the transition from the state w_1 to the state w_5, while the actions b and c are responsible for the

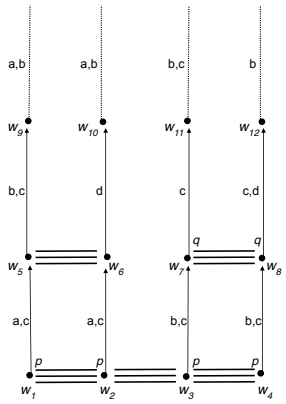


Figure 2.4: An OPDL model

transition from the state w_3 to the state w_7.

At the syntactic level, the language of OPDL includes the basic boolean constructions and the dynamic operator $[\![\alpha]\!]$, whose dual is $\langle\!\langle\alpha\rangle\!\rangle$, where α ranges over a set of programs including: (i) the standard PDL programs: atomic programs ($a, b, \ldots$), sequential composition (;), non-deterministic choice ($\cup$), iteration (*) and test ($?$), and (ii) the special 'history quantifier' program $\equiv$. The special program $\equiv$ allows us to move from a world to an alternative world included in the same moment or, alternatively, from a history to an alternative history passing through the same moment. For instance, at world w_1, the formula $[\![\equiv]\!]p$ is true since p is true at every world in the moment $\{w_1, w_2, w_3, w_4\}$. The special program $\equiv$ can be combined with the other program constructions to express, for example, what is true along those histories that are labelled with certain atomic programs. For instance, formula $[\![\equiv; b]\!]q$ is true at world w_1 since q is true at the successor world along all histories whose initial transition is labelled with the atomic program b (i.e., worlds w_7 and w_8). It is worth mentioning that in OPDL atomic programs are assumed to be linear. This means that if the atomic program a is going to be executed and φ will be true after a's execution then, for every atomic program b, if b is going to be executed then φ will be true after b's execution. This is expressed by the following OPDL validity:

$$\langle\!\langle a\rangle\!\rangle\varphi \rightarrow [\![b]\!]\varphi$$

The interesting aspect of OPDL is that it unifies different logics of action and branching time including propositional dynamic logic PDL, Full Computation Tree Logic CTL*, Coalition Logic CL and the discrete-time and bounded-choice variant of temporal STIT.[4] Indeed, we have provided polynomial embeddings of these logics into

[4]With bounded-choice variant of STIT, I mean the variant of STIT in which the number of choices available to an agent at a given moment is bounded by some integer k.

OPDL. In [**my papers: C17**], we have also proved that the satisfiability problem of OPDL is decidable.

2.3 An example of application to social reality: social influence

In [**my papers: A2,C4,C10**] we have proposed a STIT logic analysis of the concept of social influence that starts from a general view about the way rational agents make choices. Specifically, our assumption is that an agent might have several choices or alternatives *available* defining her *choice set* at a given moment, and that what the agent does is determined by her *actual* choice, which is in turn determined by the agent's *choice context* including her preferences and beliefs and the composition of her choice set. Our analysis of social influence expands this view by assuming that the agent's choice context determining the agent's actual choice might be determined by external causes. Specifically, the external conditions in which an agent finds herself or the other agents with whom the agent interacts may provide an input to the agent's decision-making process in such a way that a determinate action should follow. Specifically, influence consists in *determining* the voluntary action of an agent by modifying her *choice context*, so that a different choice becomes preferable to the influencee on comparison to what would be her preferred option without this modification. This may happen, for instance:

- by expanding the available choices (influence via choice set expansion), or
- by restricting the available choices (influence via choice set restriction) or
- by changing the payoffs associated to such choices, as when rewards or punishments are established (influence via payoff change).

The reason why we have decided to formalize influence in STIT is that this logic is capable of: (i) capturing the temporal aspect of influence, namely the fact that the influencer's choice must precede the influencee's action,[5] and (ii) addressing the strategic aspect of influencing relationships through extensive form games.

To illustrate the concept of social influence, let me consider an example about influence via choice set restriction. The example is illustrated in Figure 2.5. It represents a situation where there are three fruits on a table, an apple, a banana and a pear. The actions at issue consist in bringing about that the apple is eaten (ap), the banana is eaten (ba) or the pear is eaten (pe). Let me assume that agent 2 has certain preferences that remain constant along the tree structure. In particular, at all moments agent 2 prefers eating apples to bananas to pears. Let me also assume that 2 is rational, in the minimal sense that she acts in such a way as to achieve the outcome she prefers. Rational choices of agent 2 are depicted in grey. By choosing to eat the apple at w_1, 1 generates a situation where, given her preferences, 2 will necessarily eat the banana, rather than the pear. Indeed, although at moment m_2, 2 has two choices available, namely, the

[5]The term 'influencer' refers to the agent who exerts influence, whereas the term 'influencee' refers to the agent being influenced.

choice of eating the banana and the choice of eating the pear, only the former is rational, in the sense of being compatible with 2's preferences. In this sense, by deciding to eat the apple at w_1 and removing this option from 2's choice set, 1 influences 2 to decide to eat the banana at w_7.

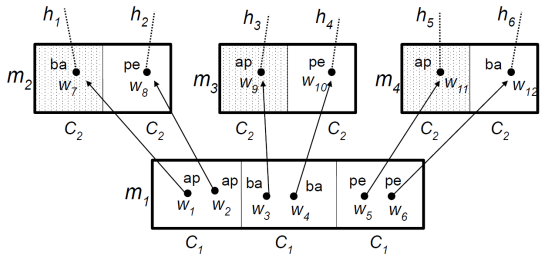


Figure 2.5: Example of influence via choice restriction

As an example of influence via payoff modification, consider the case of agent 1 reliably promising assassin 2 to pay him the reward r if 2 kills 3. The alternative at issue for 2 consists in 2 seeing to it that 3 is killed (k) or doing nothing (so that $\neg k$ holds). Let me assume that 2 prefers $k \wedge r$ to $\neg k \wedge \neg r$ and prefers $\neg k \wedge \neg r$ to $k \wedge \neg r$ (as killing may be unpleasant and involves some risk, so that it would not be preferred without the reward). Under these assumptions, agent 1's choice to reliably promise the reward at moment m_1 leads to a moment m_2 where agent 2 has the choice between producing $k \wedge r$ or $\neg k \wedge \neg r$, rather than to a moment m_3 where agent 2 would have had the choice between $k \wedge \neg r$ or $\neg k \wedge \neg r$ (no compensation been given for the action of killing). Given the preferences of 2, his rational choice at m_2 is to commit the omicide, while at m_3 it would have been not to commit it. Thus we may say that at moment m_1 agent 1, by promising the reward, influences agent 2 into killing.

As an example of influence via choice set expansion, consider a cases were agent 1 gives a gun to agent 2, who desired to kill 3, but was unable to do since he did not have an appropriate weapon. In this case, 1 makes the choice of killing available to 2, who does it according to his preferences. Also in this case, it appears that 1 has influenced 2 into killing.

These examples led us to the following informal definition of social influence:

> An agent i influences another agent j to perform a certain (voluntary) action if and only if, i sees to it that that every rational choice of j will lead j to perform the action.

In order to formalize the previous concept of social influence, we extended STIT logic with special 'rational' STIT operators of the form $[i\ \mathsf{rdstit}]$. The formula $[i\ \mathsf{rdstit}]\varphi$ has to be read "if agent i's current action is the result of a rational choice of i, then i

deliberately sees to it that φ". We adopted a minimal concept of rationality, which was sufficient for our purpose: we assumed that the choices of an agent are ranked according to the agent's preferences, and an agent is rational as long as she implements her preferred choices. The $[i\ \mathsf{rdstit}]$ operator is interpreted relatively to STIT branching time structures, like the ones illustrated in Section 2.1. Specifically, the formula $[i\ \mathsf{rdstit}]\varphi$ is true at a certain world w if and only if, *if* the actual choice to which world w belongs is a rational choice of agent i *then*, at world w agent i deliberately sees to it that φ, in the sense of deliberative STIT discussed in Section 2.1. For example, at the world w_7 in Figure 2.5, the formula $[2\ \mathsf{rdstit}]ba$ is true since the actual choice to which world w_7 belongs is a rational choice of agent 2 *and* at w_7 agent 2 deliberately sees to it that ba is the case.

To capture the idea of social influence, we introduced the following social influence operator based on the concept of deliberative STIT:

$$[i\ \mathsf{sinfl}\ j]\varphi \stackrel{\text{def}}{=} [i\ \mathsf{dstit}]\mathsf{X}[j\ \mathsf{rdstit}]\varphi.$$

In other words, we shall say that an agent i influences another agent j to make φ true, denoted by $[i\ \mathsf{sinfl}\ j]\varphi$, if and only if i deliberately sees to it that if agent i's current choice is rational then i is going to deliberately see to it that φ. The reason why the operator $[i\ \mathsf{dstit}]$ is followed by the temporal operator X is that influence requires that the influencer's choice precedes the influencee's action. On the contrary, we did not require $[j\ \mathsf{rdstit}]$ to be followed by X since in STIT the concept of action is simply captured by the deliberative STIT operator which does not necessarily need to be followed by temporal modalities. In order to illustrate the meaning of the influence operator, let me go back to the example of Figure 2.5. Since agent 2 prefers eating bananas to pears, her only rational choice at moment m_2 is $\{w_7\}$. From this assumption, it follows that formula $[1\ \mathsf{sinfl}\ 2]ba$ is true at world w_1. Indeed, at world w_1 agent 1 deliberately sees to it that, in the next world, if agent 2's choice is rational then 2 deliberately sees to it that ba is the case. Similarly, as assassin 2 prefers killing and getting the reward ($k \wedge r$) to not killing and not getting it ($\neg k \wedge \neg r$), instigator 1, by choosing to provide the reward, induces the assassin to commit the crime. This is a deliberate choice, as had the instigator chosen differently (had she not provided the reward), the rational choice for the assassin would have been a different one, namely, not to kill victim 3.

Note that $[i\ \mathsf{sinfl}\ j]\varphi$ just says that the influencee i would realize φ is she were choosing rationally, but it does not assume that i chooses rationally, and therefore it does not entail that φ would be realized. The notion of *successful* influence also requires that in the next world along the actual history, the influencee chooses rationally, as specified by the following abbreviation:

$$[i\ \mathsf{succsinfl}\ j]\varphi \stackrel{\text{def}}{=} [i\ \mathsf{sinfl}\ j]\varphi \wedge \mathsf{X}\ \mathit{rat}_i.$$

where $[i\ \mathsf{succsinfl}\ j]\varphi$ has to be read "agent i *successfully* influences agent j to make φ true". The expression rat_i means that agent i's current choice is rational. It is an abbreviation, adopted for notational convenience, of $\neg[i\ \mathsf{rdstit}]\bot$, a formula that is satisfied only when i acts rationally in the current word.

This operator of successful influence clearly implies that in the next world along the actual history the influencee performs the action for which she has been influenced.

This is expressed by the following valid formula:

$$[i \ \mathsf{succsinfl} \ j]\varphi \rightarrow \mathsf{X}[j \ \mathsf{dstit}]\varphi.$$

In [**my papers: A2**] we also provided a complete axiomatization for our STIT logic of social influence.

Chapter 3

Mind

Since the seminal work of [37] aimed at implementing Bratman's philosophical theory of intention [26], many formal logics for reasoning about mental attitudes of agents such as beliefs, desires and intentions have been developed. Among them we should mention the logics developed by [75, 84, 102, 103, 115, 131, 136, 159, 167].

The general term used to refer to this family of logics is *agent logics*. A subfamily is the family of BDI logics whose most representative example is the modal logic by [115] whose primitive constituents are the the concepts of belief (B), desire (D) and intention (I) which are expressed by corresponding modal operators. Another well-known agent logic is the so-called KARO framework developed by [102]. KARO is a multi-modal logic framework based on a blend of dynamic logic with epistemic logic, enriched with modal operators for modeling mental attitudes such as beliefs, desires, wishes, goals and intentions. More recently, the KARO framework has been used to model emotions of cognitive agents. In particular, in the KARO framework each emotion type 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 in terms of mental attitudes of agents. The latter correspond to generation rules for emotions. For instance, in [101] generation rules for four basic emotions are given: joy, sadness, anger and fear, depending on the agent's plans.

Generally speaking, agent logics are nothing but formal models of rational agency whose aim is to explain how an agent endowed with mental attitudes makes decisions on the basis of what she believes and of what she wants or prefers. In this sense, the decisions of the agent are determined by both the agent's beliefs (the agent's epistemic states) and the agent's preferences (the agent's motivational states). The output of the agent's decision-making process is either a choice about what to do in the present, also called present-directed intention, or a choice about what to do in the future, also called future-directed intention. As emphasized in the literature in philosophy [26, 100] and AI [29], a future-directed intention is the element of a partial or a complete plan of the agent: an agent may have the intention to perform a sequence of actions later (e.g., the action of going to the train station in two hours followed by the action of

taking the train from Paris to Bruxelles at 10 am) in order to achieve a certain goal (e.g., the goal of being in Bruxelles at the European Commission at 2 pm). A present-directed intention is a direct motivation to perform an action when the time point of the planned action execution is attained. The idea that the behavior of an agent can be explained by attributing mental states to the agent and by having a sophisticated account of the relationship between her epistemic states and her motivational states and of the influence of these on the agent's decision-making process is shared by many disciplines including philosophy of mind [43], cognitive sciences [113], psychology [118] and artificial intelligence [32].

Furthermore, the idea of describing rational agents in terms of their epistemic and motivational attitudes is something that agent logics share with classical decision theory and game theory. In particular, classical decision theory accounts for the criteria and principles (e.g., expected utility maximization) that a rational agent should apply in order to decide what to do on the basis of her beliefs and preferences. Game theory generalizes decision theory to the multi-agent case in which agents' decisions are interdependent and agents' actions might interfere between them so that: (i) the possibility for an agent to achieve her goals may depend on what the other agents decide to do, and (ii) agents form beliefs about the future choices of the other players and, consequently, their current decisions are influenced by what they believe the others will do. More generally, game theory involves a strategic component that is not considered by classical decision theory whose object of analysis is a single agent who makes decisions and acts in an environment she does not share with other agents. The formal representation of mental attitudes of rational agents is even more explicit in epistemic game theory [109], i.e., the branch of game theory that studies the epistemic and rationality assumptions behind solution concepts (e.g., the epistemic and rationality conditions that are necessary and/or sufficient for players in a game to play a Nash equilibrium).

Classical decision theory and game theory provide a quantitative account of individual and strategic decision-making by assuming that agents' beliefs and preferences can be respectively modeled by subjective probabilities and utilities. In particular, while subjective probability captures the extent to which a fact is *believed* by a certain agent, utility captures how much a certain state of affairs is *preferred* by the agent. In other words, subjective probability is the quantitative counterpart of the concept of belief, while utility is the quantitative counterpart of the concept of preference.

Qualitative approaches to individual and strategic decision-making have been proposed in AI [22, 81] to characterize criteria that a rational agent should adopt for making decisions when she cannot build a probability distribution over the set of possible events and her preference over the set of possible outcomes cannot be expressed by a utility function but only by a qualitative ordering over the outcomes. For example, going beyond expected utility maximization, qualitative criteria such as the maxmin principle (choose the action that will minimize potential loss) and the maxmax principle (choose the action that will maximize potential gain) have been studied and axiomatically characterized [23, 24].

In the years of my research activity I have actively contributed to the formal analysis of cognitive and affective phenomena by developing new agent logics and by using tools and techniques from logic and epistemic game theory to formalize a variety of mental attitudes and concepts, and to build theories of interaction between cognitive

agents.

In the next three sections, I will discuss my research on different topics related with the representation of mental attitudes and interaction between cognitive agents: (i) the cognitive processing leading from goal generation to action (Section 3.1), (ii) the role in decision-making of rationality and beliefs about other agents' future actions (Section 3.2), and (iii) the representation of the cognitive structure of emotions and of their influence on behaviour (Section 3.3).

3.1 Mental attitudes

The conceptual background underlying my research in the area of mental attitudes is summarized in Figure 3.1. The cognitive architecture represents the process leading from generation of desires and moral values and formation of beliefs via sensing to action performance.

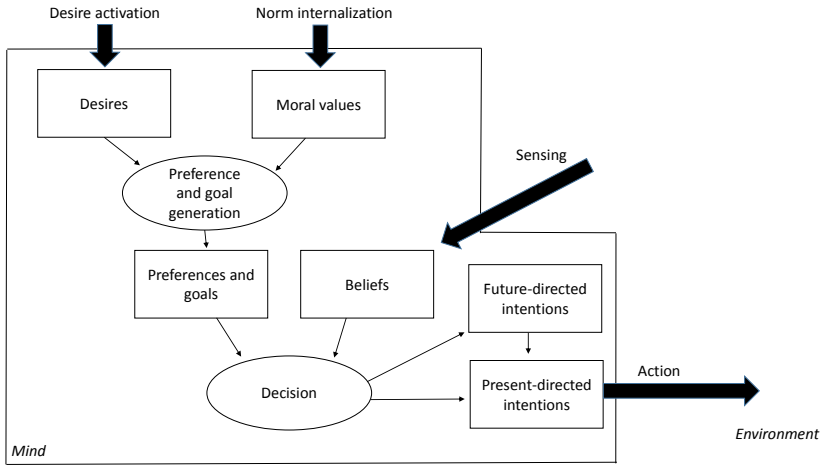


Figure 3.1: Cognitive architecture

The origin of beliefs, desires and moral values An important and general distinction in philosophy of mind is between epistemic attitudes and motivational attitudes. This distinction is in terms of the *direction of fit* of mental attitudes to the world. While epistemic attitudes aim at being true and their being true is their fitting the world, motivational attitudes aim at realization and their realization is the world fitting them [110, 4, 80]. Searle [126] calls "mind-to-world" the first kind of *direction of fit* and "world-to-mind" the second one.

There are different kinds of epistemic and motivational attitudes with different functions and properties. Examples of epistemic attitudes are beliefs, knowledge and opinions, while examples of motivational attitudes are desires, preferences, moral val-

ues and intentions. However, the most primitive and basic forms of epistemic and motivational attitudes are beliefs, desires and moral values.

Beliefs are mental representations aimed at representing how the physical, mental and social worlds are. Indeed, there are beliefs about natural facts and physical events (e.g., I believe that tomorrow will be a sunny day), introspective beliefs (e.g., I believe that I strongly wish that tomorrow will be a sunny day), and beliefs about mental attitudes of other agents (e.g., I believe that you believe that tomorrow will be a sunny day). Following the Humean conception, a desire can be viewed as an agent's attitude consisting in an anticipatory mental representation of a pleasant state of affairs (representational dimension of desires) that motivates the agent to achieve it (motivational dimension of desires). The motivational dimension of an agent's desire is realized through its representational dimension, in the sense that, a desire motivates an agent to achieve it *because* the agent's representation of the desire's content gives her pleasure. For example when an agent desires to be at the Japanese restaurant eating sushi, she imagines herself eating sushi at the Japanese restaurant and this representation gives her pleasure. This pleasant representation motivates her to go to the Japanese restaurant in order to eat sushi.

Moral values, and more generally moral attitudes (ideals, standards, etc.), originate from an agent's capability of discerning what from her point of view is (morally) good from what is (morally) bad. If an agent has a certain ideal φ, then she thinks that the realization of the state of affairs φ ought to be promoted because φ is good in itself. Differently from desires, moral values do not necessarily have a hedonistic and somatic component: their fulfillment does not necessarily give pleasure and their transgression does not necessarily give displeasure 'felt' from the body.

There are different ways to explain the origin of beliefs, desires, moral values. Beliefs are formed either via direct sensing from the external environment (e.g., I believe that there is a fire in the house since I can see it), communication (e.g., I believe that there is a fire in the house since you told me this and I trust what you say) and inference (e.g., I believe that there is a fire in the house since I already believe that smoke comes out from the house and if there is smoke coming out from the house then there is fire). One might argue that belief formation via direct sensing is more primitive than belief formation via communication and that the latter can be reduced to the former. Indeed, in the context of communication, the hearer first *perceives* the speaker's utterance, which is nothing but the performance of a physical action (e.g., uttering a certain sound, performing a certain gesture, emitting a certain light signal, etc.) and forms a belief about what the speaker has uttered. Then, she infers the meaning of the speaker's utterance (i.e., what the speaker wants to express by uttering a certain sound, by performing a certain gesture, by emitting a certain light signal, etc.). Although this is true for communication between humans and between artificial systems situated in the physical environment such as robots, it is not necessarily true for communication in an artificial domain in which there is no precise distinction between an utterance and its meaning. In the latter situation, the speaker may transmit to the hearer is a message (e.g., a propositional formula) with a precise and non-ambiguous meaning.

The concept of trust plays a fundamental role in belief formation via direct sensing and via communication. Indeed, the hearer will not believe what the speaker says unless she believes that the speaker is a reliable source of information, thereby trusting

the speaker's judgment. Similarly, for belief formation via direct sensing, an agent will not believe what she sees unless she believes that her perceptual apparatus works properly, thereby trusting it. The issue whether trust is reducible to other mental attitudes is relevant here. My favorite approach is to conceive *communication-based trust* as a belief about the reliability of a source of information, where "reliable" means that, in the normal conditions, what the source says about a certain issue is true.

As for the origin of desires, the explanation adopted in Figure 3.1 is that they are activated under certain conditions. In the case of human agents, these conditions might be physiological or epistemic. For example, the desire of drinking a glass of water could be activated by the feeling of thirst (physiological condition) and the desire of going outside for a walk might be activated by the belief that it is a sunny day (epistemic condition). In the case of artificial agents, conditions of desire activation should be specified by the system's designer. For example, a robotic assistant who has to take care of an old person could be designed in such a way that, every day at 4 pm, the desire of giving a medicine to the old person is activated in its mind.

As for the origin of moral values, social scientists (e.g., [6]) have defended the idea that there exist innate moral principles in humans such as fairness which are the product of biological evolution. Other moral values, as highlighted in Figure 3.1, have a cultural and social origin, as they are the product of the internalization of some external norm. A possible explanation is based on the hypothesis that moral judgments are true or false only in relation to and with reference to one or another agreement between people forming a group or a community. More precisely, an agent's ideals are simply norms of the group or community to which the agent belongs that have been internalized by the agent. This is the essence of the philosophical doctrine of moral relativism (see, e.g., [20]). For example, suppose that an agent believes that in a certain group or community there exists a norm (e.g., an obligation) prescribing that a given state of affairs should be true. Moreover, assume that the agent identifies herself as a member of this group or community. In this case, the agent will internalize the norm, that is, the external norm will become a moral value of the agent and will affect the agent's decisions. For example, suppose that a certain person is (and identifies herself as) citizen of a given country. As in every civil country, it is prescribed that citizens should pay taxes. Her sense of national identify will lead the person to adopt the obligation by imposing the imperative to pay taxes to herself. When deciding to pay taxes or not, she will decide to do it, not simply in order to avoid being sanctioned and being exposed to punishment, but also because she is motivated by the moral obligation to paying taxes.

From desires and moral values to preferences According to contemporary theories of human motivation both in philosophy and in economics (e.g., [129, 71]), preferences of a rational agent may originate either (i) from somatically-marked motivations such as desires or physiological needs and drives (e.g., the goal of drinking a glass of water originated from the phisiological drive of thirst), or (ii) from moral considerations and values (e.g., the goal of helping a poor person originated from the moral value of taking care of needy people). More generally, there exists desire-dependent preferences and desire-independent ones originated from moral values. This distinction allows us to identify two different kinds of moral dilemmas. The first kind of moral dilemma is

the one which is determined by the logical conflict between two moral values. The paradigmatic example is the situation of a soldier during a war. As a member of the army, the soldier feels obliged to kills his enemies, if this is the only way to defend his country. But, as a catholic, he thinks that human life should be respected. Therefore, he feels morally obliged not to kill other people. The other kind of moral dilemma is the one which is determined by the logical conflict between desires and moral values. The paradigmatic example is that of Adam and Eve in the garden of Eden. They are tempted by the desire to eat the forbidden fruit and, at the same time, they have a moral obligation not to do it.

According to the cognitive architecture represented in Figure 3.1, desires and moral attitudes of an agent are two different parameters affecting the agent's preferences. This allows us to draw the distinction between *hedonistic* agents and *moral* agents. A purely hedonistic agent is an agent who acts in order to maximize the satisfaction of her own desires, while a purely moral agent is an agent who acts in order to maximize the fulfillment of her own moral values. In other words, if an agent is purely hedonistic, the utility of an action for her coincides with the personal good the agent will obtain by performing this action, where the agent's personal good coincides with the satisfaction of the agent's own desires. If an agent is purely moral, the utility of an action for her coincides with the moral good the agent will promote by performing this action, where the agent's promotion of the moral good coincides with the accomplishment of her own moral values. Utility is just the quantitative counterpart of the concept of preference, that is, the more an agent prefers something, the higher its utility. Of course, purely hedonistic agents and purely moral agents are just extremes cases. An agent is more or less moral depending on whether the utility of a given option for her is more or less affected by her moral values. More precisely, the higher is the influence of the agent's moral values in evaluating the utility of a given decision option, the more moral the agent is. The extent to which an agent's utility is affected by her moral values can be called *degree of moral sensitivity*.[1]

Goals The reason why, in Figure 3.1, preferences and goals are included in the same box is that I conceive goals as special kinds of preferences. In particular, I conceive goal as a special kind of *preference* that satisfies two properties. An agent has φ as a goal (or wants to achieve φ) if and only if: (i) the agent prefers φ to be true to φ to be false, and (ii) the agent considers φ a possible state of affairs. The second property is called *realism* of goals by philosophers (cf.[26, 42, 99]). It is based on the idea that an agent cannot reasonably pursue a goal unless she thinks that she can *possibly* achieve it, i.e., there exists at least one possible evolution of the world (a history) that the agent considers possible along which φ will eventually become true. The first property is about the motivational aspect of goals. For φ to be a goal, the agent should not be indifferent between φ and $\neg\varphi$ in the sense that, for every situation she envisages, she prefers φ to be true in this situation than φ to be false, that is, *all other things being equal*, the agent prefers φ to be true than φ to be false. This property also defines Von

[1]This degree can be conceived as a personality trait. In the case of human agents, it is either culturally acquired or genetically determined. In the case of artificial agents, it is configured by the system designer.

Wright's concept of "preference of φ over $\neg\varphi$" [163].[2]

However not all goals have the same status. Certain goals have a motivating force while others do not have it. Indeed, the fact that the agent prefers φ to be true to φ to be false does not necessarily imply that the agent is motivated to achieve a state in which φ is true and that she decides to perform a certain action *in order to* achieve it. For φ to be a motivating goal, for every possible situation that the agent envisages in which φ is true and for every possible situation that the agent envisages in which φ is false, the agent has to prefer the former to the latter. In other words, there is no way for the agent to be satisfied without achieving φ.

An example better clarifies this point. Suppose Mary wants to buy a reflex camera Nikon and, at the same time, she would like to spend less than 300 euros. In other words, Mary has two goals in her mind:

- G1: the goal of buying a reflex camera Nikon, and
- G2: the goal of spending less than 300 euros.

She goes to the shop and it turns out that all reflex cameras Nikon cost more than 300 euros. This implies that Mary believes that she cannot achieve the two goals at the same time, as she envisages four situations in her mind but only three are considered possible by her: the situation in which only the goal G1 is achieved, the situation in which only the goal G2 is achieved and the situation in which no goal is achieved. The situation in which both goals are achieved is considered impossible by Mary. This is not inconsistent with the previous definition of goal since Mary still believes that it is possible to achieve each goal separately from the other. Figure 3.2 clearly illustrates this: the full rectangle includes all worlds that Mary envisages, so-called *information set*, while the dotted rectangle includes all worlds that Mary considers actually possible, so-called *belief set*.[3]

More details about the distinction between information set and belief set will be given in Section 3.2.

Mary decides to save her money and not to buy anything since the goal G2 is a motivating one, while the goal G1 is not. Indeed, every situation that Mary envisages in which the goal G2 is satisfied (w_3 and w_4) is preferred to every situation that Mary envisages in which the goal G2 is not satisfied (w_1 and w_2). Thus, on the basis of what she believes, Mary concludes that she can only achieve her goal G2 by saving her money and by not buying anything in the shop.

From preferences and beliefs to actions As the cognitive architecture in Figure 3.1 highlights, beliefs and preferences are those mental attitudes which determine the agent's choices and are responsible for the formation of new intentions about present actions (present-directed intentions) and future actions (future-directed intentions). In

[2]Von Wright presents a more general concept of "preference of φ over ψ" which has been recently formalized in a modal logic setting by [152].

[3]Mary's information set includes all worlds that, according to Mary, are compatible with the laws of nature. For instance, Mary can perfectly envisage a world in which she is the president of French republic even though she considers this actually impossible.

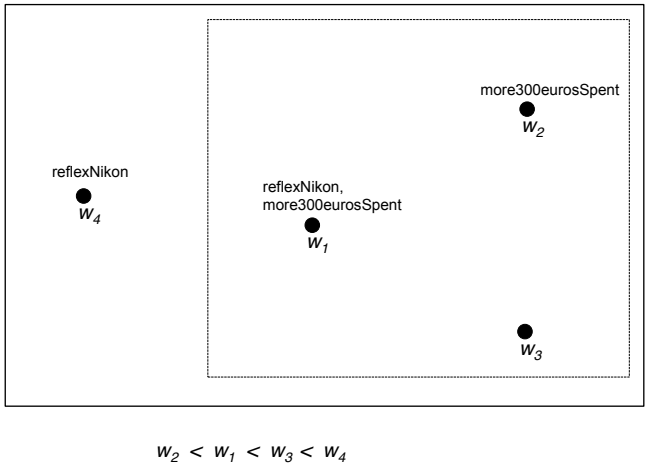


Figure 3.2: Example for goals

particular, decision is determined by beliefs, preferences and a general rationality criterion stating what an agent should do on the basis of what she believes and what she prefers. As emphasized above, different kinds of rationality criteria have been studied in the areas decision and game theory ranging from expected utility maximization, maxmin and maxmax to satisficing [135]. Once the choice has been made by the agent and the corresponding intention has been formed, the action is performed right afterwards or later. Specifically, an agent forms the intention to perform a certain action at a given point in time and, once the time of the planned action execution is attained, the agent performs the action unless before attaining it, she has reconsidered her prior intention.

My work on the logical formalization of mental attitudes Since my PhD thesis, I have been interested in the logical analysis of the cognitive processing leading from the formation of beliefs, desires and moral values to action execution via formation of preferences and generation of intentions. Specifically, we have developed a logical theory of the connection between intention and action via the concept of attempt [**my papers: A25,C53,C57**]. The theory is based on a combination of a modal logic of mental attitudes, propositional dynamic logic (PDL) and propositional linear temporal logic (PLTL). The concept of attempt corresponds to what in philosophy of action is called 'volition', namely the mental action or process of 'starting the execution of a given action'.

We have also studied the problem of intention change by developing a logic of which supports reasoning about intention formation and intention reconsideration operations [**my papers: C33,C37**]. The logic, which exploits methods and techniques from dynamic epistemic logic (DEL) [158], allows to describe the consequences of such operations on an agent's mental setting.

We have developed a logical theory of the relationship between desires, moral values and preferences which clearly distinguishes desire-dependent preferences and desire-independent ones originating from an agent's moral values [**my papers: A6**]. The theory also focuses on the connection between beliefs, preferences, goals and choices.

More recently, we have focused on the formalization of concept of desire in the framework of possibility theory and modal logic [**my papers: C7,C12,C16**]. The aim of this work is to study both the static properties and the dynamic properties of desires including: (i) how the strength of the conjunction of two desires should be computed on the basis of the strengths of the individual desires; (ii) the basic postulates characterizing desire change in relation with the basic postulates characterizing belief change [2].

We have extensively worked on the topic of trust by proposing a logic of trust based on Castelfranchi & Falcone's socio-cognitive theory of trust [33] and providing a sound and complete axiomatization for this logic [**my papers: A20,A21,B7,B14,B18,C38,C45**]. The logic adopts Castelfranchi & Falcone's reduction of trust to the more primitive concepts of belief, goal, capability and opportunity. It offers a formal and more refined version of Castelfranchi & Falcone's definition of trust by distinguishing two general types of trust: occurrent trust and dispositional trust. The former is the one studied by Castelfranchi and Falcone, and it is the trust in the occurrence of an action α of the trustee *here and now* with respect to a current goal of the truster. The latter is the truster's belief that she will possibly have a certain goal φ in the future and, whenever she will have such a goal and certain conditions will obtain, the trustee will perform α and thereby will ensure φ. In this sense, the truster is disposed to trust the trustee. An example of this kind of trust is dispositional trust in the context of trade: the truster 1 believes it to be possible that in the future she will have the goal to receive a certain product from agent 2, and believes that whenever she will have this goal and will pay the product to 2, then 2 will send the product to 1 so that she will receive the product. In [**my papers: C43,C49,C50**] we have applied the logic to security and to the theory of communication: (i) by studying the relationship between the concept trust and security properties such as privacy and integrity, and (ii) by formalizing the concepts of trust in information sources and trust in communication systems.

More recently, we have been interested in the concept of trust-based belief change, that is, belief change that depends on the degree of trust the receiver has in the source of information [**my papers: C9**]. We have proposed a logic which supports reasoning about different kinds of trust-based belief change policies including an additive policy, that cumulates information received by different information sources, and a compensatory policy that, in case different sources provide conflicting information, balances them depending on how much they are trustworthy. At the technical level, the logic consists in extending the static modal logic of belief and trust by Liau [89] in three different directions: (i) a generalization of Liau's approach to graded trust, (ii) its extension by modal operators of knowledge and by modal operators of graded belief based on Spohn's theory of uncertainty [137], and (iii) by a family of dynamic operators in the style of dynamic epistemic logic (DEL). The latter allows for the representation of the consequences of a trust-based belief change operation while the second enables to handle iterated belief change. We have provided a sound and complete axiomatization

for the variant of the logic that implements the additive policy and compensatory policy for trust-based belief change.

3.2 Rationality and decision

The fundamental concept of game theory is the concept of solution which is, at the same time, a prescriptive notion, in the sense that it prescribes how rational agents in a given interaction *should* play, and a predictive one, in the sense that it allows us to predict how the agents *will* play. There exist many different solution concepts both for games in normal form and for games in extensive form (e.g., Nash Equilibrium, iterated deletion of strongly dominated strategies, iterated deletion of weakly dominated strategies, correlated equilibrium, backward induction, forward induction, etc.) and new ones have been proposed in the recent years (see, e.g., [66]). A major issue we face when we want to use some solution concept in order either to predict human behavior or to build some practical applications (e.g., for computer security or for multi-agent systems) is to evaluate its significance. Some of the questions that arise in these situations are, for instance: given certain assumptions about the agents such as the assumption that they are rational (e.g., utility maximizers), under which conditions will the agents converge to equilibrium? Are these conditions realistic? Are they too strong for the domain of application under consideration? There is a branch of game theory, called epistemic game theory, which can help to answer these questions (cf. [109] for a general introduction to the research in this area). Indeed, the aim of epistemic game theory is to provide an analysis of the necessary and/or sufficient epistemic conditions of the different solution concepts, that is, the assumptions about the epistemic states of the players that are necessary and/or sufficient to ensure that they will play according to the prescription of the solution concept. Typical epistemic conditions which have been considered are, for example, the assumption that players have common belief (or common knowledge) about the rationality of every player,[4] the assumption that every player knows the choices of the others,[5]

Epistemic game theory shares concepts and methods with what Aumann calls interactive epistemology [5]. The latter is the research area in logic and philosophy which deals with formal models of knowledge and belief when there is more than one rational agent or "player" in the context of interaction having not only knowledge and beliefs about substantive matters, but also knowledge and beliefs about the others' knowledge and beliefs. The concept of rationality corresponds either to the optimality criterion according to which an agent should choose an action which guarantees the highest utility, given what she believes the other agents will do, or the prudential criterion according to which an agent should not choose an action which ensures the lowest utility, given what she believes the other agents will do. An example of the former is expected utility maximization, while an example of the latter is weak rationality in the sense of [151] (cf. also [104, 18]), according to which an agent should not choose an action which is

[4]This is the typical condition of iterated deletion of strongly dominated strategies (also called iterated strong dominance).

[5]This condition is required in order to ensure that the agents will converge to a Nash equilibrium.

strongly dominated by another action, given what the agent believes the other agents will do.

Much of the work in the field of epistemic game theory is based on a *quantitative* representation of uncertainty and epistemic attitudes. Notable examples are the analysis of the epistemic foundations for forward induction and for iterated admissibility based on Bayesian probabilities [138, 67], conditional probabilities [14] or lexicographic probabilities [25]. There are also few examples of *qualitative* models of epistemic attitudes and uncertainty which have been used in the field of epistemic game theory. For instance, Baltag et al. [12] have proposed an analysis of the epistemic foundation for backward induction based on a purely qualitative notion of plausibility. The distinction between quantitative and qualitative approaches to uncertainty has been widely discussed in the AI literature (cf. [60]). While in quantitative approaches belief states are characterized by classical probabilistic measures or by alternative numerical accounts, such as lexicographic probabilities or conditional probabilities [14], qualitative approaches do not use any numerical representation of uncertainty but simply a plausibility ordering on possible worlds structures inducing an epistemic-entrenchment-like ordering on propositions.

My interest in epistemic game theory is a natural consequence of my general interest in the theory mental attitudes. I believe that epistemic game theory offers the right framework to clarify how agents' mental attitudes influence behaviours of agents in a social setting. In particular, it allows us to understand the subtle connection between beliefs, preferences and decision, as represented in Figure 3.1, under the assumption that the agents' decisions are interdependent, in the sense they are affected by what they believe the others will choose.[6]

My work in the area of epistemic game theory and interactive epistemology has been devoted to develop a number of logics for epistemic game theory based on a qualitative representation of epistemic individual and collective attitudes including knowledge, belief, strong belief, common knowledge and common belief. The main motivation for this work is to show that interesting results about the epistemic foundation for solution concepts in game theory can be proved in a qualitative setting, without necessarily exploiting the complex machinery of probability theory. I have also worked on the connection between logical models of epistemic states based on Kripke semantics and formal models of epistemic states based on the concept of type space. While the former have been mainly proposed by logicians in AI [48] and philosophy [140], the latter have been proposed by game theorists in economics [72]. The main motivation for this work lies in the possibility of building a bridge between two research communities that study the same concepts and phenomena from different perspectives.

[6] As observed in Section 3.1, epistemic game theory and, more generally, game theory share with Figure 3.1 the concepts of belief and preference. However, they do not provide an account of the origin of beliefs, desires and moral values and of the connection between desires, moral values and preferences. Moreover, the concept of future-directed intention is not included in the conceptual apparatus of game theory and epistemic game theory. The same can be said for goals: the concept of goal is somehow implicit in the utility function but is not explicitly modeled.

My work on qualitative epistemic game theory and on the connection between Kripke semantics and type space semantics In [**my papers: A18**], which extends and improves over [**my papers: C39**], we have proposed some variants of a multi-modal logic of joint action, preference and knowledge that support reasoning about epistemic games in strategic form. We have given sound and complete axiomatizations of these logics as well as some complexity results for the satisfiability problem. We have considered both games with complete information and games with incomplete information. In particular, we have provided syntactic derivations of some well-known theorems in the epistemic game theory literature that specify some sufficient epistemic conditions of equilibrium notions such as Nash equilibrium and "iterated deletion of strongly dominated strategies" (IDSDS). Furthermore, we have provided a proof of Harsanyi's famous claim that all uncertainty about the structure of a game can be reduced to uncertainty about utilities [72].

In [**my papers: A16**], we have presented a branching-time epistemic temporal logic which supports reasoning about interacting agents in the context of extensive form games. The semantics of the logic is bidimensional as it quantifies over both strategies and vertices within a game tree. The logic is used to give a Hilbert-style syntactic derivation of Aumann's famous theorem that states the following: "for any non degenerate game of perfect information, common knowledge of rationality implies the backward induction solution" [6]. The interesting aspect of syntactic derivations is that they allow us to offer an in-depth formal analysis of the hypotheses that are needed to prove some well-known theorems in the epistemic game theory literature, as they make them explicit in the proofs. Specifically, while Aumann used the epistemic logic $S5^n$ to prove his theorem, thanks to our syntactic derivation, we have been able to show that the notion of belief formalized by the modal logic system $KD45^n$ is sufficient to prove this theorem. Moreover, we have shown that the characterization of epistemic conditions of IDSDS does not require introspection properties for belief and knowledge, as the modal logic systems KT^n or KD^n are both sufficient for this characterization.

In [**my papers: A12**], we have presented a combination of van Benthem et al.'s variant of PDL which gives an epistemic interpretation to programs [154] with Spohn's theory of uncertainty and belief change [137]. We have shown that the logic allows us to describe both the static and the dynamic properties of different kinds of individual and collective epistemic attitudes including knowledge, belief, graded belief, robust belief and common belief which provide the epistemic foundations of different solution concepts in game theory.

In [**my papers: A3**] we have dealt with the connection between two families of formal semantics for knowledge and belief used in the area of interactive epistemology, those based on Kripke semantics, mainly used by logicians in AI and philosophical logic, and those based on type spaces, widely used by game theorists. We have provided a formal comparison between the two and a statement of semantic equivalence with respect to three different logical systems: a doxastic logic for belief, an epistemic-doxastic logic for belief and knowledge and a logic of probabilistic beliefs and knowledge. Moreover, we have provided sound and complete axiomatizations of these logics with respect to the two equivalent Kripke semantics and type space semantics.

To better illustrate my contribution in the area of epistemic game theory, in the next paragraph I will discuss in more detail my most recent work in this area.

	C	D
C	2,2	0,3
D	3,0	1,1

Figure 3.3: Prisoner's dilemma (with player 1 being the row player and player 2 being the column player).

Epistemic foundation of DWDS^2-IDSDS In [**my papers: A5**] we have developed a minimal logic for epistemic game theory and shown that the logic is sufficiently expressive to provide epistemic foundations for various game-theoretic solution concepts including "two rounds of iterated deletion of weakly dominated strategies followed by iterated deletion of strongly dominated strategies" (DWDS^2-IDSDS). The logic is "minimal" in the sense we make as few assumptions as possible about the representation of epistemic states. The only assumption we make is that players in a game are capable of ordering the situations they envisage according to their plausibility, without necessarily being capable of assigning a numerical value to them. We have provided a complete axiomatization for this logic as well as a complexity result for its satisfiability problem.

DWDS^2-IDSDS is an interesting solution concept as it is tightly related with the concept of forward induction. The latter has been used to explain some empirical evidences and to justify the behavior of rational players in the context of extensive games such as the well-known "battle of the sexes with outside option" [39].

The formal semantics of the logic presented in [**my papers: A5**] is based on the concept of epistemic game model (EGM). EGMs extend multi-relational Kripke models commonly used by logicians and computer scientists to model epistemic concepts with a choice component. They provide a purely qualitative representation of knowledge and belief as no notion of probability is involved or other numerical representation of the strength of belief. A graphical representation of an EGM for the well-known Prisoner's Dilemma (PD) of Figure 3.3 is given in Figure 3.4. PD is a two-player game in which each player can decide either to cooperate (action C) or to defect (action D) and has an incentive to defect. Indeed, it is assumed that, if an agent defects, she gets a reward that is higher than the reward obtained in the case of cooperation, no matter what the other agent decides to do. In other words, cooperation is strongly dominated by defection. The social dilemma lies in the fact that mutual defection, the only Nash equilibrium of the game, ensures a utility for each agent that is lower than the utility obtained in the case of mutual cooperation.

EGMs require that that every player in a game has only two kinds of epistemic attitudes: (i) each player envisages a set of worlds or states (also called epistemic alternatives) that defines the player's *information set*, and (ii) among these epistemic alternatives the player can distinguish possible alternatives, that define the player's *belief set*, from impossible ones. Moreover, every world in a EGM is associated with a unique action profile where each player chooses exactly one action. Two additional constraints are given in the definition of EGM: (i) an ex-interim condition according to

which an agent chooses a certain action if and only if she knows this, and (ii) a cautiousness condition according to which every player must envisage all possible choices of the other players. The latter condition excludes that if there exists a correlation between a player's choice and the choices of the others, then the player rules out from her information set all choices of the others that are not correlated with her current choice. For instance, in Figure 3.4, full ellipses and rectangles represent, respectively, the information sets of player 1 and of player 2, while dotted ellipses and rectangles represent, respectively, the belief sets of player 1 and of player 2. At world w player 1 believes that player 2 will cooperate since for all worlds in her belief set at w player 2 chooses action C. Notice that the ex-interim condition is satisfied as, if a certain player chooses a certain action then she chooses the same action in all worlds in her information set. Notice also that the cautiousness condition is satisfied as, the information set of a certain player is sufficiently rich to include all possible choices of the other player.

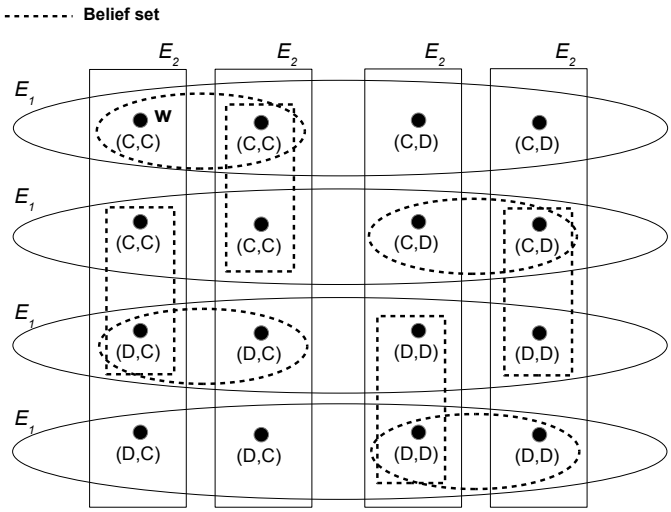


Figure 3.4: Example of EGM with two players

The basic logical language interpreted over EGMs includes standard Boolean constructions *plus*:

- choice constants of the form $\mathsf{pl}(J, s_J)$ that has to read "coalition J plays (or chooses) the action profile s_J",
- modal operators for knowledge of the form K_i, whose dual is denoted by $\widehat{\mathsf{K}}_i$, where $\mathsf{K}_i\varphi$ has to read "agent i knows that φ is true",
- modal operators for belief of the form B_i, whose dual is denoted by $\widehat{\mathsf{B}}_i$, where $\mathsf{B}_i\varphi$ has to read "agent i believes that φ is true",
- modal operators for common knowledge and common belief of the form CK and

CB where $\mathsf{CK}\varphi$ and $\mathsf{CB}\varphi$ have to read, respectively, "the agents have common knowledge that φ is true" and "the agents have common belief that φ is true".

For example at world w in the EGM of Figure 3.4 the formula $\mathsf{B}_1\mathsf{pl}(2, C)$ and $\mathsf{B}_1\mathsf{K}_2\mathsf{pl}(2, C)$ are both true.

It is shown that the logic can express the concept of 'strong belief', denoted by the symbol SB_i, which is crucial to characterize the epistemic conditions of DWDS^2-IDSDS. Specifically, $\mathsf{SB}_i\varphi$ means that all worlds that player i envisages in which φ is true are strictly more plausible than the worlds that player i envisages in which φ is false. For example, at world w, not only player 1 believes that player 2 will cooperate ($\mathsf{B}_1\mathsf{pl}(2, C)$), but also she strongly believes this ($\mathsf{SB}_1\mathsf{pl}(2, C)$), as all worlds in 1's information set at w in which 2 cooperates are strictly more plausible than the worlds in which 2 defects.

Different concepts of rationality are defined in the logic ranging from weak rationality in the sense of [151] to the lexicographic notion of perfect rationality. In particular:

- *weak rationality*: a player is weakly rational if and only if she chooses an action which is not strongly dominated within her belief set;
- *strong rationality*: a player is strongly rational if and only if she chooses an action which is not weakly dominated within her belief set;
- *perfect rationality*: a player is perfectly rational if and only if she chooses an action which is not weakly dominated within her belief set and not weakly dominated within her information set, after having discarded all weakly dominated strategies within her belief set.

It is shown that the concept of perfect rationality is essential to provide the epistemic foundation of DWDS^2-IDSDS. Specifically, the following theorem of the logic is proved:

Theorem 1 *Let*

$$\mathsf{ComplAss} \stackrel{\mathtt{def}}{=} \bigwedge_{i \in Agt} \bigwedge_{s_{-i} \in S_{-i} : s_{-i} \in S_{-i,1}^{DWDS^2-IDSDS}} \widehat{\mathsf{K}}_i(\mathsf{pl}(-i, s_{-i}) \wedge \mathsf{AllPRat}_{-i}).$$

Then:

$$\models \mathsf{CB}(\mathsf{ComplAss} \wedge \mathsf{AllPRat}_{Agt} \wedge \bigwedge_{i \in Agt} \mathsf{SB}_i\mathsf{AllPRat}_{-i}) \rightarrow \bigvee_{s \in S^{DWDS^2-IDSDS}} \mathsf{pl}(Agt, s)$$

The theorem states the following. Assume that the players have common belief that, for every player i and for every non-weakly dominated strategy of the other players, there must be a world envisaged by i in which this strategy is played and the other players are perfectly rational. This is called 'completeness assumption' and is represented by the formula $\mathsf{ComplAss}$. Then, under this assumption, the common belief that every player is perfectly rational and has a robust belief about the perfect rationality of the other players is a sufficient condition for DWDS^2-IDSDS.

The converse of the previous theorem is also proved stating that for every strategy profile that survives DWDS^2-IDSDS we can find an epistemic game model in which this strategy profile is played and the players have common belief that the completeness assumption $\mathsf{ComplAss}$ holds and that every player is perfectly rational and has a robust belief about the perfect rationality of the other players.

Theorem 2 *Let $s \in S$. Then, if $s \in S^{DWDS^2-IDSDS}$ then there exists an epistemic game model M and a world w in M such that the formula $\mathsf{pl}(Agt, s) \wedge \mathsf{CB}(\mathsf{ComplAss} \wedge \mathsf{AllPRat}_{Agt} \wedge \bigwedge_{i \in Agt} \mathsf{SB}_i \mathsf{AllPRat}_{-i})$ is true at w in M.*

The two Theorems 1 and 2 together provide a characterization of the epistemic conditions of DWDS^2-IDSDS.

3.3 Emotion

In the recent years, emotion has become a central topic in AI. The main motivation of this line of research lies in the possibility of developing computational and formal models of artificial agents who are expected to interact with humans. To ensure the accuracy of a such formal models, 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 her 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), such agents must be endowed with an adequate model of human emotions.

The most popular psychological theory of emotion in AI is the so-called appraisal theory (cf. [124] for a broad introduction to the developments in appraisal theory). This 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 a disease and considering 'I catch a disease' as an undesirable event, an agent might feel an intense fear.

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 [127, 61] 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 [118] 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 φ can be reduced to the person's belief that φ obtains and the person's desire that φ obtains. On the contrary, according to cognitive-evaluative models, a person feels happy about a certain fact φ if she believes that φ obtains and she evaluates φ to be good (desirable) for her.

The popularity of appraisal theory in logic and AI is easily explained by the fact that it perfectly fits with the concepts and level of abstraction of existing logical and computational models of cognitive agents developed in these areas. Especially cognitive-motivational models use folk-psychology concepts such as belief, knowledge, desire and intention that are traditionally used in logic and AI for modelling cognitive agents.

The conceptual background underlying my view of appraisal theory is depicted in Figure 3.5 which is nothing but the cognitive architecture of Figure 3.1 extended with an emotion component.

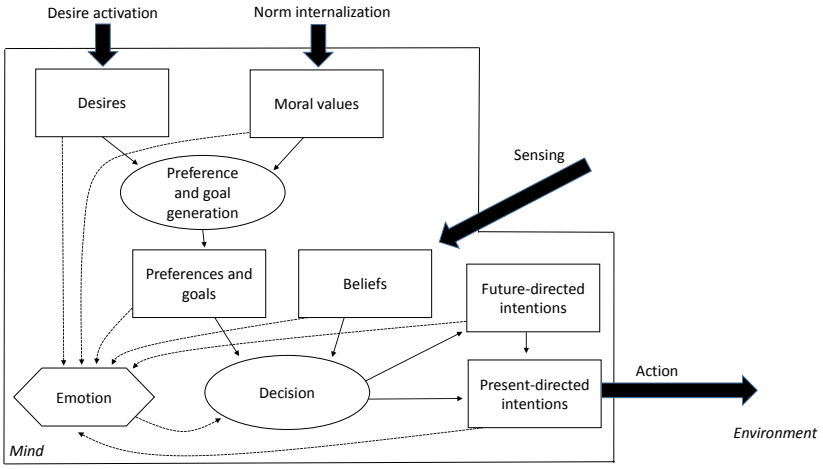


Figure 3.5: Cognitive architecture extended with emotions

Figure 3.5 highlights the role of mental attitudes in emotion. In particular, it highlights the fact that mental attitudes of different kinds such as belief, desires, preferences, goals, moral values and (present-directed or future-directed) intentions determine emotional responses. For example, as emphasized above, the emotional response of happiness is triggered by a *goal* and the *certain belief* that the content of one's goal is true. On the contrary, the emotional response of sadness is triggered by a *goal* and the *certain belief* that the content of one's goal is false. The emotional response of hope is triggered by a *goal* and the *uncertain belief* that the content of one's goal is true. On the contrary, the emotional response of fear is triggered by a *goal* and the *uncertain belief* that the content of one's goal is false. This view is consistent with a famous appraisal model, the so-called OCC psychological model of emotions [106], according to which, while joy and distress are triggered by *actual consequences*, hope

and fear are triggered by *prospective consequences* (or *prospects*). [63] interpret the term 'prospect' as synonymous of 'uncertain consequence' (in contrast with 'actual consequence' as synonymous of 'certain consequence').

Moral guilt and reproach are examples of emotion which are triggered by moral values. While moral guilt is triggered by the *belief* of being responsible for the violation of a *moral value*, reproach is triggered by the *belief* that someone else is responsible for the violation of a *moral value*. In other words, guilt is triggered by self-attribution of responsibility for the violation of a moral value, while reproach is triggered by attribution to others of responsibility for the violation of a moral value. Intentions as well might be responsible for triggering certain kinds of emotional response. For instance, as emphasized by psychological theories of anger (e.g., [85, 106, 122]), a necessary condition for an agent 1 to be angry towards another agent 2 is the agent 1's belief that agent 2 has performed an action that has damaged her, that is, 1 believes that she has been kept from attaining an important goal by an improper action of agent 2. Anger becomes more intense when agent 1 believes that agent 2 has *intentionally* caused the damage. In this sense, an agent 1's belief about another agent 2's intention may have implications on the intensity of agent 1's emotions.

Figure 3.5 also represents how emotions retroactively influence mental states and decision either (i) through coping or (ii) through anticipation and prospective thinking (i.e., the act of mentally simulating the future) in the decision-making phase.

Coping is the process of dealing with emotion, either externally by forming an intention to act in the world (problem-focused coping) or internally by changing the agent's interpretation of the situation and the mental attitudes that triggered and sustained the emotional response (emotion-focused coping) [85]. For example, when feeling an intense fear due to an unexpected and scaring stimulus, an agent starts to reconsider her beliefs and intentions in order to update her knowledge in the light of the new scaring information and to avoid running into danger (emotion-focused coping). Then, the agent forms an intention to go out of danger (problem-focused coping). Another agent can try to discharge her feeling of guilt for having damaged someone either by forming the intention to repair the damage (problem-focused coping) or by reconsidering the belief about her responsibility for the damage (emotion-focused coping). The coping process as well as its relation with appraisal is illustrated in Figure 3.6.

The influence of emotion on decision-making has been widely studied both in psychology and in economics. Rick & Loewenstein [121] distinguish the following three forms of influence:

- **Immediate emotions**: real emotions experienced at the time of decision-making:
  - **Integral influences**: influences from immediate emotions that arise from contemplating the consequences of the decision itself,
  - **Incidental influences**: influences from immediate emotions that arise from factors unrelated to the decision at hand (e.g., the agent's current mood or chronic dispositional affect);
- **Anticipated emotions**: predictions about the emotional consequences of decision outcomes (they are not experienced as emotions per se at the time of decision-making).

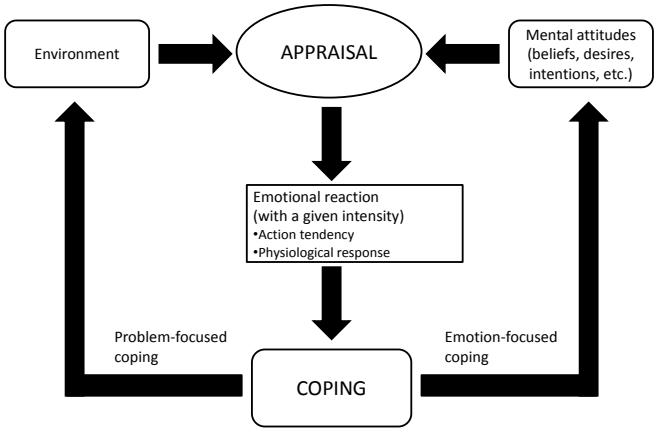


Figure 3.6: Appraisal and coping cycle

An example of integral influence of an immediate emotion is given by the following example.

Example 1 *Paul would like to eat some candies but her mother Mary has forbidden him to eat candies without her permission. Paul's fear of the sanction influences Paul's decision not to eat candies without asking permission.*

The following example illustrates incidental influence of an immediate emotion.

Example 2 *Mary has quarreled with her colleague Paul. At the end of the day she goes back home after work and on the metro a beggar asks her for money. Few hours after the quarrel with Paul, Mary is still in a bad mood and because of her current disposition she refuses the beggar's request.*

The following example illustrates the influence of anticipated emotions on decision.

Example 3 *Peter has to decide whether to leave her job as a researcher at the university of Paris and to accept a job offer as a professor at a university in the U.S. She decides to accept the job offer because she thinks that, if she refuses it, she will likely regret her decision.*

One of the most prominent theory of the integral influence of emotion on decision is Damasio's theory of the somatic marker [41]. According to this theory, decision between different courses of actions leads to potentially advantageous (positive) or harmful (negative) outcomes. These outcomes induce a somatic response used to mark them and to signal their danger or advantage. In particular, a negative somatic marker 'signals' to the agent the fact that a certain course of action should be avoided, while a positive somatic marker provides an incentive to choose a specific course of

action. According to Damasio's theory, somatic markers depend on past experiences. Specifically, pain or pleasure experienced as a consequence of an outcome are stored in memory and are felt again when the outcome is envisaged in the decision-making process. The following example clearly illustrates this.[7]

Example 4 *Mary lives in Toulouse and has to decide whether to go to Paris by plane or by train. Last time she traveled by plane she had a painful experience because of turbulence. Mary envisages the possibility of incurring again in a turbulence and gets frightened, thereby deciding to travel by train.*

Several works aimed at extending the classical expected utility model to incorporate anticipated emotions which are related to our uncertainty about the future, such as hopefulness, anxiety, and suspense [31]. Some economic models of decision-making consider how the anticipation of a future regret might affect a person's current decision [94]. In particular, according to these models, if a person believes that after choosing a certain action she will likely regret for having made this choice, she will be less willing to choose the action (than in the case in which she does not believe this). These models agree in defining regret as the emotion which 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 [52, 82, 171]. More recently, some economists have studied the influence of strategic emotions such as interpersonal guilt and anger on decision [13, 34, 77]. Following psychological theories of interpersonal guilt [15, 144], models developed in this area assume that the prototypical cause of guilt is the infliction of harm, loss, or distress on a relationship partner. Moreover, they assume that if people feel guilty for hurting their partners and for failing to live up to their expectations, they will alter their behavior (to avoid guilt) in ways that seem likely to maintain and strengthen the relationship.

My work on the logical formalization of emotions Since my PhD, I have been working on the formalization of emotions using the tools and techniques of logic and game theory. I have mainly adopted appraisal theory as the conceptual basis of my work.

The emotion I have started with is surprise. Surprise is the simplest emotion that is triggered by the mismatch between an expectation that an event will possibly occur and an incoming input (i.e., what an agent perceives). We have a provided a formal theory of surprise that clarifies two important aspects of this cognitive phenomenon [**my papers: A26,A27,B16,B19**]. First, it addresses the distinction between surprise and astonishment, the latter being the emotion triggered by something an agent could not reasonably expect. The crucial difference between surprise and astonishment is that the former necessarily requires an explicit expectation in the agent's mind, while the latter does not. One can be astonished by something since, at the moment she perceives it,

[7] Positive and negative somatic markers can operate either at a conscious level or at a unconscious/automatic level. This corresponds to Ledoux's distinction between explicit memory and implicit memory and between two possible elaborations of a stimulus inducing an emotional response [86]: conscious elaboration vs. automatic elaboration.

she realizes that it was totally unpredictable, without having formulated an expectation in advance. For example, suppose Mary is working in her office. Suddenly, someone knocks the door and enters into Mary's office. Mary sees that the person is a policeman. She is astonished by this fact even though, before perceiving it, she did not have explicit in her mind the expectation that "a policeman will not enter into the office". Secondly, the theory clarifies the role of surprise in belief change by conceiving it as a basic mechanism which is responsible for triggering belief reconsideration. The formal theory of surprise we have developed exploits methods and techniques from Dynamic Epistemic Logic (DEL). The latter is particularly suited to account for the dynamics of epistemic attitudes which is a crucial component of the surprise phenomenon.

A simplified version of the theory has been implemented in a computational architecture [**my papers: C32,C55**] by extending the classical BDI (Belief, Desire, Intention) cycle [**?**] with surprise. Surprise plays a specific role in the cognitive architecture, as it is responsible for triggering the process of belief revision. This is particularly important in the perspective of designing realistic non-omniscient cognitive agents who are situated in complex environments where many tasks must be solved. Since accurate belief revision and update require time and considerable computational costs, such agents need some mechanism which is responsible for signaling the global inconsistency of their belief bases with respect to an incoming input and for the revision of their beliefs and expectations. One of the adaptive functions of surprise is exactly this: it signals to the agent when her beliefs and expectations should be reconsidered.

More recently we have been working on the logical formalization of counterfactual emotions, that is, those emotions that are based on counterfactual reasoning about agents' choices [**my papers: A17,B15,C47**]. The prototypical counterfactual emotion is regret. Other examples are rejoicing, disappointment, and elation. We have proposed an epistemic extension of atemporal STIT logic (cf. Section 2.1) which allows to capture the cognitive structure of regret and, in particular, the conterfactual belief which is responsible for triggering this emotion, namely the *belief that a counterfactual better outcome might have been, had the agent chosen a different action*. We have also studied the mathematical as well as the computational aspects of this epistemic extension of atemporal STIT including axiomatics and complexity of the satisfiability problem. This work has been included in François Schwarzentruber's PhD thesis that I have co-supervised with other colleagues from IRIT.

In [**my papers: A8**], we have extended our STIT logical analysis of counterfactual emotions to moral emotions. The latter involve counterfactual reasoning about responsibility for the transgression of moral values. In particular, the proposed formalization accounts for the attribution of responsibility for the violation of a moral value either to the self or to the other. This is a fundamental constituent of moral emotions such as guilt, reproach, moral pride and moral approval. For example, according to the analysis we propose, guilt is triggered by the *belief that one is responsible for having behaved in a morally reprehensible way*. We have also proposed a game-theoretic account of moral guilt [**my papers: C5,C18**], in parallel with the STIT logical analysis of moral emotions.

We have also worked on the problem of emotion intensity by proposing a logical theory of the intensity of hope, fear, joy and sadness [**my papers: B13,C23,C26**]. Following existing psychological models of emotion based on appraisal theory, intensity

of these emotions is defined as a function of two cognitive parameters, the strength of the expectation and the strength of the desire which are responsible for triggering the emotional response. For instance, the intensity of hope that a certain event will occur is a monotonically increasing function of both the strength of the expectation and the strength of the desire that the event will occur. The theory also considers the behavioral aspects of such emotions: how the execution of a certain coping strategy depends on the intensity of the emotion generating it. Specifically, it is assumed that: (i) an agent is identified with a numerical value which defines her tolerance to the negative emotion, and (ii) if the intensity of the negative emotion (e.g., fear) exceeds this value then the agent will execute a coping strategy aimed at discharging the negative emotion.

In a related work [**my papers: C25,C30,C31**], we have merged speech act theory and emotion theory in order to study the expressive aspect of emotions. Specifically, we have proposed a logical formalization of expressive speech acts, that is, speech acts that are aimed at expressing a given emotion (e.g. to apologize, to thank, to reproach, etc.). This work on expressive speech acts has been included in Nadine Guiraud's PhD thesis that I have co-supervised with other colleagues from IRIT.

Chapter 4

Collectives

Agents in a society are not only individual entities sensing from the environment and acting in the pursuit of their *individual goals* and with the aim of satisfying their *individual desires* but also members of groups, teams, organizations, institutions, etc. In many circumstances, agents reason, act and think as group members and their reasoning, acting and thinking as group members radically differ from their reasoning, acting and thinking as individuals. This is typical in team sports like soccer and volleyball in which agents in the team act in the pursuit of the *common goal* of winning the match and they necessarily have to collaborate with the other team members in order to make the team performance optimal. It is well-known that sport teams in which agents in the team play individualistically are generally less performant than sport teams in which agents in the team play collaboratively. But this is also typical in our everyday life in which our behaviours are often determined by the *goals, desires and preferences of others*. When deciding what to do, we often take into account the consequences of our choices on the well-being of the others, we help the others because we empathize with them. We prefer to cut the cake in equal pieces without taking the biggest piece because we have been educated to do so, because we are sensitive to justice and averse to inequity thereby thinking that dividing the cake in equal parts between the members of our group is the right thing to do.

With the term "collective decision-making" I refer to assuming a group-oriented perspective in the decision-making phase, that is, deciding what to do (i) by taking into account the consequences of our choices on the well-being of both us and the others, (ii) by caring about social preferences, that is, not only about what we prefer but also about what the others prefer and, in more extreme cases, (iii) by identifying us and the others as members of the same group or team. Collective decision-making can be opposed to individual decision-making in which an agent only considers the consequences of her choices on her own well-being without caring about the preferences of the others.

Decision-making is not the only dimension along which "individual" can be opposed to "collective". Another important distinction, which has been deeply investigated in recent years in the philosophical area of collective intentionality [96, 147], is between individual attitudes and collective attitudes. As highlighted in Section 3.1, individual attitudes coincide with the psychological or mental states of an individual.

They can be either individual attitudes of epistemic type such as knowledge and belief or individual attitudes of motivational type such as desire, goal and moral value. While individual attitudes are attitudes of individual agents, collective attitudes are attitudes of collective entities such as groups, teams and institutions. As for individual attitudes, one can distinguish collective attitudes of epistemic type from collective attitudes of motivational type. Typical examples of the former are shared belief, common knowledge, common belief, group belief and collective acceptance. Typical examples of the latter are group goals, collective preferences and joint intentions.

In the next two sections, I will discuss in more detail the topic of collective attitudes and the topic of collective decision-making, and I will present my past research on these two topics.

4.1 Collective attitudes

Collectives such as groups, teams, coorporations, organizations, etc. do not have minds. However, we frequently ascribe intentional attitudes to them in the same way as we ascribe intentional attitudes to individuals. For example, we may speak of what our family prefers, of what the goal of a coorporation or organization is, of what the scientific community think about a certain issue, and so on.

Aggregate vs. common attitudes An important distinction in the theory of collective attitudes is between aggregate attitudes and common attitudes. As emphasized by [91] "...an aggregate attitude (of a collective) is an aggregate or summary of the attitudes of the individual members of the collective, produced by some aggregation rule or statistical criterion...". A typical example of aggregate attitude produced by a statistical criterion is shared belief, namely the fact that all agents (or most of the agents) in a set of agents believe that a certain proposition p is true. An example of aggregate attitude produced by an aggregation rule is the collective judgment of a jury about a given proposition p obtained by majority voting: the jury believes that the proposition p is true if and only if the majority of the members of the jury has expressed the individual opinion that p is true. Aggregate attitudes produced by aggregation rules are the objects of analysis of social choice theory and jugement aggregation, two important research areas in social sciences and artificial intelligence. Differently from common attitudes, aggregate attitudes do not require a level of common awareness by the members of the group. That is, a group can hold an aggregate attitude even though the members of the group do not necessarily believe so. For example, the fact that two agents share the belief that p is true does not necessarily imply that they individually believe that they share this belief. As emphasized by [91] "...a common attitude (of a collective) is an attitude held by all individual members of the collective, where their holding it is a matter of common awareness", where the term "common awareness" refers to the fact that every member of the group believes that the group has the common attitude, that every member of the group believes that every member of the group believes that the group has the common attitude, and so on. A typical example of common attitude is common belief: every agent in the group believes that p is true, every agent in the group believes that every agent in the group believes that p is true, and so on ad infinitum.

Functions of collective attitudes Collective attitudes play a crucial role in the society as: (i) they provide the basis of our common understanding through communication, (ii) they ensure coordination between agents, (iii) they are fundamental constituents of collaborative activities between agents acting as members of the same team.

In linguistic, the concept of common ground in a conversation is typically conceived as the common knowledge that the speaker and the hearer have about the rules of the language they use and about the meaning of the expressions uttered by the speaker [139]. Indeed, language use in conversation is a form of social activity which requires a certain level of coordination between what the speaker means and what the addressee understands the speaker to mean. Any utterance of the speaker is in principle ambiguous because the speaker could use it to express a variety of possible meanings. Common ground — as a mass of information and facts mutually believed by the speaker and the addressee — ensures coordination by disambiguating the meaning of the speaker's utterance. For example, suppose two different operas, "Don Giovanni" by Mozart and "Il Barbiere di Siviglia" by Rossini, are performed in the same evening at two different theaters. Mike goes to see Don Giovanni and the next morning sees Mary and asks "Did you enjoy the opera yesterday night?", identifying the referent of the word "opera" as Don Giovanni. In order to be sure that Mary will take "opera" as referring to Don Giovanni and not to Il Barbiere di Siviglia, Mike has to believe that the night before Mary too went to see Don Giovanni, that Mary believes that Mike too went to see Don Giovanni, that Mary believes that Mike believes that Mary too went to see Don Giovanni, and so on.

Moreover, since the seminal work by David Lewis [88], the concept of common belief has been show to play a central role in the formation and emergence of social conventions.

Finally, collective attitudes such as common goal and joint intention are traditionally used in in the philosophical area and in AI to account for the concept of collaborative activity [27, 64, 44, 45]. Notable examples of collaborative activity are the activities of painting a house together, dancing together a tango, or moving a heavy object together. Two or more agents acting together in a collaborative way need to have a common goal and need to form a shared plan aimed at achieving the common goal. In order to make collaboration effective, each agent has to commit to her part in the shared plan and form the corresponding intention to perform her part of the plan. Moreover, she has to monitor the behaviors of the others and, eventually, to reconsider her plan and adapt her behavior to the new circumstances.

The origin of collective attitudes Where do collective attitudes come from? How are they formed? There is no single answer to these questions, as collective attitudes can originate in many different ways.

As explained above, aggregate attitudes are the product of aggregation procedures like majority voting or unanimity (cf. [92]). The agents in a certain group decide to use a certain aggregation rule. Then, every agent expresses her opinion about a certain issue p and the aggregation rule is used to determine what the group believes or what the group accepts. Examples of collective attitudes originating from the aggregation of individual attitudes are group belief and collective acceptance.

Collective attitudes, such as shared belief and common belief, can also be formed through communication. A source of information announces to all agents in a group that a certain proposition p is true. Under the assumption that every agent perceives what the information source says and that every agent in the group trusts the information source's jugement about p, the agents will share the belief that p is true as a result of the announcement. Creation of common belief through communication requires satisfaction of certain conditions that are implicit in the concept of public announcement, as defined in the context of public announcement logic (PAL) [111], the simplest logic in the family of dynamic epistemic logics (DEL). Specifically, to ensure that an announcement will determine a common belief that the announced fact is true, every agent in the group has to perceive what the information source says, every agent in the group has to perceive that every agent in the group perceives what the information source says, and so on. The latter is called *co-presence* condition in the linguistic literature [36].

The concept of co-presence becomes particularly relevant in the perspective of designing artificial systems situated in a physical environment that need to acquire common belief of certain facts in order to achieve coordination and to make collaboration effective. For example, imagine two robots moving in the physical environment. A source of information signals to them that there is a danger. It does this by emitting a red light. The robots will be able to form different levels of mutual belief about this fact depending on: (i) their spatial positions and the orientation of their sensors with respect to the source of information, and (ii) the perception of the other robots' spatial positions and of the orientations of the other robots' sensors with respect to the source of information. The concept of co-presence applies not only to agents interacting in a physical environment but also to agents interacting in a virtual environment (e.g., virtual characters of a videogame).

Collective acceptance and institutions The problem of understanding what institutions are and how they work has been addressed both in social sciences, in philosophy and in legal theory. Computer scientists working in the area of multi-agent systems have been interested in devising artificial institutions, modeling their dynamics and the different kinds of rules and norms of an institution that agents have to deal with. Following [105, p. 3], artificial institutions can be conceived as "the rules of the game in a society or the humanly devised constraints that structure agents' interaction". In some models of artificial institutions norms are conceived as means to achieve coordination among agents and agents are supposed to comply with them and to obey the authorities of the system [47]. More sophisticated models of institutions leave to the agents' autonomy the decision whether to comply or not with the specified rules and norms of the institution [1, 95]. However, all previous models abstract away from the legislative source of the norms of an institution, and from how institutions are created, maintained and changed by their members.

What these models of artificial institutions neglect is the fundamental relationship between institutions and the collective attitudes of their members and, in particular, the fact that the existence and the dynamics of an institution (norms, rules, institutional facts, etc.) are determined by the collective attitudes of the agents which identify them-

selves as members of the institution. This aspect is emphasized in the following quote from [97, p. 77]:

> "only because institutions are anchored in peoples minds do they ever become behaviorally relevant. The *elucidation of the internal aspect is the crucial step* in adequately explaining the emergence, evolution, and effects of institutions." [Emphasis added].

Prominent philosophical theories of institutional reality conceives collective acceptance as the collective attitude on which institutions are grounded [128, 149]. The relationship between acceptance and institutions has also been emphasized in the philosophical doctrine of Legal Positivism [73]. According to Hart, the foundations of an institution consist of adherence to, or acceptance of, an ultimate rule of recognition by which the validity of any rule of the institution may be evaluated. [1]

A property that clearly distinguishes collective acceptance from common belief is that common belief implies shared belief, while collective acceptance does not: when there is a common belief in a group of agents C that a certain proposition p is true then each agent in C individually believes that p is true, while it might be the case that there is a collective acceptance in C that p is true, and at the same time one or several agents in C do not individually believe that p is true. For example, the members of a Parliament might collectively accept (*qua* members of the Parliament) that launching a military action against another country is legitimate because by majority voting the Parliament decided so, even though some of them — who voted against the military intervention — individually believe the contrary. This difference is due to the fact that collective acceptance is a kind of aggregate attitude which can be formed through aggregation procedures others than unanimity.

Another important difference between collective acceptance and common belief is the irreducibility of collective acceptance to the individual level. In particular, it has been emphasized that, while common belief is strongly linked to individual beliefs and can be reduced to them, collective attitudes such as collective acceptance cannot be reduced to a composition of individual attitudes. This aspect is particularly emphasized by Gilbert [56] who follows Durkheim's non-reductionist view of collective attitudes [46]. According to Gilbert, any proper group attitude cannot be defined only as a label on a particular configuration of individual attitudes, as common belief is. In [57, 150] it is suggested that a collective acceptance of a set of agents C is based on the fact that the agents in C identify themselves as members of a certain group, institution, team, organization, etc. and recognize each other as members of the same group, institution, team, organization, etc. Common belief and common knowledge, as traditionally defined in epistemic logic [48], do not entail this aspect of mutual recognition and identification with respect to the same group, institution, team, organization, etc.

My work on the logical formalization of collective attitudes In the recent years I have been interested in the logical theory of collective attitudes by focusing on the three aspects discussed above, namely the relationship between collective acceptance and

[1] In Hart's theory, the rule of recognition is the rule which specifies the ultimate criteria of validity in a legal system.

institutions, the perceptual origin of common belief and the formalization of common ground.

We have proposed a modal logic of collective acceptance in accordance with existing with philosophical theories of this notion [**my papers: A24,B2,C42,C46,C48,C51**]. In the logic, collective acceptance is conceived as the collective attitude that some agents have *qua* members of the same institution. In particular, a collective acceptance held by a set of agents C *qua* members of a certain institution x is the kind of acceptance the agents in C are committed to when they are "functioning together as members of the institution x", that is, when the agents in C identify and recognize each other as members of the institution x. For example, in the context of the institution Greenpeace agents (collectively) accept that their mission is to protect the Earth *qua* members of Greenpeace. The state of acceptance *qua* members of Greenpeace is the kind of acceptance these agents are committed to when they are functioning together as members of Greenpeace, that is, when they identify and recognize each other as members of Greenpeace. The logic accounts for different kinds of aggregation procedures that the members of an institution may adopt in order to build a collective acceptance of a given fact. This includes unanimity, majority and a criterion based on leadership according to which what the members of an institution collectively accept coincides with the acceptance of the legislator of the institution. Moreover, the logic clearly distinguishes collective acceptance from common belief, by emphasizing the fact that, while common belief is reducible to individual beliefs, collective acceptance cannot be reduced to individual attitudes of the members of an institution. The fact that collective acceptance is not reducible to individual attitudes is reflected in the formal semantics of the logic. While in epistemic logic common belief is commonly represented by means of the transitive closure of the union of the accessibility relations for the individual beliefs, the accessibility relation for collective acceptance is not definable in terms of the accessibility relations for individual beliefs or individual acceptances. Moreover, collective acceptance entails the notion of "group identification" that is not reducible to the individual level. We have provided a complete axiomatization for the logic of collective acceptance and a decidability result for its satisfiability problem. Following the idea of some prominent philosophical theories of institutions [128, 149] according to which institutional reality only exists in relation with the collective acceptance of institutional facts by the members of the institution. We have provided a systematic analysis of institutional concepts in the context of this logic. This includes the concepts of weak permission, strong permission, obligation and constitutive rule. We have also proposed a dynamic extension of the logic with the aim of modelling institutional change in the logic [**my papers: A23,C44**]. The idea is that institutional change is a consequence of the evolution of collective acceptance.

This work on the logic of collective acceptance has been included in Benoit Gaudou's PhD thesis that I have co-supervised with other colleagues from IRIT.

I have also been interested in the study of the the perceptual origin of common belief since the time of my PhD. The original idea was to extend the logic of common belief with a perceptual component in order to formally capture the concept of co-presence [**my papers: C59**]. More recently, we have developed a logic of attention-based announcements reflecting the intuition that an agent changes her beliefs due to an announcement only if she pays attention to the source of the announcement [**my pa-**

pers: A4,C15]. Thus, an announcement is public insofar as every agent pays attention to the source of the announcement and the agents have common belief about this. At the technical level, the update semantics of the logic corresponds to a specific type of event model in the sense of [10]. However, its interesting aspect is that it allows to represent the update operation is a more compact way that in the traditional event model semantics à la Baltag & Moss. A complete axiomatization for this logic has been given as well as a complexity result and a tableau method for its satisfiability problem.

As for the concept of common ground, we have recently proposed a logical account of the distinction between direct and indirect common belief [**my papers: B5**]. The distinction relates to the way a common belief is generated. A common belief that p is direct when it is an immediate consequence of an event e that is manifest to all those sharing the prior common belief that perceiving e entails the truth of p. A common belief that p is indirect when it is generated by what may be called a "shared inference" of the agents in the group. The inference is "constructive" in the sense that it does not presuppose the prior common belief that perceiving e entails the truth of p. This kind of shared inference is often in place in linguistic communication to guarantee that the speaker and the hearer achieve common belief about the meaning expressed by a given utterance of the speaker, relying on their mutually shared background information in the common ground.

4.2 Collective decision-making

In recent times, several experiments have demonstrated that people are often driven by other-regarding motivations, and by social preferences. For instance, it has been shown that in the one-shot version of the Prisoner's Dilemma represented in Figure 3.3 played with material incentives people often exhibit mutual cooperation [123], i.e., they choose the strategy profile (C,C). This is apparently in contrast with what game theory prescribes, as the only Nash equilibrium in this game is mutual defection, i.e., the strategy profile (D,D). Consequently, the classical game-theoretic framework has been extended in order to incorporate new important concepts such as fairness, inequity aversion, altruism and reciprocity.

Several models of other-regarding motivations and social preferences have been developed in recent years not only in economics [50], but also in sociology [98], social psychology [93] and artificial intelligence [55]. These models show that that the phenomena observed in experiments with humans can be explained in a rigorous and tractable manner. Some models assume that people are only concerned about the distributional consequences of their acts but they do not care about the intentions that lead their opponents to choose these acts. For example, according to [49], human agents could be inequity averse and care about how much material resources are allocated to other agents. Other models assume that the behavior of a person during social interaction depends on the person's beliefs and expectations about the intention of the opponent. For example, according to Rabin's model of reciprocity [114], a person is willing to be kind with another person if she believes that the other has been kind to her. On the contrary, a person has a desire to retaliate, if she believes that the other person wanted to hurt her.

Social preferences Following [58], we can distinguish self-regarding agents, or agents motivated by *individual preferences*, from other-regarding agents, or agents motivated by *social preferences*. A self-regarding agent is an agent who acts in order to achieve her private interests and to maximize the personal benefit she gets in a given situation. On the other hand, an other-regarding agent is an agent who is also concerned about the benefit other agents derive from a given situation. In other words, if an agent is self-regarding, the utility of a given state for her coincides with the personal benefit she will get in this state. If the agent is other-regarding, the utility of a given state for her also depends on the benefit the other agents will get in this state. In this sense, her utility might differ from her own personal benefit.

The notion of self-regarding agent should not be confused with the rationality assumption of classical game theory: according to classical game theory, individuals are rational in the sense that they maximize their utility. The notions of self-regarding agent and other-regarding agent are not in contradiction with this assumption. We can safely say that an other-regarding agent acts to maximize her utility even though she does not act to maximize her personal benefit (as she also cares about the benefit for the other agents). For example, Peter may decide to lend his bike to Mary in order to satisfy agent Mary's desire to go for a bike ride in the countryside (suppose Mary does not have a bike). In this situation, the utility of a given action for agent Peter depends on the extent to which this action promotes the satisfaction of agent Mary's desires. Therefore, Peter's act is other-regarding even though agent Peter is still acting to maximize his utility.

The concept of social preference is also compatible with the cognitive architecture illustrated in Figure 3.1. Nevertheless, it is worth noting that the architecture leads two different possible interpretations of this concept. According to the first interpretation, an other-regarding agent is an agent who is motivated by *the desire* of satisfying the other agents' desires or of distributing the material payoffs in a fair/equitable/impartial manner between the agents in the society. According to the second interpretation, an other-regarding agent is an agent who is motivated by *the moral value* of satisfying the other agents' desires or of distributing the material payoffs in a fair/equal/impartial way between the agents in the society. In other words, the origin of social preferences differ in these two views. While according to the first interpretation social preferences derive from desires and have an endogenous origin, according to the second one social preferences derive from moral values and could be the result of the internalization of some exogenous social norm. The latter view is compatible with the idea that an agent is sensitive to fairness norms, norms of justice or equity because these norms are part of the cultural and social environment in which the agent lives and have been internalized by the agent. These two views also have different implications on emotion analysis. While according to the first interpretation, an other-regarding agent behaving in a unfair/inequitable/partial manner should feel sad or frustrated, according to the second interpretation, an other-regarding agent in the same circumstances should feel guilty.

Team reasoning The game in Figure 4.1, called the Hi-Lo Matching game, has received quite a lot of attention in recent times.

	H	L
H	2,2	0,0
L	0,0	1,1

Figure 4.1: Hi-Lo

If both players choose the option H (the action "high"), each gains two units; if both choose L (the action "low"), each gains one unit; otherwise neither gains anything. There are two Nash equilibria in this game: the situation in which both players choose H (H,H), and the situation in which they both choose L (L,L). Hi-Lo is similar to a pure coordination game: the interests of the players are perfectly aligned and there are two Nash equilibria. But, differently from a pure coordination game, in Hi-Lo one of the two Nash equilibria is strictly better than the other, as it yields a higher utility to both players. In this case, we say that (H,H) utility-dominates (L,L). This is the reason why the coordination problem in the Hi-Lo seems trivial. It is clear that the players should coordinate on the preferred equilibrium (H,H). These intuitions have been confirmed by experiments with humans playing the Hi-Lo game with material payoffs: for instance it is shown in [8] that a very large majority of people choose the option H and coordinate on the preferred equilibrium (H,H).

The Hi-Lo game presents a fundamental problem for the best-response reasoning assumed in classical game theory (i.e., agents choose their best response to what they expect the others will do) and, consequently, for the notion of self-regarding agent we have discussed above. In fact, from the assumptions that the players are self-regarding and that they know the other players' choices, we cannot deduce that each player will choose H.[2] All we can say is that if a player expects that the other player will choose H then it is rational for her to choose H. But exactly the same can be said about L: if a player expects that the other player will choose L then it is rational for her to choose L. Theories of social preferences too fail to explain why people coordinate on the preferred equilibrium in the Hi-Lo game. Consider for instance a notion of social preference based on Harsany's criterion of distributive justice [70]:

> The social welfare of a given alternative is equal to the sum of the individual utilities.

Assume that a certain agent is other-regarding if she is motivated by the goal of maximizing social welfare. From the assumptions that the players are other-regarding and that they know what the other players choose, we cannot deduce that each player will choose H. All we can say is that if an other-regarding player expects that the other player will choose H then it is rational for her to choose H (since $2 + 2 > 0 + 0$). But exactly the same can be said about L: if an other-regarding player expects that the other player will choose L then it is rational for her to choose L (since $1 + 1 > 0 + 0$). Thus,

[2]The same deduction cannot be made even assuming that the players are self-regarding and that they have common knowledge of this.

the notion of other-regarding agent does not help to solve the problem presented by the Hi-Lo game.

In order to solve this problem some economists such as Michael Bacharach [7, 8] and Robert Sugden [142, 141] have studied a new mode of reasoning, called *team reasoning*, by which players can arrive at the conclusion that they ought to choose the option H when facing a Hi-Lo game. Team reasoning is the kind of reasoning that people use when they take themselves to be acting as members of a group or team [141]. That is, when an agent i engages in team reasoning, she identifies herself as a member of a group of agents C and conceives C as a unit of agency acting as a single entity in pursuit of some collective objective. A team reasoning agent acts for the interest of her group by identifying a strategy profile that maximizes the collective utility of the group, and then, if the maximizing strategy profile is unique, by choosing the action that forms a component of this strategy profile. As pointed out by Sugden [142], an agent has reason to act as a team member and to choose the action that forms a component of the strategy profile maximizing collective utility, conditional on assurance that the other agents also act as team members. That is, to act as a member of a team, one must be confident that the other agents act as members too. For example, in the case of the Hi-Lo, the unique profile that maximizes group utility is (H,H). Therefore, a team reasoner has an incentive to play his component in this profile, assuming that the other player will do the same.

Team-directed forms of reasoning have been studied not only by economists but also by philosophers [150, 57, 117] and social psychologists [38]. For instance, the philosopher Raimo Tuomela [150] distinguishes the 'we-mode' perspective from the 'I-mode' perspective. 'We-mode' is the mental attitude of identifying oneself as a team member and of framing the situation as a problem for the team, whereas 'I-mode' is the mental attitude of framing the problem as a classical decision-making problem that leads to maximization of (individual) expected utility. (See [65] for a comparison between Tuomela's concept of 'we-mode' and Bacharach's concept of team reasoning.)

My work on collective decision-making One of the limitations of existing theories of team reasoning is the assumption that a person in a given strategic setting can be either in the 'I-mode' or in the 'we-mode', but cannot be at the same time in the 'I-mode' and in the 'we-mode'. In other words, existing theories of team reasoning do not contemplate the possibility that a person can be *partially tied* with a given group or team so that she is motivated by two concomitant goals: (i) the goal of maximizing individual benefit, and (ii) the goal of maximizing the benefit for the group.

Consider, for example, the game in Figure 4.2 that represents a variant of the well-known dictator game [51]. In this game the row player has total control of the game, in the sense that the outcome of the game only depends on his choice.

Existing theories of team reasoning envisage only two possibilities in this game. If the row player is in the 'I-mode', he will certainly choose the first option (A), as this is the one that guarantees the maximal utility for him. On the contrary, if the row player is in the 'we-mode', she will certainly choose the second option (B), as this is the one that guarantees the maximal social welfare. In fact, any reasonable measure of social welfare, including Harsanyi's utilitarian criterion of distributive justice based

A	8,0
B	6,8
C	7,3

Figure 4.2: Dictator game with three actions

on *maximization of the sum of individual utilities* [70] and Rawls' criterion of fairness based on *maximization of the minimum of individual utilities* [116] (so-called *maximin* criterion), would prescribe that option B should be preferred to options A and C.

According to existing theories of team reasoning, option C will never be chosen by the row player. This is clearly counterintuitive, as the row player could be concerned, at the same time, with individual welfare and with social welfare, thereby preferring option C over options A and C. Indeed, option C can be seen as a fair compromise between promoting individual welfare ('I-mode') and promoting social welfare ('we-mode').

To overcome this limitation, we have recently proposed model of social ties that integrates the concept of *partial identification* with a group or team [**my papers: A9,A10**]. The model assumes that an agent will balance the goal of maximizing individual welfare with the goal of maximizing social welfare, depending on how much the agent is tied with the group. The model has been experimentally validated in collaboration with experimental economists from Toulouse School of Economics (TSE) [**my papers: A1**].

This work on social ties has been included in Frédéric Moisan's PhD thesis that I have co-supervised with other colleagues from IRIT.

My research on collective-decision making has also been devoted to the development a modal logic of social preferences in which different concepts such as fairness, inequity aversion and reciprocity can be formally represented as well as solution concepts for games in normal form corresponding to these concepts [**my papers: A15**].[3] The interesting aspect of the logic is that it allows to characterize the epistemic conditions that are necessary and/or sufficient for the agents in a game to play what a certain solution concept prescribes, depending on their types which include the self-interested type, the fair type and the reciprocator type. This is tightly related with my work on epistemic game theory discussed in Section 3.2. An example of theorem we have proved in the logic is the following one:

> If all agents are reciprocators, every agent knows the choices of the others, every agent knows that every agent knows the choices of the others, then the agents will select a strategy profile which is either a social-welfare equilibrium or a Nash equilibrium.

[3] This work extends and improves over my previous work on the logical formalization of the concept of altruism [**my papers: B21**].

where social-welfare equilibrium is the collective counterpart of the concept of Nash equilibrium: a strategy profile s is a social-welfare equilibrium if and only if, for every agent i, the action s_i is *for the entire group of agents* a best response to the other agents' joint action s_{-i}.

Chapter 5

Perspectives

The concepts and logical theories illustrated in the previous chapters constitute the *past* and the *present* of my research. I plan to refine, improve and complete them in the coming years. For instance, there are number of open issues that I would like to address such as: (i) decidability and complexity results for the satisfiability problem of both temporal STIT and Ockhamist Propositional Dynamic Logic OPDL illustrated in Chapter 2, (ii) a complete axiomatization of OPDL, (iii) decidability and complexity results for the modal logic of trust-based belief change briefly illustrated in Section 3.1, (iv) the application of the minimal logic for interactive epistemology illustrated in Section 3.2 to other solution concepts.

In this last chapter I am going to discuss some new perspectives for future research that go beyond the concepts and logical theories illustrated in the previous chapters. I do not pretend to be exhaustive. My aim is just to briefly sketch some new ideas I have recently started to elaborate.

Resource-bounded reasoning As emphasized in Section 3.1, some beliefs originate through inference. In order to formally characterize inference-based beliefs, it is necessary to drop the assumption that agents are omniscient, in the sense that: (i) their beliefs are closed under conjuction or under known implication, i.e., if φ is believed and ψ is believed then $\varphi \wedge \psi$ is believed and if φ is believed and $\varphi \rightarrow \psi$ is believed then ψ is believed; (ii) their explicit beliefs are closed under logical consequence (*alias* valid implication), i.e., if φ is believed and φ logically implies ψ, i.e., $\varphi \rightarrow \psi$ is valid, then ψ is believed as well; (iii) they believe valid sentences or tautologies; (iv) they have introspection over their beliefs, i.e., if φ is believed then it is believed that φ is believed. As pointed out by [76, 87], relaxing the assumption of logical omniscience allows for a resource-bounded agent who might fail to draw any connection between φ and its logical consequence ψ and, consequently, to believe any valid sentence and who might need time to infer and form new beliefs from her existing knowledge and beliefs.

We have recently started to study inference-based beliefs by proposing a logic which supports reasoning about the formation of beliefs through inference and through perception in non-omniscient resource-bounded agents [**my papers: C3**]. The logic

distinguishes the concept of explicit belief from the concept of background knowledge. This distinction is reflected in its formal semantics and axiomatics: (i) we use a non-standard semantics putting together a neighbourhood semantics for explicit beliefs and relational semantics for background knowledge, and (ii) we have specific axioms in the logic highlighting the relationship between the two concepts. Mental operations of perceptive type and inferential type, having effects on epistemic states of agents, are primitives in the object language of the logic. At the semantic level, they are modelled as special kinds of model-update operations, in the style of dynamic epistemic logic (DEL). We have provided results about axiomatization, decidability and complexity for this logic.

My future work will be devoted to the application of this logic to epistemic game theory. Existing analysis of the epistemic foundations of solution concepts rely on the assumption that players are omniscient and perfect reasoners. For example, the characterization of the epistemic foundation for "iterated deletion of strongly dominated strategies" (IDSDS) in terms of common knowledge (or common belief) of weak rationality only works under this assumption: if we drop it, common belief of weak rationality in the sense of [151] does not imply anymore that the agents will play a strategy profile which survives IDSDS. The logic will help to answer questions such as these: under which conditions will non-omniscient and imperfect reasoners converge to the strategy profiles prescribed by IDSDS? which beliefs are required? which kind of inference? I believe that moving from omniscient players to non-omniscient ones is an interesting move, as it allows my work on epistemic game theory to get closer to the social reality that is populated by imperfect human reasoners.

Resource-bounded reasoning is also relevant for the concept of intention. Indeed, Bratman's theory of intention [26, 28], one of the most prominent philosophical theories of intention nowadays, emphasizes that forming future-directed intentions enable agents to extend the influence of their deliberations beyond the present moment and that this is important given the limited cognitive capacities and time for deliberation of human agents. Specifically, it may be the case that at the time t an agent will have less time to deliberate and think through the options, or she may be distracted. For example, I may decide now what to do during the weekend since I know that I will have a busy week at work and will have no time to make my plan for the weekend. Existing logical accounts of the concept of intention neglect this aspect, as they assume that an agent holding an intention is a perfect reasoner with unlimited reasoning capabilities and unlimited time for deliberation. My future work will be devoted to extend the logic of beliefs for resource-bounded agents, we recently proposed in [**my papers: C3**], with a concept of intention. The objective is to come up with a satisfactory logical account of Bratman's idea that future-directed intentions only make sense from the perspective of designing agents with limited reasoning capabilities and limited time for deliberation.[1]

I also plan to extend the conceptual apparatus of epistemic game theory with the concept of intention in order to model resource-bounded players who need to plan their

[1]Notice that future-directed intention are not justified by the uncertainty of the future. Indeed, if an agent has unlimited time for deliberation, she does not need to decide in advance what to do in the *future*, even if the future is uncertain. At any moment, she is able to make the optimal decision about what to do in the *present*.

future actions in advance in an extensive game setting, since they have limited cognitive capacities and limited time for deliberation. But this is more a long-term perspective.

"Mentalizing" logics of actions In the future I intend to fill an existing gap in my past and present research, namely the gap between my work on the logic of action, illustrated in Chapter 2, and my work on the logic of mental attitudes, illustrated in Chapter 3. The expected plan to fill this gap will consist in "mentalizing" the temporal STIT logic presented in Chapter 2. First of all, I plan to enrich temporal STIT with an explicit representation of both the agents' beliefs and the agents' preferences over histories. Secondly, I plan to provide a sophisticated account of the connection between beliefs, preferences and choices, the latter being primitive elements in the STIT semantics. This connection will be based on the the rationality criteria studied in epistemic game theory and illustrated in Section 3.2. More generally, it is based on the idea that, in a multi-agent setting, an agent's choice is determined by the agent's preferences over the possible courses of action and by her beliefs about the choices of the others.

I plan to compare the Kripke semantics of this "mental" extension of temporal STIT with semantics that are traditionally used in the area of epistemic game theory to represent beliefs of players in extensive form games and that are based on the concept of type space (see [109] for a general introduction to these semantics). This work will be a natural evolution of my past research, that I have briefly illustrated in Section 3.2, on the comparison between Kripke semantics and type space semantics for a variety of epistemic logical languages.

The possibility of "mentalizing" temporal STIT will also allow me to greatly improve the work on social influence I have presented in Section 2.3 by:

- capturing the subtle relationship between influence on beliefs and influence on actions, namely the fact that an agent can induce another agent to behave in a certain way by changing his beliefs, since beliefs are the input of decision-making and provide reasons for deciding and for acting,
- distinguishing mere influence from the more specific concept of manipulation, in the sense of influencing someone to do something by lying and by providing false information.

Influence-based opinion diffusion As emphasized in Section 4.1, collective attitudes such as collective acceptance, shared belief and common belief can originate in different ways. I have mentioned aggregation of individual attitudes and communication as two possible ways of forming collective attitudes.

In recent times, I have been interested in the formation of collective attitudes through influence-based opinion diffusion.

Influence is the process which consists in forming an opinion about a given issue or proposition on the basis of the opinions expressed by the other individuals in the society. It ensures that opinions can spread in a society. Specifically, by expressing her opinions, an agent affects the opinions of other agents who in turn will affect the opinions of other agents, and so on. This generates a propagation of individual opinions in the society (a sort of cascade) that can lead to the formation of a shared belief (or

consensus) about a given issue. This kind of influence process is mediated by trust, as for an agent to be affected by the expressed opinions of another agent, she has to trust his judgment.

Recently, I have got interested in the epistemic, dynamic and strategic aspects of influence-based opinion diffusion. My plan is to develop a number of logics and formal theories that clarify:

- how the influence process works by changing beliefs on the basis of the expressed opinions of other agents in the society;
- how certain structural properties of the trust relations (e.g., whether the trust network induced by the trust relations between the agents is acyclic or has hubs) guarantee convergence of opinions or stabilization to consensus in the long run;
- how disclosure of opinions has a strategic component that affects stabilization to consensus in the long run either by promoting it or by hindering it.

Some preliminary results in this direction are presented in [**my papers: C4,C8**] in which methods and techniques from judgment aggregation [90, 62] are applied to the analysis of opinion diffusion and stabilization to consensus in social networks. I plan to improve this work by exploiting methods and techniques from dynamic epistemic logic (DEL) in order to enrich the representation of the agents' opinions and, in particular, to represent high-order opinions, i.e., opinions about other agents' opinions.

Also, less recently, I have focused on the problem of representing information sources explicitly in epistemic logic [**my papers: C27**]. This allows to keep track of the evidences obtained from the social environment that are responsible for influence and, consequently, for belief formation. I plan to pursue further this line of research in the coming years.

Grounding of mental attitudes Most of existing logics of mental attitudes and their evolution are "ungrounded" in the sense that their formal semantics are based on abstract notions such as the concept of Kripke model or the notions of envisaged world and possible world, as illustrated in Section 3.1, which are not built from the agents' primitive mental states and perceptual data. The grounding problem becomes particularly relevant for AI practitioners, when moving from theory to practice. When designing an artificial intelligent agent, we normally start from an explicit representation of the agent's mental states including her perceptual data base (i.e., all facts that the agent can see), her belief base (i.e., all facts that the agent believes) and her goal base (i.e., all states of affairs that the agent wants to achieve). From this perspective, the concepts of undistinguishability and possible world only make sense if they are built on the top of them and are grounded on them.

I have recently started to work on grounded semantics for logics of mental attitudes. Specifically, in [**my papers: B4,C2,C6**] we have developed a logic of visibility-based knowledge, which generalizes previous work by [155, 156]. This logic is a simple variant of epistemic logic in which knowledge of agents is grounded on the concept of visibility: an agent knows that a certain fact p is true if and only if p is true and the truth value of p is "visible" to her. Its interesting aspect is that it supports reasoning about

high-order visibility, namely the fact that an agent 1 sees that another agent 2 sees the truth value of p or that an agent 1 sees that another agent 2 sees that another agent 3 sees the truth value of p, and so on. An agent's high-order knowledge about what the other agents know depends on the agent's high-order visibility of what the other agents can see. For example, an agent 1 knows that another agent 2 knows that a certain fact p is true if and only if agent 1 knows that p is true and can see that agent 2 can see the truth value of p. In [**my papers: C1**], we have applied the logic to the analysis of strategic contexts in which agents can have epistemic goals, change what other agents can see and act in order to change the epistemic states of other agents.

I plan to pursue further this line of research by integrating spatial aspects in the logics of mental attitudes I have considered so far. This choice is justified by the objective of implementing such logics in robots interacting in a physical environment. In order to use such logics for robotic implementations, their semantics have to be grounded on space. Specifically, a logic is required that explains how epistemic states of the robots are determined from what the robots perceive from the physical environment. To this aim, existing epistemic logics and dynamic epistemic logics need to incorporate the concept of perception, and to provide an explicit representation of the space in which the robots' actions and perceptions are situated.

Conventions The last concept I want to discuss is convention, a concept that has been widely studied in economics [143], philosophy [19, 148] and computer science [164, 161, 133, 130], given the fundamental role it plays in the regulation of both human and artificial societies.

Eating manners, the kind of clothes we wear in office, and the side of the road on which we drive are mundane examples of convention. Roughly, a social convention is a customary, arbitrary and self-enforcing rule of behavior that is generally followed and expected to be followed in a group or in a society at large [88]. When a social convention is established, everybody behaves in an agreed-upon way even if they did not in fact explicitly agree to behave in this way. A social convention can thus be seen as a kind of tacit agreement that has evolved out of a history of previous interactions [143, 148].

Since the seminal contribution by David Lewis [88], the modern approach to conventions is rooted both in epistemic logic and in evolutionary game theory. The *epistemic approach* to the study of conventions has focused on the characterization of the kind of mutual beliefs and expectations that are required for a group to adopt a certain convention [40, 134, 160] and on the distinction between the epistemic conditions of conventions in contrast with the epistemic conditions of social norms [17]. For instance, according to the well-known definition of convention by David Lewis [88, pp. 76], a given regularity of behavior R is a convention for a population of agents P at a recurrent situation S, only if the agents in the population P *mutually expect* everyone in P to conform to the regularity R in the situation S (and commonly believe so). The *evolutionary approach* to the study of conventions has focused on the conditions under which a certain convention can emerge on a given population of agents depending on the agents' learning capabilities. Notable examples of this approach are the models by Kandori et al. [83] and Young [169] which make predictions about the conditions

under which agents converge to equilibrium in a certain coordination game by learning the others' play and adjusting their strategies over time. For instance, Kandori et al.'s model investigates the dynamic process that leads the agents to converge to the risk dominant equilibrium in a repeated 2×2 coordination game.

I have always been interested in the topic of convention since the time of my PhD. I believe it is a fascinating topic at the intersection of the three axes of my research: actions, mind and collectives. I believe I have now acquired the conceptual and technical skills to clearly understand it and to be able to address it formally.

One of the challenges of my future research is to develop a formal model of convention which is able to reconcile the epistemic approach and the evolutionary approach to the study of conventions. Indeed, none of the existing evolutionary models of conventions deals with the epistemic aspect of conventions, as they do not assume agents to be cognitive and only consider a simplified notion of convention as a mere regularity of behavior. I would like my model of conventions to be able to clarify how the agents' reasoning and learning capabilities as well as the structure of the environment (e.g., who can see whom, whether actions of agents are public or private) and the communication network (e.g., who can communicate with whom) can either promote or hinder the formation of a system of mutual expectations about each other's behavior in a certain population of cognitive agents. As emphasized above, such a system of mutual expectations provides the epistemic foundation of the Lewisian concept of convention. Thus, the main innovation of my research project on convention will be to go beyond existing models of the emergence of conventions that completely neglect the epistemic aspect of the concept of convention, which is of paramount importance according to the Lewisian tradition.

At the technical level, the development of this model will require the combination of methods and techniques from epistemic logic and dynamic epistemic logic (DEL) with methods and techniques from the areas of evolutionary game theory and the theory of learning games [53, 165]. I am not new to them. In a recent work [**my papers: C5**], we have used them to develop a game-theoretic model of guilt and the evolution of fairness norms.

Relevance for the local research community Before concluding this document, I would like to emphasize that the perspectives for my future research perfectly fit the objectives of the LILaC (Logique, Interaction, Langue, et Calcul) team at the Institut de Recherche en Informatique de Toulouse (IRIT), the research team in which I have spent the last ten years of my professional activity and in which I plan to spend the coming years. My research project is at the core of the four axes of LILaC, namely logic, interaction, language and computation. The combination of logic and game theory, the two methodologies of my research, has opened new research avenues that are currently explored at the team level and that, I believe, will lead to interesting results in the future. Moreover, each perspective is expected to lead to fruiful collaborations with the other members of the LILaC team. Perspective "Resource-bounded reasoning" is currently explored in collaboration with with Philippe Balbiani and David Fernandez-Duque, whereas the work on intention will be a follow-up of my joint work with Andreas Herzig and Laurent Perrussel. I am currently co-supervising with them the PhD

thesis of Zhanhao Xiao on the topic of intention. Perspectives "Mentalizing logics of actions" and "Grounding of mental attitudes" will directly involve Philippe Balbiani, Olivier Gasquet and Andreas Herzig given their interest in the grounding problem and their expertise in logical modelling of mental attitudes and in logics of space. I am currently co-supervising with Andreas Herzig the PhD thesis of Faustine Maffre on the logic of visibility-based knowledge. Perspective "Influence-based opinion diffusion" is currently explored in collaboration with Umberto Grandi and Laurent Perrussel. We plan to involve Dominique Longin in the work. Indeed, there is an interesting connection between our work on influence-based opinion and Dominique's previous work on emotional contagion and emotion dynamics at the population level. Andreas Herzig and Dominique Longin will also be directly involved in the perspective "Convention", given their interest in the logic of collective attitudes and their dynamics.

I am currently involved in a project coordinated by Astrid Hopfensitz, an experimental economist from Toulouse School of Economics (TSE). I have collaborated with her in the past on collective decision-making (cf. Section 4.2) and we have co-supervised together with Andreas Herzig the PhD thesis of Frédéric Moisan. The project is on the general topic of social intelligence. Given the relevance of the concept of convention for the project, I plan to collaborate with Astrid on the empirical validation of the formal model of convention that I intend to develop during the next few years.

Bibliography

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

[2] C. Alchourron, P. Gardenfors, and D. Makinson. On the logic of theory change: Partial meet contraction and revision functions. *The Journal of Symbolic Logic*, 50:510–530, 1985.

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

[4] G. E. M. Anscombe. *Intention*. Basil Blackwell, 1957.

[5] R. Aumann. Interactive epistemology I: Knowledge. *International Journal of Game Theory*, 28(3):263–300, 1999.

[6] R. J. Aumann. Backward induction and common knowledge of rationality. *Games and Economic Behavior*, 8:6–19, 1995.

[7] M. Bacharach. Interactive team reasoning: a contribution to the theory of cooperation. *Research in economics*, 23:117–147, 1999.

[8] M. Bacharach. *Beyond individual choice: teams and frames in game theory*. Princeton University Press, Oxford, 2006.

[9] P. Balbiani, A. Herzig, and N. Troquard. Dynamic logic of propositional assignments: A well-behaved variant of pdl. In *Proceedings of the 2013 28th Annual ACM/IEEE Symposium on Logic in Computer Science (LICS 2013)*, pages 143–152. Morgan Kaufmann, 2013.

[10] A. Baltag and L. S. Moss. Logics for epistemic programs. *Synthese*, 139 (2):165–224, 2004.

[11] A. Baltag and S. Smets. A qualitative theory of dynamic interactive belief revision. In *Proceedings of LOFT 7*, volume 3 of *Texts in Logic and Games*, pages 13–60. Amsterdam University Press, 2008.

[12] A. Baltag, S. Smets, and J. A. Zvesper. Keep hoping for rationality: a solution to the backward induction paradox. *Synthese*, 169(2):301–333, 2009.

[13] P. Battigalli and M. Dufwenberg. Guilt in games. *The American Economic Review*, 97(2):170–176, 2007.

[14] P. Battigalli and M. Siniscalchi. Strong belief and forward induction reasoning. *J. of Economic Theory*, 106(2):356–391, 2002.

[15] R. F. Baumeister, A. M. Stillwell, and T. F. Heatherton. Guilt: an interpersonal approach. *Psychological Bullettin*, 115(2):243–267, 1994.

[16] N. Belnap, M. Perloff, and M. Xu. *Facing the future: agents and choices in our indeterminist world*. Oxford University Press, New York, 2001.

[17] C. Bicchieri. *The grammar of society: the nature and dynamics of social norms*. Cambridge University Press, 2006.

[18] K. Binmore. *Fun and Games: A Text on Game Theory*. D. C. Heath and Company, 1991.

[19] K. Binmore. *Natural Justice*. Oxford University Press, 2005.

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

[21] E. Bonzon, M.-C. Lagasquie-Schiex, J. Lang, and B. Zanuttini. Boolean games revisited. In G. Brewka, S. Coradeschi, A. Perini, and P. Traverso, editors, *Proceedings of the 17th European Conference on Artificial Intelligence*, pages 265–269, 2006.

[22] C. Boutilier. Towards a logic for qualitative decision theory. In *Proceedings of International Conference on Principles of Knowledge Representation and Reasoning (KR'94)*, pages 75–86. AAAI Press, 1994.

[23] R. I. Brafman and Moshe Tennenholtz. An axiomatic treatment of three qualitative decision criteria. *Journal of the ACM*, 47(3):452–482.

[24] R. I. Brafman and Moshe Tennenholtz. On the foundations of qualitative decision theory. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI'96)*, pages 1291–1296. AAAI Press, 1996.

[25] A. Brandenburger, A. Friedenberg, and J. Keisler. Admissibility in games. *Econometrica*, 76:307–352, 2008.

[26] M. Bratman. *Intentions, plans, and practical reason*. Harvard University Press, Cambridge, 1987.

[27] M. Bratman. Shared cooperative activity. *The Philosophical Review*, 101(2):327–41, 1992.

[28] M. Bratman. Intention, belief, and instrumental rationality. In D. Sobel and S. Wall, editors, *Reasons for action*, pages 13–36. Cambridge University Press, 2009.

[29] M. Bratman, D. J. Israel, and M. E. Pollack. Plans and resource-bounded practical reasoning. *Computational Intelligence*, 4:349–355, 1988.

[30] J. Broersen. Deontic epistemic stit logic distinguishing modes of mens rea. *Journal of Applied Logic*, 9(2):137–152, 2011.

[31] A. Caplin and J Leahy. Psycological expected utility theory and anticipatory feelings. *Quarterly Journal of Economics*, 116(1):55–79, 2001.

[32] C. Castelfranchi. Modelling social action for ai agents. *Artificial Intelligence*, 103:157–182, 1998.

[33] C. Castelfranchi and R. Falcone. *Trust Theory: A Socio-Cognitive and Computational Model*. Wiley, 2010.

[34] G. Charness and M. Dufwenberg. Guilt in games. *Econometrica*, 74(6):1579–1601, 2009.

[35] B. F. Chellas. Time and modality in the logic of agency. *Studia Logica*, 51(3-4):485–518, 1992.

[36] H. Clark and C. Marshall. Definite reference and mutual knowledge. In A. K. Joshi, B. L. Webber, and I. A. Sag, editors, *Elements of discourse understanding*. 1981.

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

[38] A. M. Colman, B. N. Pulford, and J. Rose. Collective rationality in interactive decisions: evidence for team reasoning. *Acta Psychologica*, 128:387–397, 2008.

[39] R. Cooper, D. De Jong, R. Forsythe, and T. Ross. Forward induction in the battle-of-the-sexes games. *The American Economic Review*, 83(5):1303–1316, 1993.

[40] R. P. Cubitt and R. Sugden. Common knowledge, salience and convention: a reconstruction of david lewis' game theory. *Economics and Philosophy*, 19:175–210, 2003.

[41] A. Damasio. *Descartes Error: Emotion, Reason and the Human Brain*. Putnam Publishing, New York, 1994.

[42] D. Davidson. Intending. In *Essays on Actions and Events*. Oxford University Press, New York, 1980.

[43] D. C. Dennett. *The Intentional Stance*. MIT Press, Cambridge, Massachusetts, 1987.

[44] B. Dunin-Keplicz and R. Verbrugge. Collective intentions. *Fundamenta Informaticae*, 51(3):271–295, 2002.

[45] B. Dunin-Keplicz and R. Verbrugge. *Teamwork in Multi-Agent Systems: A Formal Approach*. Wiley, 2010.

[46] E. Durkheim. *The rules of Sociological Method*. Free Press, New York, 1982. first published in French in 1895.

[47] M. Esteva, J. Padget, and C. Sierra. Formalizing a language for institutions and norms. In *Intelligent Agents VIII (ATAL'01)*, volume 2333 of *LNAI*, pages 348–366, Berlin, 2001. Springer Verlag.

[48] R. Fagin, J. Halpern, Y. Moses, and M. Vardi. *Reasoning about Knowledge*. MIT Press, Cambridge, 1995.

[49] E. Fehr and K. M. Schmidt. A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics*, 114:817–868, 1999.

[50] E. Fehr and K. M. Schmidt. Theories of fairness and reciprocity: Evidence and economic applications. In *Advances in Economics and Econometrics*. Cambridge University Press, 2003.

[51] R. Forsythe, J. L. Horowitz, N. E. Savin, and M. Sefton. Fairness in simple bargaining experiments. *Games and Economic Behavior*, 6:347–369, 1994.

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

[53] D. Fudenberg and D. K. Levine. *The theory of learning in games*. MIT Press, 1998.

[54] D. M. Gabbay, I. M. Hodkinson, and M. A. Reynolds. *Temporal Logic: Mathematical Foundations and Computational Aspects*, volume 1. Clarendon Press, Oxford, 1994.

[55] Y. Gal, A. Pfeffer, F. Marzo, and B. J. Grosz. Learning social preferences in games. In *Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI'04)*, pages 226–231. AAAI Press, 2004.

[56] M. Gilbert. Modelling collective belief. *Synthese*, 73(1):185–204, 1987.

[57] M. Gilbert. *On Social Facts*. Routledge, London and New York, 1989.

[58] H. Gintis. *The bounds of reason: game theory and the unification of the behavioral sciences*. Princeton University Press, Cambridge, 2009.

[59] R. Goldblatt. *Logics of Time and Computation, 2nd edition*. CSI Lecture Notes, Stanford, California, 1992.

[60] M. Goldszmidt and J. Pearl. Qualitative probability for default reasoning, belief revision and causal modeling. *Artificial Intelligence*, 84:52–112, 1996.

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

[62] U. Grandi and U. Endriss. Lifting integrity constraints in binary aggregation. *Artificial Intelligence*, 199-200:45–66, 2013.

[63] J. Gratch and S. Marsella. A domain independent framework for modeling emotion. *Journal of Cognitive Systems Research*, 5(4):269–306, 2004.

[64] B. Grosz and S. Kraus. Collaborative plans for complex group action. *Artificial Intelligence*, 86(2):269–357, 1996.

[65] R. Hakli, K. Miller, and R. Tuomela. Two Kinds of We-Reasoning. *Economics and Philosophy*, 26:291–320, 2010.

[66] J. Y. Halpern. Beyond nash equilibrium: Solution concepts for the 21st century. In K. R. Apt and E. Gradel, editors, *Lectures in Game Theory for Computer Scientists*, pages 264–289. 2011.

[67] J. Y. Halpern and R. Pass. A logical characterization of iterated admissibility. In A. Heifetz, editor, *Proc. of TARK 2009*, pages 146–155, 2009.

[68] D. Harel, D. Kozen, and J. Tiuryn. *Dynamic Logic*. MIT Press, Cambridge, 2000.

[69] P. Harrenstein, W. van der Hoek, J.-J. Meyer, and C. Witteveen. Boolean games. In *Proceedings of the 8th Conference on Theoretical Aspects of Rationality and Knowledge (TARK)*, pages 287–298. Morgan Kaufmann Publishers Inc., 2001.

[70] J. Harsanyi. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy*, 63:309–321, 1955.

[71] J. Harsanyi. Morality and the theory of rational behaviour. In A.K. Sen and B. Williams, editors, *Utilitarianism and Beyond*. Cambridge University Press, Cambridge, 1982.

[72] J. C. Harsanyi. Games with incomplete information played by 'bayesian' players. *Management Science*, 14:159–182, 1967.

[73] H. L. A. Hart. *The concept of law*. Clarendon Press, Oxford, 1992. new edition.

[74] A. Herzig and F. Schwarzentruber. Properties of logics of individual and group agency. *Advances in modal logic*, 7:133–149, 2008.

[75] Andreas Herzig and Dominique Longin. C&L intention revisited. In Didier Dubois, Chris Welty, and Mary-Anne Williams, editors, *Proceedings 9th Int. Conf. on Principles on Principles of Knowledge Representation and Reasoning(KR2004)*, pages 527–535. AAAI Press, 2004.

[76] J. Hintikka. Impossible possible worlds vindicated. *Journal of Philosophical Logic*, 4:475–484, 1975.

[77] A. Hopfensitz and E. Reuben. The importance of emotions for the effectiveness of social punishment. *The Economic Journal*, 119(540):1534–1559, 2009.

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

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

[80] I. L. Humberstone. Direction of fit. *Mind*, 101(401):59–83, 1992.

[81] Doyle J. and Thomason R. Background to qualitative decision theory. *The AI Magazine*, 20(2):55–68, 1999.

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

[83] M. Kandori, G. Mailath, and R. Rob. Learning, mutation, and long run equilibria in games. *Econometrica*, 61:29–56, 1993.

[84] K. Konolige and M. E. Pollack. A representationalist theory of intention. In R. Bajcsy, editor, *Proceedings 13th International Joint Conference on Artificial Intelligence (IJCAI 93)*, pages 390–395, San Francisco, CA, 1993. Morgan Kaufmann Publishers.

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

[86] J. LeDoux. *The emotional Brain*. Simon and Schuster, New York, 1996.

[87] H. Levesque. A logic of implicit and explicit belief. In *Proceedings of AAAI-84*, pages 198–202. AAAI Press, 1984.

[88] D. K. Lewis. *Convention: a philosophical study*. Harvard University Press, Cambridge, 1969.

[89] C. J. Liau. Belief, information acquisition, and trust in multi-agent systems: a modal logic formulation. *Artificial Intelligence*, 149:31–60, 2003.

[90] C. List. The theory of judgment aggregation: an introductory review. *Synthese*, 187:179–207, 2012.

[91] C. List. Three kinds of collective attitudes. *Erkenntnis*, 79(9):1601–1622, 2014.

[92] C. List and P. Pettit. *Group Agency: The Possibility, Design, and Status of Corporate Agents*. Oxford University Press, 2011.

[93] G. F. Loewenstein, L. Thompson, and M. H. Bazerman. Social utility and decision making in interpersonal contexts. *Journal of Personality and Social Psychology*, 57:426–441, 1989.

[94] G. Loomes and R. Sugden. Testing for regret and disappointment in choice under uncertainty. *Economic J.*, 97:118–129, 1987.

[95] F. Lopez y Lopez, M. Luck, and M. d'Inverno. Normative agent reasoning in dynamic societies. In *Proceedings of the Third International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS'04)*, pages 732–739. ACM Press, 2004.

[96] K. Ludwig and M. Jankovic. Collective intentionality. In L. McIntyre and A. Rosenberg, editors, *The Routledge Companion to the Philosophy of Social Science*. Routledge, New York, 2016.

[97] C. Mantzavinos, D.C. North, and S. Shariq. Learning, institutions, and economic performance. *Perspectives on Politics*, 2:75–84, 2004.

[98] H. Margolis. *Selfishness, Altruism, and Rationality: A Theory of Social Choice*. University of Chicago Press, Chicago, 1982.

[99] H. McCann. Settled objectives and rational constraints. *American Philosophical Quarterly*, 28:25–36, 1991.

[100] A. R. Mele. *Springs of Action: Understanding Intentional Behavior*. Oxford University Press, Oxford, 1992.

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

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

[103] K. Miller and G. Sandu. Weak commitments. In G. Holmstron-Hintikka and R. Tuomela, editors, *Contemporary Action Theory, vol.2: Social Action*. Kluwer Academic Publishers, Dordrecht, 1997.

[104] R. Myerson. *Game Theory: Analysis of Conflict*. Harvard University Press, 1991.

[105] D.C. North. *Institutions, Institutional Change, and Economic Performance*. Cambridge University Press, Cambridge, 1990.

[106] Andrew Ortony, G.L. Clore, and A. Collins. *The cognitive structure of emotions*. Cambridge University Press, Cambridge, MA, 1988.

[107] M. J. Osborne and A. Rubinstein. *A course in game theory*. MIT Press, 1994.

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

[109] A. Perea. *Epistemic game theory: reasoning and choice*. Cambridge University Press, 2012.

[110] M. Platts. *Ways of meaning*. Routledge and Kegan Paul, 1979.

[111] J. A. Plaza. Logics of public communications. In M. Emrich, M. Pfeifer, M. Hadzikadic, and Z. Ras, editors, *Proceedings of the 4th International Symposium on Methodologies for Intelligent Systems*, 201-216, 1989.

[112] A. Prior. *Past, Present, and Future*. Clarendon Press, Oxford, 1967.

[113] Z. Pylyshyn. *Computation and Cognition: Toward a Foundation for Cognitive Science*. MIT Press, Cambridge, Massachusetts, 1984.

[114] M. Rabin. Incorporating fairness into game theory and economics. *American Econonic Review*, 83(5):1281–1302, 1993.

[115] A. S. Rao and M. P. Georgeff. Modelling rational agents within a BDI-architecture. In *Proceedings of KR'91*, pages 473–484, San Mateo, CA, 1991. Morgan Kaufmann Publishers.

[116] J. Rawls. *A theory of Justice*. Harvard University Press, Cambridge, 1971.

[117] D. Regan. *Utilitarianism and cooperation*. Clarendon Press, Oxford, 1980.

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

[119] M. Reynolds. An axiomatization of full computation tree logic. *Journal of Symbolic Logic*, 66(3):1011–1057, 2001.

[120] M. A. Reynolds. An axiomatization of prior's ockhamist logic of historical necessity. In *Advances in Modal Logic*, volume 4, pages 355–370. King's College Publications, 2003.

[121] S. Rick and G. Loewenstein. The role of emotion in economic behavior. In M. Lewis, J. Haviland-Jones, and L. Feldman-Barrett, editors, *The Handbook of Emotion*. Guilford, New York, 2008.

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

[123] D. Sally. Conversation and cooperation in social dilemmas: a meta-analysis of experiments from 1958 to 1992. *Rationality and Society*, 7:58–92, 1995.

[124] K. R. Scherer, A. Schorr, and T. Johnstone, editors. *Appraisal Processes in Emotion: Theory, Methods, Research*. Oxford University Press, Oxford, 2001.

[125] F. Schwarzentruber. Complexity results of STIT fragments. *Studia Logica*, 100(5):1001–1045, 2012.

[126] J. Searle. *Expression and meaning*. Cambridge University Press, 1979.

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

[128] J. Searle. *The Construction of Social Reality*. The Free Press, New York, 1995.

[129] J. Searle. *Rationality in Action*. MIT Press, Cambridge, 2001.

[130] S. Sen and S. Airiau. Emergence of norms through social learning. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI 2007)*, pages 1507–1512. ACM Press, 2007.

[131] Y. Shoham. Agent-oriented programming. *Artificial Intelligence*, 60:51–92, 1993.

[132] Y. Shoham and K. Leyton-Brown. *Multiagent Systems: Algorithmic, Game Theoretic and Logical Foundations*. Cambridge University Press, 2009.

[133] Y. Shoham and M. Tennenholtz. On the emergence of social conventions: modeling, analysis, and simulations. *Artificial Intelligence*, 94(1-2):139–166, 1997.

[134] G. Sillari. A logical framework for convention. *Synthese*, 147(2):379–400, 2005.

[135] H. A. Simon. Rational choice and the structure of the environment. *Psychological Review*, 63(2):129–138, 1956.

[136] M. Singh and N. Asher. A logic of intentions and beliefs. *Journal of Philosophical Logic*, 22:513–544, 1993.

[137] W. Spohn. Ordinal conditional functions: a dynamic theory of epistemic states. In *Causation in decision, belief change and statistics*, pages 105–134. Kluwer, 1998.

[138] R. Stalnaker. Belief revision in games: forward and backward induction. *Mathematical Social Sciences*, 36:31–56, 1998.

[139] R. Stalnaker. Common ground. *Linguistics and Philosophy*, 25(5-6):701–721, 2002.

[140] R. Stalnaker. On logics of knowledge and belief. *Philosophical Studies*, 128:169–199, 2006.

[141] R. Sugden. Team preferences. *Economics and Philosophy*, 16:175–204, 2000.

[142] R. Sugden. The logic of team reasoning. *Philosophical Explorations*, 6(3):165–181, 2003.

[143] R. Sugden. *Economics of rights, co-operation and welfare (2nd Edition)*. Palgrave Macmillan, 2004.

[144] J. P. Tangney. Recent advances in the empirical study of shame and guilt. *American Behavioral Scientist*, 38(8):1132–1145, 1995.

[145] R. Thomason. Combinations of tense and modality. In D. Gabbay and F. Guenthner, editors, *Handbook of Philosophical Logic*, volume 2, pages 135–165. Reidel, Dordrecht, 1984. 2nd edition.

[146] M. L. Tiomkin and J. A. Makowsky. Propositional dynamic logic with local assignments. *Theoretical Computer Science*, 36(71-87):71–87, 1985.

[147] D. P. Tollefsen. Collective intentionality and the social sciences. *Philosophy of the Social Sciences*, 32(1):25–50, 2002.

[148] L. Tummolini, G. Andrighetto, C. Castelfranchi, and R. Conte. A convention or (tacit) agreement betwixt us: on reliance and its normative consequences. *Synthese*, 190(4):585–618, 2013.

[149] R. Tuomela. *The Philosophy of Social Practices: A Collective Acceptance View*. Cambridge University Press, Cambridge, 2002.

[150] R. Tuomela. *The Philosophy of Sociality*. Oxford University Press, Oxford, 2007.

[151] J. van Benthem. Rational dynamics and epistemic logic in games. *International Game Theory Review*, 9(1):13–45, 2007.

[152] J. van Benthem, P. Girard, and O. Roy. Everything else being equal: A modal logic for ceteris paribus preferences. *Journal of Philosophical Logic*, 38:83–125, 2009.

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

[154] J. van Benthem, J. van Eijck, and B. Kooi. Logics of communication and change. *Information and Computation*, 204(11):1620–1662, 2006.

[155] W. van der Hoek, P. Iliev, and M. Wooldridge. A logic of revelation and concealment. In Wiebe van der Hoek, Lin Padgham, Vincent Conitzer, and Michael Winikoff, editors, *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 1115–1122. IFAAMAS, 2012.

[156] W. van der Hoek, N. Troquard, and M. Wooldridge. Knowledge and control. In Liz Sonenberg, Peter Stone, Kagan Tumer, and Pinar Yolum, editors, *Proceedings of the 10th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 719–726. IFAAMAS, 2011.

[157] W. van der Hoek and M. Wooldridge. On the logic of cooperation and propositional control. *Artificial Intelligence*, 164(1-2):81–119, May 2005.

[158] H. van Ditmarsch, W. van der Hoek, and B. Kooi. *Dynamic Epistemic Logic*. Springer Verlag, 2007.

[159] B. van Linder, van der Hoek, and J.-J. Ch. W., Meyer. Formalising abilities and opportunities. *Fundamenta Informaticae*, 34:53–101, 1998.

[160] P. Vanderschraaf. Convention as correlated equilibrium. *Erkenntnis*, 42(1):65–87, 1995.

[161] D. Villatoro, S. Sen, and J. Sabater-Mir. Exploring the dimensions of convention emergence in multiagent systems. *Advances in Complex Systems*, 14(2):201–227, 2011.

[162] F. von Kutschera. T x W completeness. *Journal of Philosophical Logic*, 26(3):241–250, 1997.

[163] G. H. Von Wright. *The logic of preference*. Edinburgh University Press, 1963.

[164] W. Walker and M. Wooldridge. Understanding the emergence of conventions in multi-agent systems. In *Proceedings of the First International Conference on Multi-Agent Systems (ICMAS-95)*, pages 384–389. AAAI Press, 1995.

[165] J. W. Weibull. *Evolutionary Game Theory*. MIT Press, 1995.

[166] S. Wölf. Propositional Q-logic. *Journal of Philosophical Logic*, 31:387–414, 2002.

[167] M. Wooldridge. *Reasoning about rational agents*. MIT Press, Cambridge, 2000.

[168] M. Xu. Axioms for deliberative STIT. *Journal of Philosophical Logic*, 27:505–552, 1998.

[169] H. P. Young. The evolution of conventions. *Econometrica*, 61:57–84, 1993.

[170] A. Zanardo. Branching-time logic with quantification over branches: The point of view of modal logic. *Journal of Symbolic Logic*, 61(1):143–166, 1996.

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

Appendix A

My publications

(A) **Journal papers**

1. Attanasi, G., Hopfensitz, A., Lorini, E., Moisan, F. (forthcoming). Social connectedness improves co-ordination on individually costly, efficient outcomes. *European Economic Review*.[1]
2. Lorini, E., Sartor, G. (forthcoming). A STIT logic for reasoning about social influence. *Studia Logica*.[2]
3. Galeazzi, P., Lorini, E. (forthcoming). Epistemic logic meets epistemic game theory: a comparison between multi-agent Kripke models and type spaces. *Synthese*.[3]
4. Bolander, T., van Ditmarsch, H., Herzig, A., Lorini, E., Pardo, P., Schwarzentruber, F. (2016). Announcements to Attentive Agents. *Journal of Logic, Language and Information*, 25(1), pp. 1-35.
5. Lorini, E. (2016). A minimal logic for interactive epistemology. *Synthese*, 193(3), pp. 725-755.
6. Lorini, E. (2016). A logic for reasoning about moral agents. *Logique & Analyse*, 58(230).
7. Grossi, D., Lorini, E., Schwarzentruber, F. (2015). The Ceteris Paribus Structure of Logics of Game Forms. *Journal of Artificial Intelligence Research*, 53, 91-126.
8. Lorini, E., Longin, D., Mayor, E. (2014). A logical analysis of responsibility attribution: emotions, individuals and collectives. *Journal of Logic and Computation*, 24(6), pp. 1313-1339.
9. Hopfensitz, A., Lorini, E., Moisan, F. (2014). Conflicting goals and their impact on games where payoffs are more or less ambiguous . Commentary on "Mapping Collective Behavior in the Big-data Era". *Behavioral and Brain Sciences*, 37(1), pp. 85-87.
10. Attanasi, G., Hopfensitz, A., Lorini, E., Moisan, F. (2014). The Effects of Social Ties on Coordination: Conceptual Foundations for an Empirical Analysis. *Phenomenology and the Cognitive Sciences*, 13(1), pp. 47-73.
11. Lorini, E. (2013). Temporal STIT logic and its application to normative reasoning. *Journal of Applied Non-Classical Logics*, 23(4), pp. 372-399.

[1] Pre-publication online version available at
http://www.sciencedirect.com/science/article/pii/S0014292116300186

[2] Pre-publication online version available at
http://link.springer.com/article/10.1007/s11225-015-9636-x

[3] Pre-publication online version available at
http://link.springer.com/article/10.1007/s11229-015-0834-x

12. Lorini, E. (2013). On the epistemic foundation for iterated weak dominance: an analysis in a logic of individual and collective attitudes. *Journal of Philosophical Logic*. 42(6), pp. 863-904.
13. Reisenzein, R., Hudlicka, E., Dastani, M., Gratch, J., Hindriks, K., Lorini, E., Meyer, J.-J. (2013). Computational Modeling of Emotion: Towards Improving the Inter- and Intradisciplinary Exchange. *IEEE Transactions on Affective Computing*, 4(3), pp. 246-266.
14. Lorini, E., Troquard, N., Herzig, A., Broersen, J. (2012). Grounding power on actions and mental attitudes. *Logic Journal of the IGPL*, 21 (3), pp. 311-331.
15. Lorini, E. (2011). From self-regarding to other-regarding agents in strategic games: a logical analysis. *Journal of Applied Non-Classical Logics*, 21 (3-4), pp. 443-476.
16. Lorini, E., Moisan, F. (2011). An epistemic logic of extensive games. *Electronic Notes in Theoretical Computer Science*, 278, pp. 245-260.
17. Lorini, E., Schwarzentruber, F. (2011). A logic for reasoning about counterfactual emotions. *Artificial Intelligence*, 175(3-4), pp. 814-847.
18. Lorini, E., Schwarzentruber, F. (2010). A Modal Logic of Epistemic Games. *Games*, 1(4), pp. 478-526.
19. Lorini, E. (2010). A dynamic logic of agency II: deterministic DLA, coalition logic, and game theory. *Journal of Logic, Language and Information*, 19(3), pp. 327-351.
20. Ben-Naïm, J., Bonnefon, J. F., Herzig, A., Leblois, S., Lorini, E. (2010). Computer-Mediated Trust in Self-Interested Expert Recommendations. *AI & Society*, 25(4), pp. 413-422.
21. Herzig, A., Lorini, E., Hübner, J. F., Vercouter, L. (2010). A logic of trust and reputation. *Logic Journal of the IGPL*, 18(1), pp. 214-244.
22. Herzig, A., Lorini, E. (2010). A dynamic logic of agency I: STIT, abilities and powers. *Journal of Logic, Language and Information*, 19(1), pp. 89-121.
23. Herzig, A., de Lima, T., Lorini, E. (2009). On the dynamics of institutional agreements. *Synthese*, 171(2), pp. 321-355.
24. Lorini, E., Longin, D., Gaudou, B., Herzig, A. (2009). The logic of acceptance: grounding institutions on agents' attitudes. *Journal of Logic and Computation*, 19(6), pp. 901-940.
25. Lorini, E., Herzig, A. (2008). A logic of intention and attempt. *Synthese*, 163(1), pp. 45-77.
26. Lorini, E., Castelfranchi, C. (2007). The cognitive structure of surprise: looking for basic principles. *Topoi: An International Review of Philosophy*, 26(1), pp. 133-149.
27. Lorini, E., Castelfranchi, C. (2006). The unexpected aspects of surprise. *International Journal of Pattern Recognition and Artificial Intelligence*, 20 (6), pp. 817-833.

(B) **Book chapters**

1. Herzig, A., Lorini, E., Troquard, N. (forthcoming). Action theories. In S. O. Hansson, V. F. Hendricks (Eds.), *Handbook of Formal Philosophy*, Springer.
2. Gaudou, B., Herzig, A., Longin, D., Lorini, E. (2015). On modal logics of group belief. In A. Herzig, E. Lorini (Eds.), *The cognitive foundations of group attitudes and social interaction*, series "Studies in the Philosophy of Sociality", Springer.
3. Ben-Naïm, J., Longin, D., Lorini, E. (2014). Formalisation de systèmes d'agent cognitif, de la confiance, et des émotions. In P. Marquis, O. Papini, H. Prade (Eds.), *Panorama actuel de l'IA: ses bases méthodologiques et ses développements*, Cépaduès, Toulouse.
4. Andreas Herzig, Emiliano Lorini (2014). A modal logic of perceptual belief. In F. Lihoreau, M. Rebuschi (Eds.), *Epistemology, Context and Formalism*, Springer, Synthese Library, Volume 369, pp. 197-211.
5. Lorini, E., Herzig, A. (2014). Direct and Indirect Common Belief. In A. Konzelmann Ziv, H. B. Schmid, U. Schmid (Eds.), *Institutions, Emotions, and Group Agents*, Springer book series "Studies in the Philosophy of Sociality", Springer, Volume 2, pp. 355-372.

6. Herzig, A., de Lima, T., Lorini, E., Troquard, N. (2014). Three traditions in the logic of action: bringing them together. In R. Trypuz (Eds.), *Krister Segerberg on Logic of Actions*, Studia Logica book series "'Trends in Logic", subseries "Outstanding contributions", Springer, Volume 1, pp. 61-84.

7. Ben-Naïm, J., Bonnefon, J.-F., Herzig, A., Leblois, S., Lorini, E. (2013). Computer-mediated trust in self-interested expert recommendations. In S. Cowley, F. Vallee-Tourangeau (Eds.), *Cognition beyond the brain - Interactivity and human thinking*, Springer, pp. 53-70.

8. Broersen, J., Gabbay, D., Herzig, A., Lorini, E., Meyer, J.-J., Parent, X., van der Torre, L. (2013). Deontic logic. In S. Ossowski (Eds.), *Agreement Technologies*, Springer, pp. 108-127.

9. Balke, T., da Costa Pereira, C., Dignum, F., Lorini, E., Rotolo, A., Vasconcelos, W., Villata, S. (2013). Norms in MAS: Definitions and Related Concepts. In G. Andrighetto, G. Governatori, P. Noriega, L. van der Torre (Eds.), *Normative Multi-Agent Systems 2013*, Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, pp. 1-31.

10. Broersen, J., Cranefield, S., Elrakaiby, Y., Gabbay, D., Grossi, D., Lorini, E., Parent, X., van der Torre, L., Tummolini, L., Turrini, P., Schwarzentruber, F. (2013). Normative Reasoning and Consequence. In G. Andrighetto, G. Governatori, P. Noriega, L. van der Torre (Eds.), *Normative Multi-Agent Systems 2013*, Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, pp. 33-70.

11. Lorini, E. (2013). Modal Logic and Intentional Agency. In B. Kaldis (Eds.), *The Encyclopedia of Philosophy and the Social Sciences*, SAGE Publications, pp. 621-623.

12. Lorini, E. (2013). Mutual Beliefs. In B. Kaldis (Eds.), *The Encyclopedia of Philosophy and the Social Sciences*, SAGE Publications, pp. 640-642.

13. Lorini, E. (2012). The Cognitive Anatomy and Functions of Expectations Revisited. In F. Paglieri, L. Tummolini, R. Falcone, M. Miceli (Eds.), *The Goals of Cognition. Essays in Honor of Cristiano Castelfranchi*, College Publications, pp. 273-288.

14. Herzig, A., Lorini, E., Moisan, F. (2012). A simple logic of trust based on propositional assignments. In F. Paglieri, L. Tummolini, R. Falcone, M. Miceli (Eds.), *The Goals of Cognition. Essays in Honor of Cristiano Castelfranchi*, College Publications, pp. 407-419.

15. Adam, C., Gaudou, B., Longin, D., Lorini, E. (2011). Logical Modeling of Emotions for Ambient Intelligence. In F. Mastrogiovanni, N.-Y. Chong (Eds.), *Handbook of Research on Ambient Intelligence and Smart Environments: Trends and Perspectives*, IGI Publishing, New York, pp. 108-127.

16. Macedo, L., Cardoso, A., Reisenzein, R., Lorini, E., Castelfranchi, C. (2009). Artificial Surprise. In J. Vallverdú, D. Casacuberta (Eds.), *Handbook of Research on Synthetic Emotions and Sociable Robotics*, IGI Publishing, New York, pp. 261-285.

17. Lorini, E., Castelfranchi, C. (2009). Intentional agents in defense. In M. Barley, H. Mouratidis, A. Unruh, D. Spears, F. Martinelli, P. Scerri (Eds.), *Safety and Security in Multi-Agent Systems: the Early Years*, Lecture Notes in Computer Science, vol. 4324, Springer-Verlag, Berlin, pp. 293-307.

18. Castelfranchi, C., Falcone, R., Lorini, E. (2008). A non-reductionist approach to trust. In J. Golbeck (Eds.), *Computing with Social Trust*, Human-Computer Interaction Series, Springer, Berlin, pp. 45-72.

19. Lorini, E., Piunti, M., Castelfranchi, C., Falcone, R., Miceli, M. (2008). Anticipation and emotions for goal-directed agents. In G. Pezzulo, M. Butz, R. Falcone, C. Castelfranchi, *The Challenge of Anticipation*, Lecture Notes in Computer Science, vol. 5225, Springer-Verlag, Berlin, pp. 135-160.

20. Castelfranchi, C., Giardini, F., Lorini, E., Tummolini, L. (2007). The prescriptive destiny of predictive attitudes: from expectations to norms via conventions. In G. Sartor, C. Cevenini, G. Quadri di Cardano (Eds.), *Agenti software e commercio elettronico: profili giuridici, tecnologici e psico-sociali*, GEDIT, Bologna, Italy, pp. 43-55.

21. Lorini, E., Marzo, F., Castelfranchi, C. (2005). A cognitive model of the altruistic mind. In B. Kokinov (Eds.), *Advances in Cognitive Economics*, NBU Press, Sofia, Bulgaria, pp. 282-294 (ISBN: 9545354046).

(C) **Conferences and workshops with published proceedings**

1. Herzig, A., Lorini, E., Maffre, F., Schwarzentruber, F. (forthcoming). Epistemic boolean games based on a logic of visibility and control. *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI 2016)*, AAAI press.
2. Charrier, T., Herzig, A., Lorini, E., Maffre, F., Schwarzentruber, F. (forthcoming). Building epistemic logic from observations and public announcements. *Proceedings of the Fifteenth International Conference on Principles of Knowledge Representation and Reasoning (KR 2016)*, AAAI press.
3. Balbiani, P., Fernandez-Duque, D., Lorini, E. (forthcoming). A Logical Theory of Belief Dynamics for Resource-Bounded Agents. *Proceedings of the Fifteenth International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2016)*, ACM press.
4. Grandi, U., Lorini, E., Perrussel, L. (forthcoming). Games of Influence. *Proceedings of the Fifteenth International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2016)*, ACM press.

   Lorini, E., Sartor, G. (2015). Influence and Responsibility: A Logical Analysis. In *Proceedings of the Twenty-Eighth Annual Conference on Legal Knowledge and Information Systems (JURIX 2015)*, Frontiers in Artificial Intelligence and Applications 279, IOS Press, pp. 51-60.
5. Lorini, E., Muehlenbernd, R. (2015). The long-term benefits of following fairness norms: a game-theoretic analysis. In *Proceedings of the 18th Conference on Principles and Practice of Multi-Agent Systems (PRIMA 2015)*, LNCS, Springer-Verlag, Berlin, pp. 301-318.
6. Herzig, A., Lorini, E., Maffre, F. (2015). A poor man's epistemic logic based on assignment and higher-order propositional observation. In *Proceedings of the LORI-V Workshop on Logic, Rationality and Interaction*, LNCS, Springer- Verlag, Berlin, pp. 156-168.
7. Dubois, D., Lorini, E., Prade, H. (2015). Revising desires: a possibility theory viewpoint. *Proceedings of the 11th International Conference on Flexible Query Answering Systems (FQAS 2015)*, Advances in Intelligent Systems and Computing, Volume 400, Springer, pp. 3-13.
8. Grandi, U., Lorini, E., Perrussel, L. (2015). Propositional Opinion Diffusion. In *Proceedings of the Fourteenth International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2015)*, ACM press, pp. 989-997.
9. Lorini, E., Jiang, G., Perrussel, L. (2014). Trust-based belief change. In *Proceedings of the 21st European Conference on Artificial Intelligence (ECAI 2014)*, IOS press, pp. 549–554.
10. Lorini, E., Sartor, G. (2014). A STIT Logic Analysis of Social Influence. In *Proceedings of the Thirteenth International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2014)*, ACM press, pp. 853-860.
11. Ben-Naim, J., Lorini, E. (2014). Evaluating Power of Agents from Dependence Relations in Boolean Games. In *Proceedings of the Thirteenth International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2014)*, ACM press, pp. 885-892.
12. Dubois, D., Lorini, E., Prade, H. (2014). Nonmonotonic desires: a possibility theory viewpoint. In *Proceedings of the International Workshop on Defeasible and Ampliative Reasoning (DARe@ECAI 2014)*, CEUR Workshop Proceedings, Vol. 1212.
13. Adam, C., Lorini, L. (2014). A BDI Emotional Reasoning Engine for an Artificial Companion. In *Proceedings of the Workshop on Agents and multi-agent Systems for AAL and e-HEALTH (PAAMS 2014)*, Springer, CCIS Series, pp. 66-78.
14. Herzig, A., Lorini, E., Walther, D. (2013). Reasoning about actions meets strategic logics. In *Proceedings of the LORI-IV Workshop on Logic, Rationality and Interaction*, Lecture Notes in Artificial Intelligence, vol. 8196, Springer-Verlag, Berlin, pp. 162-175.

15. van Ditmarsch, H., Herzig, A., Lorini, E., Schwarzentruber, F. (2013). Listen to me! Public announcements to agents that pay attention - or not. In *Proceedings of the LORI-IV Workshop on Logic, Rationality and Interaction*, Lecture Notes in Artificial Intelligence, vol. 8196, Springer-Verlag, Berlin, pp. 96-109.

16. Dubois, D., Lorini, E., Prade, H. (2013). Bipolar possibility theory as a basis for a logic of desires and beliefs. In *Proceedings of the International Conference on Scalable Uncertainty Management (SUM 2013)*, Lecture Notes in Computer Science, Springer-Verlag, pp. 204-218.

17. Balbiani, P., Lorini, E. (2013). Ockhamist Propositional Dynamic Logic: A Natural Link between PDL and CTL*. In *Proceedings of the 20th International Workshop on Logic, Language, Information and Computation (WOLLIC 2013)*, Lecture Notes in Computer Science, Springer-Verlag, pp. 251-265.

18. Gaudou, B., Lorini, E., Mayor, E. (2013). Moral Guilt: An Agent-Based Model Analysis. In *Proceedings of the 9th Conference of the European Social Simulation Association (ESSA 2013)*, Series "Advances in Social Simulation-Advances in Intelligent Systems and Computing", Vol. 229, pp. 95-106.

19. Grossi, D., Lorini, E., Schwarzentruber, F. (2013). Ceteris Paribus Structure in Logics of Game Forms. In B. C. Schipper (Eds.), *Proceedings of the International Conference on Theoretical Aspects of Rationality and Knowledge (TARK 2013)*, University of California, p. 94-104 (ISBN: 978-0-615-74716-3).

20. Lorini, E. (2012). Logics for Reasoning About Agents' Attitudes in Strategic Contexts. In *Proceedings of the Thirteenth International Workshop on Computational Logic in Multi-Agent Systems (CLIMA 2012)*, Lecture Notes in Artificial Intelligence, vol. 7486, Springer-Verlag, Berlin, pp. 26.

21. Lorini, E. (2012). On the logical foundations of moral agency. In *Proceedings of the Eleventh International Conference on Deontic Logic in Computer Science (DEON 2012)*, Lecture Notes in Computer Science, vol. 7393, Springer-Verlag, Berlin, pp. 108-122.

22. Herzig, A., de Lima, T., Lorini, E., Troquard, N. (2012). A Computationally Grounded Dynamic Logic of Agency, with an Application to Legal Actions. In *Proceedings of the Eleventh International Conference on Deontic Logic in Computer Science (DEON 2012)*, Lecture Notes in Computer Science, vol. 7393, Springer-Verlag, Berlin, pp. 170-183.

23. Dastani, M., Lorini, E. (2012). A logic of emotions: from appraisal to coping. In *Proceedings of the Eleventh International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2012)*, ACM Press, New York, pp. 1133-1140.

24. Gaudou, B., Herzig, A., Lorini, E., Sibertin-Blanc, C. (2011). How to do social simulation in logic: modelling the segregation game in a dynamic logic of assignments. In *Proceedings of the Twelfth International Workshop on Multi-Agent-Based Simulation (MABS 2011), Revised Selected Papers*, Lecture Notes in Artificial Intelligence, vol. 7124, Springer-Verlag, Berlin, pp. 1-12.

25. Riviere, J., Adam, C., Pesty, S., Pelachaud, C., Guiraud, N., Longin, D., Lorini, E. (2011). Expressive Multimodal Conversational Acts for SAIBA Agents. In *Proceedings of the Eleventh International Conference on Intelligent Virtual Agents (IVA 2011)*, Lecture Notes in Computer Science, vol. 6895, Springer-Verlag, Berlin, pp. 316-323.

26. Lorini, E. (2011). A Dynamic Logic of Knowledge, Graded Beliefs and Graded Goals and Its Application to Emotion Modelling. In *Proceedings of the LORI-III Workshop on Logic, Rationality and Interaction*, Lecture Notes in Artificial Intelligence, vol. 6953, Springer-Verlag, Berlin, pp. 165-178.

27. Lorini, E., Perrussel, L., Thévenin, J.-M. (2011). A Modal Framework for Relating Belief and Signed Information. In *Proceedings of the Twelfth Computational Logic in Multi-Agent Systems (CLIMA 2011)*, Lecture Notes in Artificial Intelligence, vol. 6814, Springer-Verlag, Berlin, pp. 58-73.

28. Herzig, A., Lorini, E., Troquard, N. (2011). A dynamic logic of institutional actions. In *Proceedings of the Twelfth Computational Logic in Multi-Agent Systems (CLIMA 2011)*, Lecture Notes in Artificial Intelligence, vol. 6814, Springer-Verlag, Berlin, pp. 228-233.

29. Herzig, A., Lorini, E., Moisan, F., Troquard, N. (2011). A dynamic logic of normative systems. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI 2011)*, AAAI Press, pp. 228-233.

30. Balbiani, P., Guiraud, N., Herzig, A., Lorini, E. (2011). Agents that speak: modelling communicative plans and information sources in a logic of announcements. In *Proceedings of the Tenth International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2011)*, ACM Press, New York, pp. 1207-1208.

31. Guiraud, N., Longin, D., Lorini, E., Pesty, S., Riviere, J. (2011). The face of emotions: a logical formalization of expressive speech acts. In *Proceedings of the Tenth International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2011)*, ACM Press, New York, pp. 1031-1038.

32. Lorini, E., Piunti, M. (2010). Introducing Relevance Awareness in BDI Agents. In *Proceedings of the Seventh International Workshop on Programming Multi-Agent Systems (ProMAS 2009), Revised Invited and Selected Papers*, Lecture Notes in Computer Science, vol. 5919, Springer-Verlag, Berlin, pp. 219-236.

33. van Ditmarsch, H., de Lima, T., Lorini, E. (2010). Intention Change via Local Assignments. In *Proceedings of the Third International Workshop on Languages, Methodologies, and Development Tools for Multi-Agent Systems (LADS 2010), Revised Invited and Selected Papers*, Lecture Notes in Computer Science, vol. 6822, Springer-Verlag, Berlin, pp. 136-151.

34. Lorini, E., Verdicchio, M. (2010). Towards a Logical Model of Social Agreement for Agent Societies. In *Proceedings of the 5th international conference on Coordination, Organizations, Institutions, and Norms in Agent Systems (COIN'09), Revised Selected Papers*, Lecture Notes in Computer Science, vol. 6069, Springer-Verlag, Berlin, pp. 147-162.

35. Lorini, E. (2010). The hidden path from delegation to obligations: a logical analysis. In *Proceedings of the Artificial Intelligence and Social Behavior Society Convention - Second Symposium on Social Networks and Multiagent Systems (AISB - SNAMAS 2010)*, The Society for the Study of Artificial Intelligence and Simulation of Behaviour (SSAISB Press).

36. Lorini, E. (2010). A Logical Analysis of Commitment Dynamics. In *Proceedings of the Tenth International Conference on Deontic Logic in Computer Science (DEON 2010)*, Lecture Notes in Artificial Intelligence, vol. 6181, Springer-Verlag, Berlin, pp. 288-305.

37. Lorini, E., de Lima, T., van Ditmarsch, H. (2010). A Logical Model of Intention and Plan Dynamics. In *Proceedings of the 19th European Conference on Artificial Intelligence (ECAI 2010)*, IOS Press, pp. 1075-1076.

38. Bourdon, J., Fueled, G., Herzig, A., Lorini, E. (2010). Trust in complex actions. In *Proceedings of the 19th European Conference on Artificial Intelligence (ECAI 2010)*, IOS Press, pp. 1037-1038.

39. Lorini, E., Schwarzentruber, F., Herzig, A. (2009). Epistemic Games in Modal Logic: Joint Actions, Knowledge and Preferences all together. In X. He, J. Horty, E. Pacuit (Eds.), *Proceedings of the LORI-II Workshop on Logic, Rationality and Interaction*, Lecture Notes in Artificial Intelligence, Springer-Verlag, Berlin, pp. 212-226.

40. Aucher, G., Grossi, D., Herzig, A., Lorini, E. (2009). Dynamic Context Logic. In X. He, J. Horty, E. Pacuit (Eds.), *Proceedings of the LORI-II Workshop on Logic, Rationality and Interaction*, Lecture Notes in Artificial Intelligence, Springer-Verlag, Berlin, pp. 15-26.

41. Lorini, E., Dastani, M., van Ditmarsch, H., Herzig, A. (2009). Intentions and Assignments. In X. He, J. Horty, E. Pacuit (Eds.), *Proceedings of the LORI-II Workshop on Logic, Rationality and Interaction*, Lecture Notes in Artificial Intelligence, Springer-Verlag, Berlin, pp. 198-211.

42. de Boer, M., Herzig, A., de Lima, T., Lorini, E. (2009). Tableaux for Acceptance Logic. In M. Baldoni, J. Bentahar, J. Lloyd, M. B. van Riemsdijk, *Declarative Agent Languages and Technologies VII (DALT 2009)*, Lecture Notes in Artificial Intelligence, Springer-Verlag, Berlin, pp. 85-100.

43. Lorini, E., Demolombe, R. (2009). From trust in information sources to trust in communication systems: an analysis in modal logic. In J. Broersen, J.-J. Meyer (Eds.), *Proceedings of*

the First International Workshop on Knowledge Representation for Agents and Multi-agent Systems (KRAMAS 2008), Revised Selected Papers, Lecture Notes in Artificial Intelligence, Springer-Verlag, Berlin, pp. 81-98.

44. Herzig, A., de Lima, T., Lorini, E. (2009). On the Dynamics of Institutional Agreements. In J. Broersen, J.-J. Meyer (Eds.), *Proceedings of the First International Workshop on Knowledge Representation for Agents and Multi-agent Systems (KRAMAS 2008), Revised Selected Papers*, Lecture Notes in Artificial Intelligence, Springer-Verlag, Berlin, pp. 66-80.
45. Lorini, E., Falcone, R., Castelfranchi, C. (2009). Trust within the context of organizations: a formal approach. In P. Degano, J. Guttman, F. Martinelli (Eds.), *Proceedings of the Fifth International Workshop on Formal Aspects in Security and Trust (FAST 2008), Revised Selected Papers*, Lecture Notes in Computer Science, vol. 5491, Springer-Verlag, Berlin, pp. 114-128.
46. Lorini, E., Longin, D., Gaudou, B. (2009). Anchoring the institutional dimension of speech acts in agents' attitudes: a logical approach. In *Proceedings of the IEEE International Conference on Research, Innovation and Vision for the Future (RIVF 2009)*, IEEE Press.
47. Lorini, E., Schwarzentruber, F. (2009). A logic for reasoning about counterfactual emotions. In C. Boutilier (Eds.), *Proceedings of the Twenty-first International Joint Conference on Artificial Intelligence (IJCAI'09)*, AAAI Press, pp. 867-872.
48. Lorini, E., Longin, D. (2008). A logical account of institutions: from acceptances to norms via legislators. In G. Brewka, J. Lang (Eds.), *Proceedings of the 11th International Conference on Principles of Knowledge Representation and Reasoning (KR 2008)*, AAAI Press, pp. 38-48.
49. Lorini, E., Demolombe, R. (2008). Trust and norms in the context of computer security: toward a logical formalization. In R. van der Meyden, L. van der Torre, *Proceedings of the Ninth International Conference on Deontic Logic in Computer Science (DEON 2008)*, Lecture Notes in Computer Science, vol. 5076, Springer-Verlag, Berlin, pp. 50-64.
50. Lorini, E., Demolombe, R. (2008). From Binary Trust to Graded Trust in Information Sources: A Logical Perspective. In S. Barber, J. Sabater-Mir, M. P. Singh (Eds.), *Proceedings of the Eleventh International Workshop on Trust in Agent Societies (TRUST 2008), Revised Selected Papers*, Lecture Notes in Artificial Intelligence, vol. 5396, Springer-Verlag, Berlin, pp. 205-225.
51. Gaudou, B., Longin, D., Lorini, E., Tummolini, L. (2008). Anchoring Institutions in Agents' Attitudes: Towards a Logical Framework for Autonomous MAS. In S. Parsons, J. P. Mueller (Eds.), *Proceedings of the Seventh International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2008)*, ACM Press, New York, pp. 728-735.
52. Herzig, A., Lorini, E., Hübner, J. F., Ben-Naim, J., Boissier, O., Castelfranchi, C., Demolombe, R., Longin, D., Perrussel, L. (2008). Prolegomena for a logic of trust and reputation. In G. Boella, G. Pigozzi, M. P. Singh, H. Verhagen (Eds.), *Proceedings of the Third International Workshop on Normative Multiagent Systems (NorMAS 2008)*, University of Luxembourg Press, Luxembourg, pp. 143-157 (ISBN: 2919940481).
53. Trypuz, R., Lorini, E., Vieu, L. (2007). Solving Bratman's video game puzzle in two formalisms. In X. Arrazola, J. M. Larrazabal (Eds.), *Proceedings of ILCLI International Workshop on Logic and Philosophy of Knowledge, Communication and Action (LogKCA 2007)*, University of the Basque Country Press, Donostia, pp. 411-426 (ISBN: 9788498600223).
54. Lorini, E., Herzig, A., Broersen, J., Troquard, N. (2007). Grounding power on actions and mental attitudes. In R. Verbrugge, B. Duniz-Keplicz (Eds.), *Proceedings of the Third International Workshop on Formal Approaches to Multi-Agent Systems (FAMAS 2007)*, Durham University Press, Durham, UK, pp. 19-37.
55. Lorini, E., Piunti, M. (2007). The benefits of surprise in dynamic environments. In A. Paiva, R. Prada, R. W. Picard (Eds.), *Proceedings of the Second International Conference on Affective Computing and Intelligent Interaction (ACII 2007)*, Lecture Notes in Computer Science, vol. 4738, Springer-Verlag, Berlin, pp. 362-373.
56. Lorini, E., Troquard, N., Herzig, A., Castelfranchi, C. (2007). Delegation and mental states. In E. H. Durfee, M. Yokoo (Eds.), *Proceedings of the Sixth International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2007)*, ACM Press, New York, pp. 610-612.

57. Lorini, E., Herzig, A., Castelfranchi, C. (2006). Introducing "attempt" in a modal logic of intentional action. In M. Fisher, W. van der Hoek (Eds.), *Proceedings of the 10th European Conference on Logics in AI (JELIA 2006)*, Lecture Notes in Artificial Intelligence, vol. 4160, Springer-Verlag, Berlin, pp. 280-292.

58. Lorini, E., Falcone, R. (2005). Modeling Expectations in Cognitive Agents. In C. Castelfranchi, C. Balkenius, M. Butz, A. Ortony (Eds.), *Proceedings of AAAI 2005 Fall Symposium-From Reactive to Anticipatory Cognitive Embodied Systems*, AAAI Press, Menlo Park, pp. 114-121.

59. Lorini, E., Tummolini, L., Herzig, A. (2005). Establishing Mutual Beliefs by Joint Attention: towards a Formal Model of Public Events. In B. Bara, L. Barsalou, M. Bucciarelli (Eds.), *Proceedings of the 27th Annual Conference of the Cognitive Science Society (CogSci 2005)*, Lawrence Erlbaum Associates, Mahwah, NJ, pp. 1325-1330.

60. Pezzulo, G., Lorini, E., Calvi, G. (2004). How do I know how much I don't know? A cognitive approach about Uncertainty and Ignorance. In K. D. Forbus, D. Gentner, T. Regier (Eds.), *Proceedings of the 26th Annual Meeting of the Cognitive Science Society (CogSci 2004)*, Lawrence Erlbaum Associates, Mahwah, NJ, pp. 1095-1100.

61. Castelfranchi, C., Giardini, F., Lorini, E., Tummolini, L. (2003). The prescriptive destiny of predictive attitudes: From Expectations to Norms via Conventions. In R. Alterman, D. Kirsh (Eds.), *Proceedings of the 25th Annual Conference of the Cognitive Science Society (CogSci 2003)*, Lawrence Erlbaum Associates, Mahwah, NJ, pp. 222-227.

62. Castelfranchi, C., Lorini, E. (2003). Cognitive Anatomy and Functions of Expectations. In F. Schmalhofer, R. M. Young, G. Katz (Eds.), *Proceedings of the First European Cognitive Science Conference (EuroCogSci 2003)*, Lawrence Erlbaum Associates, Mahwah, NJ, pp. 377-379.

Journal of Applied Non-Classical Logics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/tncl20

Temporal logic and its application to normative reasoning

Emiliano Lorini[a]
[a] IRIT-CNRS, Toulouse, France.
Published online: 09 Oct 2013.

To cite this article: Emiliano Lorini (2013) Temporal logic and its application to normative reasoning, Journal of Applied Non-Classical Logics, 23:4, 372-399, DOI: 10.1080/11663081.2013.841359

To link to this article: http://dx.doi.org/10.1080/11663081.2013.841359

Journal of Applied Non-Classical Logics, 2013
Vol. 23, No. 4, 372–399, http://dx.doi.org/10.1080/11663081.2013.841359

Temporal STIT logic and its application to normative reasoning

Emiliano Lorini*

IRIT-CNRS, Toulouse, France

I present a variant of STIT with time, called T-STIT (Temporal STIT), interpreted in standard Kripke semantics. On the syntactic level, T-STIT is nothing but the extension of atemporal individual STIT by: (i) the future tense and past tense operators, and (ii) the operator of group agency for the *grand coalition* (the coalition of all agents). A sound and complete axiomatisation for T-STIT is given. Moreover, it is shown that T-STIT supports reasoning about interesting normative concepts such as the concepts of achievement obligation and commitment.

Keywords: STIT logic; temporal logic; axiomatisation; deontic logic

1. Introduction

STIT logic (the logic of *Seeing to it That*; Belnap, Perloff, & Xu, 2001) is one of the most prominent formal accounts of agency. It is the logic of sentences of the form 'the agent i sees to it that φ is true'. Horty (2001) extends the STIT logic of Belnap et al. (2001) with operators of group agency in order to express sentences of the form 'the group of agents C sees to it that φ is true'. Following Lorini and Schwarzentruber (2011), one might use the terms 'individual STIT logic' and 'group STIT logic' to designate respectively the STIT logic of Belnap et al. (2001) (in which only the actions of agents are described) and Horty's (2001) variant of STIT logic (in which both actions of agents and joint actions of groups are represented).

The original semantics for STIT given by Belnap et al. (2001) is defined in terms of **BT**+**AC** structures: branching-time structures (**BT**) augmented by agent choice functions (**AC**). A **BT** structure is made of a set of moments and a tree-like ordering over them. An **AC** for a certain agent i is a function mapping each moment m into a partition of the set of histories passing through that moment, a history h being a maximal set of linearly ordered moments and the equivalence classes of the partition being the possible choices for agent i at moment m. As shown by Balbiani, Herzig, and Troquard (2008), Lorini and Schwarzentruber (2011), and Herzig and Schwarzentruber (2008), however, both 'atemporal individual STIT' (i.e., individual STIT without tense operators in the object language) axiomatised by Belnap et al. (2001, Chapter 17), and 'atemporal group STIT' (i.e., group STIT without tense operators) can be 'simulated' in standard Kripke semantics. A similar idea is proposed by Kooi and Tamminga (2008), who introduce the concept of 'consequential model'. Consequential models are equivalent to the Kripke atemporal group STIT models used by Herzig and Schwarzentruber (2008) and Lorini and Schwarzentruber (2011), in which the authors abstract away from the branching-time account of STIT.

*Email: lorini@irit.fr

The present article goes beyond previous work on atemporal STIT by presenting a variant of STIT logic with time interpreted in standard Kripke semantics and by providing a sound and complete axiomatisation for this logic. I call this variant of STIT logic T-STIT (Temporal STIT). On the syntactic level, the logic T-STIT is nothing but the extension of atemporal individual STIT by: (i) the future tense and past tense operators, and (ii) the operator of group agency in Horty's (2001) sense for the *grand coalition* (the coalition of all agents).

The main motivation of this work is that past research on the mathematical properties of STIT (e.g., completeness and decidability) has mainly focused on atemporal STIT, while extensions of STIT by tense operators are far less well studied and understood. The interest of studying them is that they offer valuable formal languages for representing a variety of normative concepts such as commitment (Bentahar, Moulin, Meyer, & Chaib-draa, 2004; Desai, Narendra, & Singh, 2008; Singh, 2008) and achievement obligation (i.e., the obligation to perform a given action at some point in the future; Broersen, Dastani, & van der Torre, 2003; Governatori & Rotolo, 2010) that have an intrinsic *temporal* nature. These concepts are fundamental for understanding normative relationships between individuals in a society and have become useful abstractions for the design of multi-agent systems since they can be used to model a variety of interactive situations like contracts, agreements, negotiation, dialogue, and argumentation.

The rest of the article is organised as follows. In Section 2, the logic T-STIT is presented and a complete axiomatisation for this logic is given. An application of T-STIT to the formalisation of normative concepts such as achievement obligation and commitment is given in Section 3. It is shown that that T-STIT allows us to capture subtle temporal properties of these normative concepts such as *persistence* (i.e., the conditions under which a commitment persists over time). In Section 4, related work on STIT logic is discussed.

2. A STIT logic with time

This section presents the syntax and a Kripke-style semantics for T-STIT (Subsections 2.1 and 2.2). An axiomatisation of T-STIT is provided in Subsection 2.3, while in Subsection 2.4 it is proved that this axiomatisation is sound and complete with respect to the given semantics.

In the logic T-STIT, the so-called Chellas's STIT operators (Chellas, 1992) are taken as primitive operators of agency. As pointed out by Xu (1998) and Horty and Belnap (1995), so-called deliberative STIT operators and Chellas's STIT operators are interdefinable and just differ in the choice of primitive operators.

2.1. Syntax

Assume a countably infinite set of propositional atoms denoting basic facts $Atm = \{p, q, \ldots\}$ and a finite set of agents $Agt = \{1, \ldots, n\}$.

The language $\mathcal{L}_{\text{T-STIT}}(Atm, Agt)$ of the logic T-STIT is the set of formulas defined by the following BNF:

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid [i]\varphi \mid [Agt]\varphi \mid \Box\varphi \mid \mathsf{G}\varphi \mid \mathsf{H}\varphi$$

where p ranges over Atm and i ranges over Agt. The other Boolean constructions $\top$, $\bot$, $\vee$, $\rightarrow$ and $\leftrightarrow$ are defined from $\neg$ and $\wedge$ in the standard way.

Operators of the form $[i]$ are Chellas's STIT operators. The formula $[i]\varphi$ captures the fact that φ is guaranteed by a present action of agent i, and has to be read 'agent i sees to

it that φ regardless of what the other agents do'. I shorten the reading of $[i]\varphi$ to 'agent i sees to it that φ'. The crucial aspect of STIT theory is that an agent i's action is described in terms of the result that agent i brings about by her 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. I define the dual of the operator $[i]$ as follows: $\langle i\rangle\varphi \stackrel{\text{def}}{=} \neg[i]\neg\varphi$.

$[Agt]$ is a group STIT operator which captures the fact that φ is guaranteed by a present choice of all agents, and has to be read 'all agents see to it that φ by acting together'. The dual of the operator $[Agt]$ is defined as expected: $\langle Agt\rangle\varphi \stackrel{\text{def}}{=} \neg[Agt]\neg\varphi$. The modal operator $[Agt]$ will be fundamental in Section 2.2 in order to axiomatise a basic property relating action and time studied in STIT: the so-called property of *no choice between undivided histories* (Belnap et al., 2001, Chapter 7).

$\Box\varphi$ stands for 'φ is true regardless of what every agent does' or 'φ is true no matter what the agents do' or simply 'φ is necessarily true'. I define the dual of $\Box$ as follows: $\Diamond\varphi \stackrel{\text{def}}{=} \neg\Box\neg\varphi$. Note that the operators $[i]$ and $\Diamond$ can be combined in order to express what agents can do: $\Diamond[i]\varphi$ means 'agent i can see to it that φ'. Moreover, the operators $[i]$ and $\Box$ can be combined in order define the deliberative STIT operators $[i\ dstit{:}\ ]$ studied by Horty and Belnap (1995): $[i\ dstit{:}\ \varphi] \stackrel{\text{def}}{=} [i]\varphi \wedge \neg\Box\varphi$.

Finally, G and H are tense operators that are respectively used to express facts that are always true in the *strict* future and facts that are always true in the past. $\mathsf{G}\varphi$ means 'φ will always be true in the future' and $\mathsf{H}\varphi$ means 'φ has always been true in the past'. I define the dual of the future tense operator G as follows: $\mathsf{F}\varphi \stackrel{\text{def}}{=} \neg\mathsf{G}\neg\varphi$. $\mathsf{F}\varphi$ means 'φ will be true at some point in the future'. Moreover, I define the dual of the past tense operator H as follows: $\mathsf{P}\varphi \stackrel{\text{def}}{=} \neg\mathsf{H}\neg\varphi$. $\mathsf{P}\varphi$ means 'φ has been true at some point in the past'.

The following abbreviations will also be convenient:

$$\mathsf{G}^*\varphi \stackrel{\text{def}}{=} \varphi \wedge \mathsf{G}\varphi;$$
$$\mathsf{F}^*\varphi \stackrel{\text{def}}{=} \neg\mathsf{G}^*\neg\varphi.$$

$\mathsf{G}^*\varphi$ stands for 'φ is true in the present and will always be true', whereas $\mathsf{F}^*\varphi$ stands for 'φ is true in the present or will be true at some point in the future'.

2.2. A Kripke semantics for STIT logic with time

The basic notion in the semantics is the notion of a temporal Kripke STIT model that is nothing but a multi-relational Kripke model with special constraints on the accessibility relations. For notational convenience, in what follows I am going to use the following abbreviations. Given a set of elements W, an arbitrary binary relation $\mathcal{R}$ on W and an element w in W, let $\mathcal{R}(w) = \{v \in W | (w, v) \in \mathcal{R}\}$. Moreover, given two binary relations $\mathcal{R}_1$ and $\mathcal{R}_2$ on W let $\mathcal{R}_1 \circ \mathcal{R}_2$ be the standard operation of composition between binary relations. Temporal Kripke STIT models can be seen as extensions of Zanardo's (1996) Ockhamist frames by a choice component, i.e., by accessibility relations for the individual choices of the agents and an accessibility relation for the collective choice of the grand coalition Agt.

Definition 1 (Temporal Kripke STIT model). *The class of temporal Kripke STIT models includes all tuples* $M = (W, \mathcal{R}_\Box, \{\mathcal{R}_i | i \in Agt\}, \mathcal{R}_{Agt}, \mathcal{R}_\mathsf{G}, \mathcal{R}_\mathsf{H}, \mathcal{V})$ *where:*

- W *is a nonempty set of possible worlds;*
- $\mathcal{R}_\Box$*, every* $\mathcal{R}_i$ *and* $\mathcal{R}_{Agt}$ *are equivalence relations between worlds in* W *such that:*

  (C1) $\mathcal{R}_i \subseteq \mathcal{R}_\Box$*;*

(C2) *for all* $u_1, \ldots, u_n \in W$*: if* $(u_i, u_j) \in \mathcal{R}_\Box$ *for all* $i, j \in \{1, \ldots, n\}$ *then* $\bigcap_{1 \leq i \leq n} \mathcal{R}_i(u_i) \neq \emptyset$*;*

(C3) *for all* $w \in W$*:* $\mathcal{R}_{Agt}(w) = \bigcap_{i \in Agt} \mathcal{R}_i(w)$*;*

- $\mathcal{R}_\mathsf{G}$ *and* $\mathcal{R}_\mathsf{H}$ *are binary relations between worlds in* W *such that* $\mathcal{R}_\mathsf{G}$ *is serial and transitive,* $\mathcal{R}_\mathsf{H}$ *is the inverse relation of* $\mathcal{R}_\mathsf{G}$ *(i.e.,* $\mathcal{R}_\mathsf{H} = \mathcal{R}_\mathsf{G}^{-1} = \{(w, v) | (v, w) \in \mathcal{R}_\mathsf{G}\}$*), and:*

(C4) *for all* $w, v, u \in W$*: if* $v, u \in \mathcal{R}_\mathsf{G}(w)$ *then* $u \in \mathcal{R}_\mathsf{G}(v)$ *or* $v \in \mathcal{R}_\mathsf{G}(u)$ *or* $u = v$*;*

(C5) *for all* $w, v, u \in W$*: if* $v, u \in \mathcal{R}_\mathsf{H}(w)$ *then* $u \in \mathcal{R}_\mathsf{H}(v)$ *or* $v \in \mathcal{R}_\mathsf{H}(u)$ *or* $u = v$*;*

(C6) $\mathcal{R}_\mathsf{G} \circ \mathcal{R}_\Box \subseteq \mathcal{R}_{Agt} \circ \mathcal{R}_\mathsf{G}$*;*

(C7) *for all* $w \in W$*: if* $v \in \mathcal{R}_\Box(w)$ *then* $v \notin \mathcal{R}_\mathsf{G}(w)$*;*

- $\mathcal{V} : Atm \longrightarrow 2^W$ *is a valuation function for atomic formulas.*

The valuation function $\mathcal{V}$ is used to identify those states in a model in which a given atomic proposition is true. Specifically, $w \in \mathcal{V}(p)$ means that p is true at world w.

$\mathcal{R}_\Box(w)$ is the set of worlds that are alternative to the world w. Following the Ockhamist's view of time (Prior, 1967; Thomason, 1984; Zanardo, 1996), I call the equivalence classes induced by the equivalence relation $\mathcal{R}_\Box$ *moments*.[1] The set of all moments in the model M is denoted by *Mom* and the elements in *Mom* are denoted by $m, m', \ldots$.

$\mathcal{R}_\mathsf{G}(w)$ defines the set of worlds that are in the *strict* future of world w, where the strict future does not include the present. $\mathcal{R}_\mathsf{H}(w)$ defines the set of worlds that are in the *past* of world w. The Constraint C7 ensures that if two worlds belong to the same moment then one of them cannot be in the future of the other. Since the relation $\mathcal{R}_\Box$ is reflexive, the Constraint C7 implies the irreflexivity of the relation $\mathcal{R}_\mathsf{G}$, i.e., for all $w \in W$ we have $w \notin \mathcal{R}_\mathsf{G}(w)$. The fact that the relation $\mathcal{R}_\mathsf{G}$ is transitive and irreflexive just means that it is a strict partial order on the set W. The Constraint C4 ensures that time is connected towards the future, while the Constraint C5 ensures that time is connected towards the past.

Let $\mathcal{T}(w) = \mathcal{R}_\mathsf{H}(w) \cup \{w\} \cup \mathcal{R}_\mathsf{G}(w)$ be the set of worlds that are temporally related with world w. The fact that the relation $\mathcal{R}_\mathsf{G}$ is irreflexive and transitive together with the Constraints C4 and C5 ensure that $\mathcal{R}_\mathsf{G}$ is a strict linear (or total) order on the set $\mathcal{T}(w)$. For every world w in W, I call the linearly ordered set $(\mathcal{T}(w), \mathcal{R}_\mathsf{G})$ the history going through w. For notational convenience, I write h_w instead of $(\mathcal{T}(w), \mathcal{R}_\mathsf{G})$. Note that, because of the seriality of the relation $\mathcal{R}_\mathsf{G}$, every history h_w is infinite.

This highlights that there is a one-to-one correspondence between worlds and histories, as for every world w there exists a unique history going through it. In other words, one can interchangeably use the term 'world' and 'history going through a certain world' without lost of generality. As every world in a model is identified with a unique history going through it, the equivalence relation $\mathcal{R}_\Box$ can also be understood as an equivalence relation between historic alternatives: $(w, v) \in \mathcal{R}_\Box$ means that the history going through v is alternative to the history going through w.

From the temporal relation $\mathcal{R}_\mathsf{G}$ over worlds in W, we can define the following relation $<$ over moments in *Mom*, where $m < m'$ means that moment m' is in the strict future of moment m.

Definition 2 (Ordering of moments). *For all* $m, m' \in Mom$*: let* $m < m'$ *if and only if there are* $w \in m$ *and* $v \in m'$ *such that* $(w, v) \in \mathcal{R}_\mathsf{G}$.

For every world w, the set $\mathcal{R}_i(w)$ identifies agent i's *actual* choice at w, that is to say, the set of all alternatives that is forced by agent i's actual choice at w. Because of the

one-to-one correspondence between worlds and histories, one can also identify agent i's *actual* choice at w with the set of histories $\{h_v : v \in \mathcal{R}_i(w)\}$. In other words, in T-STIT an agent chooses among different sets of histories.

Constraint C1 in Definition 1 just means that an agent can only choose among possible alternatives. This constraint ensures that, for every world w, the equivalence relation $\mathcal{R}_i$ induces a partition of the set $\mathcal{R}_\Box(w)$. An element of this partition is a choice that is *possible (or available)* for agent i at w.

Constraint C2 expresses the so-called assumption of *independence of agents* or *independence of choices*: if $\mathcal{R}_1(u_1)$ is a possible choice for agent 1, $\mathcal{R}_2(u_2)$ is a possible choice for agent 2,…, $\mathcal{R}_n(u_n)$ is a possible choice for agent n, then their intersection is nonempty. More intuitively, this means that agents can never be deprived of choices due to the choices made by other agents.

For every world w, the set $\mathcal{R}_{Agt}(w)$ identifies the *actual* choice of group Agt at w – that is to say, the set of all alternatives that is forced by the collective choice of all agents at w. Constraint C3 just says that the set of alternatives that is forced by the collective choice of all agents at w is equal to the pointwise intersection of the sets of alternatives that are forced by the individual choices of the agents in Agt at w. In other words, the choice of a group corresponds to the intersection of the choices of the individuals in the group. This corresponds to the notion of joint action proposed by Horty (2001), where the joint action of a group is described in terms of the result that the agents in the group bring about by acting together.

The Constraint C6 expresses a basic relation between action and time: if v is in the future of w and u and v are in the same moment, then there exists an alternative z in the collective choice of all agents at w such that u is in the future of z. This constraint corresponds to the property of *no choice between undivided histories* given in STIT logic (Belnap et al., 2001, Chapter 7). It captures the idea that if two histories come together in some future moment then, in the present, each agent does not have a choice between these two histories. This implies that if an agent can choose between two histories at a later stage then she does not have a choice between them in the present. The Constraint C6 is crucial in order to prove that the relation $<$ defined above is a tree-like ordering of the moments in *Mom*.

Proposition 3. *The relation $<$ satisfies the following four properties for all $m, m', m'' \in Mom$:*
(Irreflexivity) $m \not< m$;
(Transitivity) if $m < m'$ and $m' < m''$ then $m < m''$;
(Asymmetry) if $m < m'$ then $m' \not< m$;
(No backward branching) if $m' < m$ and $m'' < m$ then $m' < m''$ or $m'' < m'$ or $m' = m''$.

Proof. The irreflexivity of $<$ follows from the Constraint C7. Indeed, suppose that $m < m$. This implies that there are $w, v \in W$ such that $(w, v) \in \mathcal{R}_\Box$ and $(w, v) \in \mathcal{R}_{\mathsf{G}}$. But this is in contradiction with the Constraint C7.

Transitivity of $<$ follows from the transitivity of the relation $\mathcal{R}_{\mathsf{G}}$ and from the Constraints C1, C3 and C6. Suppose that $m < m'$ and $m' < m''$. Therefore, for some arbitrary worlds w_1, w_2, w_3, w_4 we have $w_1 \in m$, $w_2, w_3 \in m'$, $w_4 \in m''$, $w_2 \in \mathcal{R}_{\mathsf{G}}(w_1)$ and $w_4 \in \mathcal{R}_{\mathsf{G}}(w_3)$. By the Constraint C6, it follows that there is $w_5 \in \mathcal{R}_{Agt}(w_1)$ such that $w_3 \in \mathcal{R}_{\mathsf{G}}(w_5)$. The Constraints C1 and C3 together imply that $\mathcal{R}_{Agt} \subseteq \mathcal{R}_\Box$. Hence, there is $w_5 \in m$ such that $w_3 \in \mathcal{R}_{\mathsf{G}}(w_5)$. Thus, by the transitivity of $\mathcal{R}_{\mathsf{G}}$, there is $w_5 \in m$ such that $w_4 \in \mathcal{R}_{\mathsf{G}}(w_5)$. It follows that $m < m''$.

The asymmetry of $<$ follows from its irreflexivity.

No backward branching follows from the Constraints C5 and C6. Suppose that $m' < m$ and $m'' < m$. Therefore, for some arbitrary worlds w_1, w_2, w_3, w_4 we have $w_1, w_2 \in m$, $w_3 \in m'$, $w_4 \in m''$, $w_3 \in \mathcal{R}_{\mathsf{H}}(w_1)$ and $w_4 \in \mathcal{R}_{\mathsf{H}}(w_2)$. By the Constraint C6, it follows that there is $w_5 \in \mathcal{R}_{Agt}(w_3)$ such that $w_5 \in \mathcal{R}_{\mathsf{H}}(w_2)$. Thus, since $\mathcal{R}_{Agt} \subseteq \mathcal{R}_{\Box}$, there is $w_5 \in m'$ such that $w_5 \in \mathcal{R}_{\mathsf{H}}(w_2)$. Moreover, by the Constraint C5 and the fact that $w_4 \in \mathcal{R}_{\mathsf{H}}(w_2)$, it follows that there is $w_5 \in m'$ such that $w_5 \in \mathcal{R}_{\mathsf{H}}(w_4)$ or $w_4 \in \mathcal{R}_{\mathsf{H}}(w_5)$ or $w_4 = w_5$. Since $w_4 \in m''$, the latter implies that $m' < m''$ or $m'' < m'$ or $m' = m''$. ■

Another interesting property of temporal Kripke STIT models that follows from the Constraints C6 and C7 is the so-called property of *past isomorphism* (PI), which is similar to the property of *past isomorphism* of Zanardo's (1996) Ockhamist frames.

Proposition 4. *For all $w, v \in W$, if $(w, v) \in \mathcal{R}_{\Box}$ then there exists an order-isomorphism f between $\mathcal{R}_{\mathsf{H}}(w)$ and $\mathcal{R}_{\mathsf{H}}(v)$ such that, for all $u \in \mathcal{R}_{\mathsf{H}}(w)$, $(u, f(u)) \in \mathcal{R}_{Agt}$.*[2]

Proof. First of all note that, by Constraints C1 and C3, $\mathcal{R}_{Agt} \subseteq \mathcal{R}_{\Box}$.

Assume $(w, v) \in \mathcal{R}_{\Box}$. By Constraint C7 and the fact that $\mathcal{R}_{Agt} \subseteq \mathcal{R}_{\Box}$, every world has at most one $\mathcal{R}_{Agt}$-equivalent in any history and, hence, by Constraint C6, every world u such that $u \in \mathcal{R}_{\mathsf{H}}(w)$ has exactly one $\mathcal{R}_{Agt}$-equivalent in $\mathcal{R}_{\mathsf{H}}(v)$. Therefore, the restriction of $\mathcal{R}_{Agt}$ to the set $\mathcal{R}_{\mathsf{H}}(w) \times \mathcal{R}_{\mathsf{H}}(v)$ is an order-preserving bijective function. ■

Given a temporal Kripke STIT model $M = (W, \mathcal{R}_{\Box}, \{\mathcal{R}_i | i \in Agt\}, \mathcal{R}_{Agt}, \mathcal{R}_{\mathsf{G}}, \mathcal{R}_{\mathsf{H}}, \mathcal{V})$, a world w and a formula φ, I write $M, w \models \varphi$ to mean that φ is true at world w in M. The truth conditions of formulas are then defined as follows:

$$
\begin{aligned}
M, w \models p &\iff w \in \mathcal{V}(p) \\
M, w \models \neg\varphi &\iff M, w \not\models \varphi \\
M, w \models \varphi \wedge \psi &\iff M, w \models \varphi \text{ and } M, w \models \psi \\
M, w \models \Box\varphi &\iff \forall v \in \mathcal{R}_{\Box}(w) : M, v \models \varphi \\
M, w \models [i]\varphi &\iff \forall v \in \mathcal{R}_{i}(w) : M, v \models \varphi \\
M, w \models [Agt]\varphi &\iff \forall v \in \mathcal{R}_{Agt}(w) : M, v \models \varphi \\
M, w \models \mathsf{G}\varphi &\iff \forall v \in \mathcal{R}_{\mathsf{G}}(w) : M, v \models \varphi \\
M, w \models \mathsf{H}\varphi &\iff \forall v \in \mathcal{R}_{\mathsf{H}}(w) : M, v \models \varphi
\end{aligned}
$$

Example 5. *Figure 1 provides an example that clearly illustrates the semantics of T-STIT. At world w in the temporal Kripke STIT model M represented in the figure, agent 2 sees to it that p is true (i.e., $M, w \models [2]p$). Indeed, p holds at every world in agent 2's choice at w. Moreover, at w agent 1 sees to it that p or q is true (i.e., $M, w \models [1](p \vee q)$) because either p or q hold at every world in agent 1's choice at w. Finally, at w the group $\{1, 2\}$ sees to it that q will be true at some point in the future (i.e., $M, w \models [\{1, 2\}]\mathsf{F}q$) because q holds at some future point of every history in the group $\{1, 2\}$'s choice at w.*

Given a T-STIT formula φ, I say that φ is T-STIT *valid*, denoted by $\models_{\text{T-STIT}} \varphi$, if and only if for every temporal Kripke STIT model M and for every world w in M we have $M, w \models \varphi$. I say that φ is satisfiable in T-STIT if and only if $\neg\varphi$ is not T-STIT valid.

The following proposition provides an example of interesting T-STIT validities.

Proposition 6. *The following two formulas are T-STIT valid:*

$$
\begin{aligned}
\mathsf{G}\Diamond\mathsf{G}^*\varphi &\rightarrow \langle Agt\rangle\mathsf{G}\varphi \\
\mathsf{G}\Diamond(\mathsf{G}^*\varphi \wedge \mathsf{F}^*\psi) &\rightarrow \langle Agt\rangle(\mathsf{G}\varphi \wedge \mathsf{F}\psi)
\end{aligned}
$$

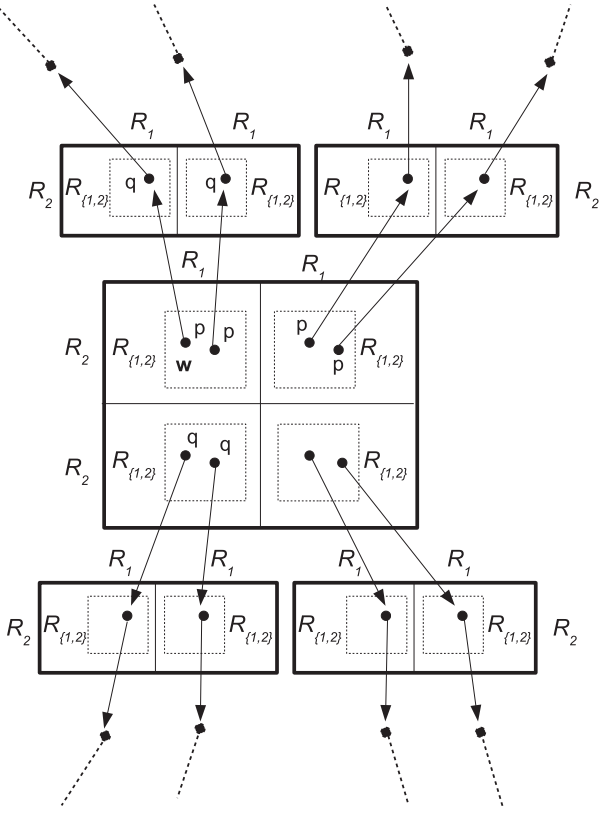


Figure 1. Example of temporal Kripke STIT model.
Note: The world w is the actual world. Choices of agent 1 are represented by columns whereas choices of agent 2 are represented by rows. Choices of the group $\{1, 2\}$ are represented by dotted rectangles. The temporal relation $\mathcal{R}_{\mathsf{G}}$ is represented by the arrows.

Proof. I only prove the first validity, as the second one can be proved in a similar way. I prove its contrapositive, namely I prove that $[Agt]\mathsf{F}\varphi \rightarrow \mathsf{F}\Box\mathsf{F}^*\varphi$ is T-STIT valid. Suppose $M, w \models [Agt]\mathsf{F}\varphi$. This means that:

(A) for all $v \in \mathcal{R}_{Agt}(w)$ there is $u \in \mathcal{R}_{\mathsf{G}}(v)$ such that $M, u \models \varphi$.

Let

$$A^{\varphi}_{w} = \left\{u | M, u \models \varphi \text{ and } \exists v \in \mathcal{R}_{Agt}(w) \text{ such that } u \in \mathcal{R}_{\mathsf{G}}(v)\right\}$$

and let

$$PA^{\varphi}_{w} = \left\{\bar{u} \in \mathcal{R}_{\mathsf{G}}(w) | \exists u \in A^{\varphi}_{w} \text{ such that } (\bar{u}, u) \in \mathcal{R}_{\Box}\right\}.$$

Moreover let

$$NB^{\varphi}_{w} = \mathcal{R}_{\mathsf{G}}(w) \setminus \bigcup_{\bar{u} \in PA^{\varphi}_{w}} \mathcal{R}_{\mathsf{G}}(\bar{u}).$$

I am going to show that $NB^{\varphi}_{w} \neq \emptyset$. Suppose $NB^{\varphi}_{w} = \emptyset$. Thus, there exists $\bar{u} \in PA^{\varphi}_{w}$ such that $\mathcal{R}_{\mathsf{G}}(w) = \mathcal{R}_{\mathsf{G}}(\bar{u})$. Hence, by Constraint C5, there exists $\bar{u} \in PA^{\varphi}_{w}$ such that $w = \bar{u}$.

The latter implies that there exists $\bar{u} \in \mathcal{R}_{\mathsf{G}}(w)$ such that $w = \bar{u}$, which is in contradiction with the fact that $\mathcal{R}_{\mathsf{G}}$ is irreflexive.

Take any world $z \in NB^{\varphi}_{w}$. Clearly, $z \in \mathcal{R}_{\mathsf{G}}(w)$. I am going to show that

(B) for all $v \in \mathcal{R}_{\Box}(z)$ either $M, v \models \varphi$ or there is $u \in \mathcal{R}_{\mathsf{G}}(v)$ such that $M, u \models \varphi$.

Suppose $v \in \mathcal{R}_{\Box}(z)$. From $z \in \mathcal{R}_{\mathsf{G}}(w)$, by Constraint C6 and the previous observation (A), we have that there are z', v' such that $z' \in \mathcal{R}_{Agt}(w)$, $v \in \mathcal{R}_{\mathsf{G}}(z')$, $v' \in \mathcal{R}_{\mathsf{G}}(z')$ and $M, v' \models \varphi$. Moreover, by Constraint C4, $v' \in \mathcal{R}_{\mathsf{G}}(v)$ or $v \in \mathcal{R}_{\mathsf{G}}(v')$ or $v = v'$. I am going to show that $v = v'$ or $v' \in \mathcal{R}_{\mathsf{G}}(v)$ by reductio ad absurdum.

Suppose $v \in \mathcal{R}_{\mathsf{G}}(v')$. Clearly, $v' \in A^{\varphi}_{w}$. From $z' \in \mathcal{R}_{Agt}(w)$, $v \in \mathcal{R}_{\Box}(z)$, $z \in \mathcal{R}_{\mathsf{G}}(w)$, $v \in \mathcal{R}_{\mathsf{G}}(z')$, $v' \in \mathcal{R}_{\mathsf{G}}(z')$ and $v \in \mathcal{R}_{\mathsf{G}}(v')$, by Proposition 4, it follows that there exists $\bar{v}' \in \mathcal{R}_{\mathsf{G}}(w)$ such that $v' \in \mathcal{R}_{Agt}(\bar{v}')$ and $z \in \mathcal{R}_{\mathsf{G}}(\bar{v}')$. Since $\mathcal{R}_{Agt} \subseteq \mathcal{R}_{\Box}$ and $v' \in A^{\varphi}_{w}$, the latter implies that there exists $\bar{v}' \in \mathcal{R}_{\mathsf{G}}(w)$ such that $\bar{v}' \in PA^{\varphi}_{w}$, $v' \in \mathcal{R}_{Agt}(\bar{v}')$ and $z \in \mathcal{R}_{\mathsf{G}}(\bar{v}')$. $\bar{v}' \in PA^{\varphi}_{w}$ and $z \in \mathcal{R}_{\mathsf{G}}(\bar{v}')$ together imply that $z \notin NB^{\varphi}_{w}$. But this is in contradiction with the initial hypothesis that $z \in NB^{\varphi}_{w}$.

By the previous item (B), we have that $M, z \models \Box\mathsf{F}^{*}\varphi$. Since $z \in \mathcal{R}_{\mathsf{G}}(w)$, $M, w \models \mathsf{F}\Box\mathsf{F}^{*}\varphi$. ■

2.3. Axiomatisation

Figure 2 contains a complete axiomatisation with respect to the class of temporal Kripke STIT models.

PC	All tautologies of classical propositional calculus
S5(i)	All S5-principles for the operators $[i]$
S5$(\Box)$	All S5-principles for the operator $\Box$
S5(Agt)	All S5-principles for the operator $[Agt]$
KD4(G)	All KD4-principles for the operator G
K(H)	All K-principles for the operator H
$(\Box \to i)$	$\Box\varphi \to [i]\varphi$
$(i \to Agt)$	$([1]\varphi_1 \wedge \ldots \wedge [n]\varphi_n) \to [Agt](\varphi_1 \wedge \ldots \wedge \varphi_n)$
(AIA)	$(\Diamond[1]\varphi_1 \wedge \ldots \wedge \Diamond[n]\varphi_n) \to \Diamond([1]\varphi_1 \wedge \ldots \wedge [n]\varphi_n)$
$(\mathbf{Conv}_{\mathsf{G},\mathsf{H}})$	$\varphi \to \mathsf{GP}\varphi$
$(\mathbf{Conv}_{\mathsf{H},\mathsf{G}})$	$\varphi \to \mathsf{HF}\varphi$
$(\mathbf{Connected}_{\mathsf{G}})$	$\mathsf{PF}\varphi \to (\mathsf{P}\varphi \vee \varphi \vee \mathsf{F}\varphi)$
$(\mathbf{Connected}_{\mathsf{H}})$	$\mathsf{FP}\varphi \to (\mathsf{P}\varphi \vee \varphi \vee \mathsf{F}\varphi)$
(NCUH)	$\mathsf{F}\Diamond\varphi \to \langle Agt\rangle\mathsf{F}\varphi$
(MP)	$\dfrac{\varphi,\ \varphi \to \psi}{\psi}$
(IRR)	$\dfrac{(\Box\neg p \wedge \Box(\mathsf{G}p \wedge \mathsf{H}p)) \to \varphi}{\varphi}$, provided p does not occur in φ

Figure 2. Axiomatisation of T-STIT.

This includes all tautologies of classical propositional calculus **(PC)** as well as modus ponens **(MP)**. Moreover, we have all the principles of the normal modal logic S5 for every operator $[i]$, for the operator $[Agt]$ and for the operator $\Box$, all principles of the normal modal logic KD4 for the future tense operator G and all principles of the normal modal logic K for the past tense operator H. That is, we have Axiom K for each operator: $(\blacksquare\varphi \wedge \blacksquare(\varphi \rightarrow \psi)) \rightarrow \blacksquare\psi$ with $\blacksquare \in \{\Box, \mathsf{G}, \mathsf{H}, [Agt]\} \cup \{[i] | i \in Agt\}$. We have Axiom D for the future tense modality G: $\neg(\mathsf{G}\varphi \wedge \mathsf{G}\neg\varphi)$. We have Axiom 4 for $\Box$, G, $[Agt]$ and for every $[i]$: $\blacksquare\varphi \rightarrow \blacksquare\blacksquare\varphi$ with $\blacksquare \in \{\Box, [Agt], \mathsf{G}\} \cup \{[i] | i \in Agt\}$. Furthermore, we have Axiom T for $\Box$, $[Agt]$ and for every $[i]$: $\blacksquare\varphi \rightarrow \varphi$ with $\blacksquare \in \{\Box, [Agt]\} \cup \{[i] | i \in Agt\}$. We have Axiom B for $\Box$, $[Agt]$ and for every $[i]$: $\varphi \rightarrow \blacksquare\neg\blacksquare\neg\varphi$ with $\blacksquare \in \{\Box, [Agt]\} \cup \{[i] | i \in Agt\}$. Finally we have the rule of necessitation for each modal operator: $\frac{\varphi}{\blacksquare\varphi}$ with $\blacksquare \in \{\Box, [Agt], \mathsf{G}, \mathsf{H}\} \cup \{[i] | i \in Agt\}$. In what follows, I write $\vdash_{\text{T-STIT}} \varphi$ if φ is a T-STIT theorem. Moreover, I say that φ is T-STIT consistent if $\not\vdash_{\text{T-STIT}} \neg\varphi$.

$(\Box \rightarrow i)$ and **(AIA)** are the two central principles in Xu's (1998) axiomatisation of the Chellas's STIT operators $[i]$. According to Axiom **$(\Box \rightarrow i)$**, if φ is true regardless of what every agent does, then every agent sees to it that φ. In other words, an agent brings about those facts that are inevitable.[3] According to Axiom **$(i \rightarrow Agt)$**, all agents bring about together what each of them brings about individually.

We have principles for the tense operators and for the relationship between time and action. **$(\mathbf{Connected_G})$** and **$(\mathbf{Connected_H})$** are the basic axioms for the linearity of the future and the linearity of the past (Goldblatt, 1992). **$(\mathbf{Conv_{G,H}})$** and **$(\mathbf{Conv_{H,G}})$** are the basic interaction axioms between future and past of minimal tense logic according to which 'what is, will always have been' and 'what is, has always been going to be'.

Axiom **(NCUH)** establishes a fundamental relationship between action and time and corresponds to the semantic constraint of 'no choice between undivided histories' over temporal Kripke STIT models (Constraint C6): if in some future world φ will be possible then the actual collective choice of all agents will possibly result in a state in which φ is true.

(IRR) is a variant of the well-known Gabbay's irreflexivity rule that has been widely used in the past for proving completeness results for different kinds of temporal logic in which time is supposed to be irreflexive (see, e.g., Gabbay, Hodkinson, & Reynolds, 1994; Reynolds, 2003; von Kutschera, 1997; Zanardo, 1996). The idea is that the special kind of irreflexivity for the relation $\mathcal{R}_{\mathsf{G}}$ expressed by the Constraint C7 in Definition 1, although not definable in terms of an axiom, can be characterised in an alternative sense by means of the rule **(IRR)**. This rule is perhaps more comprehensible if we consider its contrapositive: if p does not occur in φ and φ is T-STIT consistent, then $\Box\neg p \wedge \Box(\mathsf{G}p \wedge \mathsf{H}p) \wedge \varphi$ is T-STIT consistent.

Theorem 7. *The set of T-STIT validities is completely axiomatised by the principles given in Figure 2.*

2.4. Proof of Theorem 7

It is a routine task to show that the axioms given in Figure 2 are valid with respect to the class of temporal Kripke T-STIT models and that the rules of inference preserve validity. Thus, if φ is a T-STIT theorem then φ is T-STIT valid.

I am going to prove that if φ is T-STIT consistent then φ is T-STIT satisfiable. The proof is divided into two steps.

First of all, I introduce the class of *superadditive* temporal Kripke STIT models. While in the temporal Kripke STIT models the collective choice of the grand coalition is *equal*

to the pointwise intersection of the individual choices, in superadditive temporal Kripke STIT models the collective choice of the grand coalition is merely *included* in the pointwise intersection of the individual choices. I prove that the logic T-STIT does not distinguish the semantics in terms of temporal Kripke STIT models from the more 'liberal' semantics in terms of superadditive temporal Kripke STIT models. That is to say, the set of validities with respect to the class of temporal Kripke STIT models is equal to the set of validities with respect to the class of superadditive temporal Kripke STIT models (Lemma 9).

Secondly, I prove that the set of validities in the class of superadditive temporal Kripke STIT models is completely axiomatised by the principles given in Figure 2 (Lemma 10).

Theorem 7 directly follows from Lemma 10 and Lemma 9.

Let me define the class of *superadditive* temporal Kripke STIT models.

Definition 8 (Superadditive temporal Kripke STIT model). *The class of superadditive temporal Kripke STIT models includes all tuples* $M = (W, \mathcal{R}_\Box, \{\mathcal{R}_i | i \in Agt\}, \mathcal{R}_{Agt}, \mathcal{R}_\mathsf{G}, \mathcal{R}_\mathsf{H}, \mathcal{V})$ *where:*

- W *is a nonempty set of possible worlds;*
- $\mathcal{R}_\Box$*, every* $\mathcal{R}_i$ *and* $\mathcal{R}_{Agt}$ *are equivalence relations between worlds in* W *such that:*
  - **(C1)** $\mathcal{R}_i \subseteq \mathcal{R}_\Box$*;*
  - **(C2)** *for all* $u_1, \ldots, u_n \in W$*: if* $(u_i, u_j) \in \mathcal{R}_\Box$ *for all* $i, j \in \{1, \ldots, n\}$ *then* $\bigcap_{1 \leq i \leq n} \mathcal{R}_i(u_i) \neq \emptyset$*;*
  - **(C3*)** *for all* $w \in W$*:* $\mathcal{R}_{Agt}(w) \subseteq \bigcap_{i \in Agt} \mathcal{R}_i(w)$*;*
- $\mathcal{R}_\mathsf{G}$ *and* $\mathcal{R}_\mathsf{H}$ *are binary relations between worlds in* W *such that* $\mathcal{R}_\mathsf{G}$ *is serial and transitive,* $\mathcal{R}_\mathsf{H}$ *is the inverse relation of* $\mathcal{R}_\mathsf{G}$*, and:*
  - **(C4)** *for all* $w, v, u \in W$*: if* $v, u \in \mathcal{R}_\mathsf{G}(w)$ *then* $u \in \mathcal{R}_\mathsf{G}(v)$ *or* $v \in \mathcal{R}_\mathsf{G}(u)$ *or* $u = v$*;*
  - **(C5)** *for all* $w, v, u \in W$*: if* $v, u \in \mathcal{R}_\mathsf{H}(w)$ *then* $u \in \mathcal{R}_\mathsf{H}(v)$ *or* $v \in \mathcal{R}_\mathsf{H}(u)$ *or* $u = v$*;*
  - **(C6)** $\mathcal{R}_\mathsf{G} \circ \mathcal{R}_\Box \subseteq \mathcal{R}_{Agt} \circ \mathcal{R}_\mathsf{G}$*;*
  - **(C7)** *for all* $w \in W$*: if* $v \in \mathcal{R}_\Box(w)$ *then* $v \notin \mathcal{R}_\mathsf{G}(w)$*;*
- $\mathcal{V} : Atm \longrightarrow 2^W$ *is a valuation function for atomic formulas.*

The only difference between the class of temporal Kripke STIT models and superadditive temporal Kripke STIT models is in the Constraint C3*.

Lemma 9. *Let* φ *be a formula in* $\mathcal{L}_{T\text{-}STIT}(Atm, Agt)$*. Then,* φ *is satisfiable in the class of temporal Kripke STIT models if and only if it is satisfiable in the class of superadditive temporal Kripke STIT models.*

Proof. ($\Rightarrow$) The left-to-right direction of the equivalence is obvious as the class of temporal Kripke STIT models is included in the class of superadditive temporal Kripke STIT models.

($\Leftarrow$) As to the right-to-left direction, I am going to show how to transform a superadditive temporal Kripke STIT model into a temporal Kripke STIT model without affecting the satisfiability of a formula. The technique used here is inspired by Vakarelov (1992).

Consider a superadditive temporal Kripke STIT model $M = (W, \mathcal{R}_\Box, \{\mathcal{R}_i | i \in Agt\}, \mathcal{R}_{Agt}, \mathcal{R}_\mathsf{G}, \mathcal{R}_\mathsf{H}, \mathcal{V})$ and a world $\underline{w}$ in M such that $M, \underline{w} \models \varphi$. I am going to define a temporal Kripke STIT model $M' = (W', \mathcal{R}'_\Box, \{\mathcal{R}'_i | i \in Agt\}, \mathcal{R}'_{Agt}, \mathcal{R}'_\mathsf{G}, \mathcal{R}'_\mathsf{H}, \mathcal{V}')$ that satisfies φ.

For every $i \in Agt$, let

$$\Delta_i = \{\mathcal{R}_i(w) | w \in W\}$$

be the partition of W induced by the equivalence class $\mathcal{R}_i$. Elements of Δ_i are called i's choices. Moreover, let

$$\Delta = \{\delta \in \prod\nolimits_{i \in Agt} \Delta_i | \bigcap\nolimits_{i \in Agt} \delta_i \neq \emptyset\}$$

be the set of collective choices. Elements of Δ are denoted by $\delta, \delta', \ldots$. For every $\delta \in \Delta$, δ_i denotes the element in the vector δ corresponding to the agent i. Furthermore, for notational convenience I write δ^w to denote the collective choice in Δ that includes the world w. That is, δ^w is the collective choice in Δ such that $w \in \bigcap_{i \in Agt} \delta_i^w$.

For every $\delta \in \Delta$, let

$$\Gamma_\delta = \{\mathcal{R}_{Agt}(w) | w \in \bigcap\nolimits_{i \in Agt} \delta_i\}$$

be the set of $\mathcal{R}_{Agt}$-equivalence classes that are included in $\bigcap_{i \in Agt} \delta_i$.

For every $\delta \in \Delta$, let

$$k_\delta : [1, \ldots, card(\Gamma_\delta)] \longrightarrow \Gamma_\delta$$

be a *bijection* associating every integer between 1 and $card(\Gamma_\delta)$ to a unique element of Γ_δ. In the sequel I write Γ_δ^n to indicate the element $k_\delta(n)$ for every $1 \leq n \leq card(\Gamma_\delta)$.

For every $w \in W$ let

$$TC_w = \{\delta \in \Delta | \exists v \in \bigcap\nolimits_{i \in Agt} \delta_i \text{ such that } v \in \mathcal{T}(w)\}$$

be the set of collective choices that are temporally related with w, with $\mathcal{T}(w) = \mathcal{R}_{\mathsf{H}}(w) \cup \{w\} \cup \mathcal{R}_{\mathsf{G}}(w)$.

Moreover, let

$$PC_w = \{\delta \in \Delta | \exists v \in \bigcap\nolimits_{i \in Agt} \delta_i \text{ such that } v \in \mathcal{R}_{\mathsf{H}}(w)\}$$

be the set of collective choices that are in the past of w.

Finally, let

$$FC_w = \{\delta \in \Delta | \exists v \in \bigcap\nolimits_{i \in Agt} \delta_i \text{ such that } v \in \mathcal{R}_{\mathsf{G}}(w)\}$$

be the set of collective choices that are in the future of w. Note that:
(A1) if $v \in \mathcal{T}(w)$ then $TC_w = TC_v$;
(A2) if $v \in \mathcal{R}_{\Box}(w)$ then $PC_w = PC_v$ (because of Proposition 4. in Section 2.2);
(A3) if $v \in \mathcal{R}_{Agt}(w)$ then $\delta^w = \delta^v$ and $PC_w = PC_v$ (because of (A2)).

Moreover, let Λ_w be the set of all total functions $f : TC_w \longrightarrow \mathbb{Z}^n$ that satisfy the following constraint:

- for all $\delta \in TC_w$ and for all $\overrightarrow{x} \in \mathbb{Z}^n$, if $f(\delta) = \overrightarrow{x}$ then there exists $v \in \mathcal{T}(w)$ such that $v \in \Gamma_\delta^{\Sigma_{x_i \in \{x_1, \ldots, x_n\}} x_i}$,

where $\mathbb{Z}$ is the set of integers and $\overrightarrow{x} = \langle x_1, \ldots, x_n \rangle$. I write $f_i(\delta)$ to denote the i-element in the vector $f(\delta)$, with $1 \leq i \leq n$.

Because of the Constraint C7, we have that for all $w \in W$ and for all $\overrightarrow{x} \in \mathbb{Z}^n$:

(B) if $v \in \mathcal{T}(w)$ and $v \in \Gamma_{\delta^w}^{\Sigma_{x_i \in \{x_1, \ldots, x_n\}} x_i}$ then $w = v$.

Furthermore, because of Proposition 4 in Section 2.2, we have that for all $w, v \in W$ and for all $\overrightarrow{x} \in \mathbb{Z}^n$:

(C) if $v \in \mathcal{R}_{\Box}(w)$ then there exists $u_1 \in \mathcal{R}_{\mathsf{H}}(w)$ such that $u_1 \in \Gamma_{\delta}^{\Sigma_{x_i \in \{x_1,\ldots,x_n\}} x_i}$ if and only if there exists $u_2 \in \mathcal{R}_{\mathsf{H}}(v)$ such that $u_2 \in \Gamma_{\delta}^{\Sigma_{x_i \in \{x_1,\ldots,x_n\}} x_i}$.

From the previous item (C) it follows that for all $w, v \in W$:

(D) if $f \in \Lambda_w$ and $v \in \mathcal{R}_{\Box}(w)$ then there exists $f' \in \Lambda_v$ such that, for all $\delta \in PC_w$, $f'(\delta) = f(\delta)$.

Moreover, from the previous item (A1), for all $w, v \in W$ we have that:

(E) if $f \in \Lambda_w$ and $v \in \mathcal{T}(w)$ then there exists $f' \in \Lambda_v$ such that, for all $\delta \in TC_w$, $f'(\delta) = f(\delta)$.

I am now able to define the model $M' = (W', \mathcal{R}'_{\Box}, \{\mathcal{R}'_i | i \in Agt\}, \mathcal{R}'_{Agt}, \mathcal{R}'_{\mathsf{G}}, \mathcal{R}'_{\mathsf{H}}, \mathcal{V}')$:

- $W' = \{w_f | w \in W \text{ and } f \in \Lambda_w\}$;
- for all $w_f, v_{f'} \in W'$, $(w_f, v_{f'}) \in \mathcal{R}'_{\Box}$ iff $(w, v) \in \mathcal{R}_{\Box}$ and $f(\delta) = f'(\delta)$ for all $\delta \in PC_w$;
- for all $i \in Agt$ and for all $w_f, v_{f'} \in W'$, $(w_f, v_{f'}) \in \mathcal{R}'_i$ iff $\delta_i^w = \delta_i^v$, $f_i(\delta^w) = f'_i(\delta^v)$ and $f(\delta) = f'(\delta)$ for all $\delta \in PC_w$;
- for all $w_f, v_{f'} \in W'$, $(w_f, v_{f'}) \in \mathcal{R}'_{Agt}$ iff $\delta^w = \delta^v$, $f(\delta^w) = f'(\delta^v)$ and $f(\delta) = f'(\delta)$ for all $\delta \in PC_w$;
- for all $w_f, v_{f'} \in W'$, $(w_f, v_{f'}) \in \mathcal{R}'_{\mathsf{G}}$ iff $(w, v) \in \mathcal{R}_{\mathsf{G}}$ and $f(\delta) = f'(\delta)$ for all $\delta \in TC_w$;
- for all $w_f, v_{f'} \in W'$, $(w_f, v_{f'}) \in \mathcal{R}'_{\mathsf{H}}$ iff $(v_{f'}, w_f) \in \mathcal{R}'_{\mathsf{G}}$;
- for all $p \in Atm$, $\mathcal{V}'(p) = \{w_f \in W' | w \in \mathcal{V}(p)\}$.

It is a routine task to check that the mapping $f : w_f \mapsto w$ defines a *bounded morphism* from M' to M (Blackburn, De Rijke, & Venema, 2001, Definition 2.12). Indeed, it follows from the definitions of $\mathcal{R}'_{\Box}$, $\mathcal{R}'_i$, $\mathcal{R}'_{Agt}$, $\mathcal{R}'_{\mathsf{G}}$ and $\mathcal{R}'_{\mathsf{H}}$ that for all $w_f, v_{f'} \in W'$:

- $(w_f, v_{f'}) \in \mathcal{R}'_i$ implies $(w, v) \in \mathcal{R}_i$;
- $(w_f, v_{f'}) \in \mathcal{R}'_{Agt}$ implies $(w, v) \in \mathcal{R}_{Agt}$;
- $(w_f, v_{f'}) \in \mathcal{R}'_{\Box}$ implies $(w, v) \in \mathcal{R}_{\Box}$;
- $(w_f, v_{f'}) \in \mathcal{R}'_{\mathsf{G}}$ implies $(w, v) \in \mathcal{R}_{\mathsf{G}}$;
- $(w_f, v_{f'}) \in \mathcal{R}'_{\mathsf{H}}$ implies $(w, v) \in \mathcal{R}_{\mathsf{H}}$.

For instance, suppose that $(w_f, v_{f'}) \in \mathcal{R}'_i$. By definition of $\mathcal{R}'_i$ the latter implies that $\delta_i^w = \delta_i^v$ and $w \in \delta_i^w$ and $v \in \delta_i^v$. The latter implies that $(w, v) \in \mathcal{R}_i$.

Now, suppose that $(w_f, v_{f'}) \in \mathcal{R}'_{Agt}$. This implies that $\delta^w = \delta^v$ and $f(\delta^w) = f'(\delta^v)$. By the previous observation (B), it follows that $w, v \in \Gamma_{\delta^w}^{f_1(\delta^w)+\ldots+f_n(\delta^w)}$. Thus, $(w, v) \in \mathcal{R}_{Agt}$.

The other way around we have that for all $w_f \in W'$:

- if $(f(w_f), v) \in \mathcal{R}_{\Box}$ then there is $v_{f'}$ such that $(w_f, v_{f'}) \in \mathcal{R}'_{\Box}$;
- if $(f(w_f), v) \in \mathcal{R}_i$ then there is $v_{f'}$ such that $(w_f, v_{f'}) \in \mathcal{R}'_i$;
- if $(f(w_f), v) \in \mathcal{R}_{Agt}$ then there is $v_{f'}$ such that $(w_f, v_{f'}) \in \mathcal{R}'_{Agt}$;
- if $(f(w_f), v) \in \mathcal{R}_{\mathsf{G}}$ then there is $v_{f'}$ such that $(w_f, v_{f'}) \in \mathcal{R}'_{\mathsf{G}}$;
- if $(f(w_f), v) \in \mathcal{R}_{\mathsf{H}}$ then there is $v_{f'}$ such that $(w_f, v_{f'}) \in \mathcal{R}'_{\mathsf{H}}$.

Let me prove the previous items. The first item is trivial and I do not prove it here.

Suppose that $w_f \in W'$ and $(f(w_f), v) \in \mathcal{R}_i$. This implies that $(w, v) \in \mathcal{R}_i$. Therefore, $\mathcal{R}_i(w) = \mathcal{R}_i(v)$ as $\mathcal{R}_i$ is an equivalence relation. Hence, $\delta_i^v = \delta_i^w$. It follows that $(w_f, v_{f'}) \in \mathcal{R}'_i$ where the function f' is defined as follows: (i) $f'_i(\delta^v) = f_i(\delta^w)$ and all $f'_j(\delta^v)$ with $j \neq i$ are such that $v \in \Gamma_{\delta^v}^{f'_1(\delta^v)+\ldots+f'_n(\delta^v)}$; (ii) for all $\delta \in PC_w$, $f'(\delta) = f(\delta)$; (iii) for all $\delta \in FC_v$, $f'(\delta)$ is such that there exists $u \in \mathcal{R}_{\mathsf{G}}(v)$ with $u \in \Gamma_{\delta}^{f'_1(\delta)+\ldots+f'_n(\delta)}$. This function f' is guaranteed to exist because of the observation (D) above and the fact that $\mathcal{R}_i \subseteq \mathcal{R}_{\Box}$.

Suppose that $w_f \in W'$ and $(f(w_f), v) \in \mathcal{R}_{Agt}$. This implies that $(w, v) \in \mathcal{R}_{Agt}$. Therefore, $\mathcal{R}_{Agt}(w) = \mathcal{R}_{Agt}(v)$ as $\mathcal{R}_{Agt}$ is an equivalence relation. Hence, $\delta^v = \delta^w$. It follows that $(w_f, v_{f'}) \in \mathcal{R}'_{Agt}$ where the function f' is defined as follows: (i) $f'(\delta^v) = f(\delta^w)$; (ii) for all $\delta \in PC_w$, $f'(\delta) = f(\delta)$; (iii) for all $\delta \in FC_v$, $f'(\delta)$ is such that there exists $u \in \mathcal{R}_{\mathsf{G}}(v)$ with $u \in \Gamma_{\delta}^{f'_1(\delta)+\ldots+f'_n(\delta)}$. This function f' is guaranteed to exist because of the observation (D) above and the fact that $\mathcal{R}_{Agt} \subseteq \mathcal{R}_{\Box}$.

Suppose that $w_f \in W'$ and $(f(w_f), v) \in \mathcal{R}_{\mathsf{G}}$. This implies that $(w, v) \in \mathcal{R}_{\mathsf{G}}$. It follows that $(w_f, v_{f'}) \in \mathcal{R}'_{\mathsf{G}}$ where the function f' is such that, for all $\delta \in TC_w$, $f'(\delta) = f(\delta)$. This function f' is guaranteed to exist because of the observation (E) above. Thus, $(w_f, v_{f'}) \in \mathcal{R}'_{\mathsf{G}}$.

Finally, suppose that $w_f \in W'$ and $(f(w_f), v) \in \mathcal{R}_{\mathsf{H}}$. This implies that $(w, v) \in \mathcal{R}_{\mathsf{H}}$ which is equivalent to $(v, w) \in \mathcal{R}_{\mathsf{G}}$. It follows that $(v_{f'}, w_f) \in \mathcal{R}'_{\mathsf{G}}$ where the function f' is such that, for all $\delta \in TC_v$, $f(\delta) = f'(\delta)$. This function f' is guaranteed to exist because of the observation (E) above. Thus, $(w_f, v_{f'}) \in \mathcal{R}'_{\mathsf{H}}$.

As f is a bounded morphism it holds that $M, \underline{w} \models \varphi$ if and only if $M', \underline{w}_f \models \varphi$. Thus, $M', \underline{w}_f \models \varphi$ for all $\underline{w}_f \in W'$.

In order to terminate the proof we also need to be sure that M' is a temporal Kripke STIT model in the sense of Definition 1. It is a routine task to show that $\mathcal{R}'_i$, $\mathcal{R}'_{Agt}$ and $\mathcal{R}'_{\Box}$ are equivalence relations, that the model transformation preserves transitivity and seriality for the relation $\mathcal{R}'_{\mathsf{G}}$, and that the model M' satisfies the Constraints C1, C2, C4, C5 and C7. Let me prove that it also satisfies Constraints C3 and C6.

$v_{f'} \in \bigcap_{i \in Agt} \mathcal{R}'_i(w_f)$ if and only if (i) $\delta_i^w = \delta_i^v$ and $f_i(\delta^w) = f'_i(\delta^v)$ for all $i \in Agt$; and (ii) $f(\delta) = f'(\delta)$ for all $\delta \in PC_w$. The latter is equivalent to $\delta^w = \delta^v$, $f(\delta^w) = f'(\delta^v)$ and $f(\delta) = f'(\delta)$ for all $\delta \in PC_w$, which in turn is equivalent to $v_{f'} \in \mathcal{R}'_{Agt}(w_f)$. This proves that the model M' satisfies the Constraint C3.

Now suppose that $v_{f'} \in \mathcal{R}'_{\mathsf{G}}(w_f)$ and $u_{f''} \in \mathcal{R}'_{\Box}(v_{f'})$. It follows that: (i) $f(\delta) = f'(\delta)$ for all $\delta \in TC_w$; and (ii) $f'(\delta) = f''(\delta)$ for all $\delta \in PC_v$. Moreover, since the model M satisfies the Constraint C6, there exists $z \in W$ such that $z \in \mathcal{R}_{Agt}(w)$ and $u \in \mathcal{R}_{\mathsf{G}}(z)$. Let $f''' \in \Lambda_z$ be the function that for all $\delta \in TC_z$, $f''(\delta) = f'''(\delta)$. It follows that $u_{f''} \in \mathcal{R}'_{\mathsf{G}}(z_{f'''})$. I am going to prove that $z_{f'''} \in \mathcal{R}'_{Agt}(w_f)$. From the previous conditions (i) and (ii) it follows that $f(\delta) = f''(\delta)$ for all $\delta \in TC_w \cap PC_v$. Since $v \in \mathcal{R}_{\mathsf{G}}(w)$, $TC_w \cap PC_v = PC_v$. Hence, $f(\delta) = f''(\delta)$ for all $\delta \in PC_v$. Furthermore, $f(\delta) = f''(\delta)$ for all $\delta \in PC_w \cup \{\delta^w\}$, because $v \in \mathcal{R}_{\mathsf{G}}(w)$. From the latter and the fact that $f''(\delta^w) = f'''(\delta^z)$ for all $\delta \in TC_z$, it follows that $f(\delta^w) = f'''(\delta^z)$ for all $\delta \in TC_z \cap (PC_w \cup \{\delta^w\})$. Since $z \in \mathcal{R}_{Agt}(w)$, by the previous observation (A3), it follows that $\delta^w = \delta^z$ and $f(\delta^w) = f'''(\delta^z)$ for all $\delta \in PC_w \cup \{\delta^w\}$. Hence, $f(\delta^w) = f'''(\delta^z)$ and $f(\delta) = f'''(\delta)$ for all $\delta \in PC_w$. Thus, $z_{f'''} \in \mathcal{R}'_{\mathsf{G}}(w_{f'})$. This proves that the model M' satisfies the Constraint C6. ■

The following lemma provides an axiomatisation result for the class of superadditive temporal Kripke STIT models.

Lemma 10. *The set of T-STIT formulas that are valid in the class of superadditive temporal Kripke STIT models is completely axiomatised by the principles given in Figure 2.*

Proof. In order to prove Lemma 10, I use a technique similar to the one used by Gabbay, Hodkinson and Reynolds (1994) for proving completeness of linear temporal logic and of logic of historical necessity by means of a slightly different variant of the *irreflexivity rule* **(IRR)**.

Let me start with the following standard definition of canonical model for T-STIT.

Definition 11 (Canonical model for T-STIT). *The canonical model M^c for T-STIT is the tuple $M^c = (W^c, \mathcal{R}^c_{\Box}, \{\mathcal{R}^c_i | i \in Agt\}, \mathcal{R}^c_{Agt}, \mathcal{R}^c_{\mathsf{G}}, \mathcal{R}^c_{\mathsf{H}}, \mathcal{V}^c)$ where:*

- *W^c is the set of all maximal consistent sets of T-STIT formulas (MCSs);*
- *$\mathcal{R}^c_{\Box}, \mathcal{R}^c_i, \mathcal{R}^c_{Agt}, \mathcal{R}^c_{\mathsf{G}}$ and $\mathcal{R}^c_{\mathsf{H}}$ are respectively the canonical relations for $\Box$, $[i]$, $[Agt]$, G and H, that is:*
  - *for all $\Gamma, \Delta \in W^c$, $(\Gamma, \Delta) \in \mathcal{R}^c_{\Box}$ iff for all formulas ψ, $\psi \in \Delta$ implies $\Diamond\psi \in \Gamma$;*
  - *for all $\Gamma, \Delta \in W^c$ and for all $i \in Agt$, $(\Gamma, \Delta) \in \mathcal{R}^c_i$ iff for all formulas ψ, $\psi \in \Delta$ implies $\langle i\rangle\psi \in \Gamma$;*
  - *for all $\Gamma, \Delta \in W^c$, $(\Gamma, \Delta) \in \mathcal{R}^c_{Agt}$ iff for all formulas ψ, $\psi \in \Delta$ implies $\langle Agt\rangle\psi \in \Gamma$;*
  - *for all $\Gamma, \Delta \in W^c$, $(\Gamma, \Delta) \in \mathcal{R}^c_{\mathsf{G}}$ iff for all formulas ψ, $\psi \in \Delta$ implies $\mathsf{F}\psi \in \Gamma$;*
  - *for all $\Gamma, \Delta \in W^c$, $(\Gamma, \Delta) \in \mathcal{R}^c_{\mathsf{H}}$ iff for all formulas ψ, $\psi \in \Delta$ implies $\mathsf{P}\psi \in \Gamma$;*
- *$\mathcal{V}^c$ is the valuation defined by $\mathcal{V}^c(p) = \{\Gamma \in W^c | p \in \Gamma\}$ for all $p \in Atm$.*

Furthermore, let me introduce the notion of 'diamond'-saturation for maximal consistent sets of T-STIT formulas.

Definition 12 (Diamond saturated set of MCSs). *Given a set X of maximal consistent sets of T-STIT formulas, I say that X is a* diamond saturated *set of MCSs if and only if for each $\Gamma \in X$ the following five conditions are satisfied:*

- *for each formula $\Diamond\varphi \in \Gamma$ there is $\Delta \in X$ such that $(\Gamma, \Delta) \in \mathcal{R}^c_{\Box}$ and $\varphi \in \Delta$;*
- *for each formula $\langle i\rangle\varphi \in \Gamma$ there is $\Delta \in X$ such that $(\Gamma, \Delta) \in \mathcal{R}^c_i$ and $\varphi \in \Delta$;*
- *for each formula $\langle Agt\rangle\varphi \in \Gamma$ there is $\Delta \in X$ such that $(\Gamma, \Delta) \in \mathcal{R}^c_{Agt}$ and $\varphi \in \Delta$;*
- *for each formula $\mathsf{F}\varphi \in \Gamma$ there is $\Delta \in X$ such that $(\Gamma, \Delta) \in \mathcal{R}^c_{\mathsf{G}}$ and $\varphi \in \Delta$;*
- *for each formula $\mathsf{P}\varphi \in \Gamma$ there is $\Delta \in X$ such that $(\Gamma, \Delta) \in \mathcal{R}^c_{\mathsf{H}}$ and $\varphi \in \Delta$.*

The following truth lemma is provable in the standard way by induction on φ (see Blackburn et al., 2001, Lemma 4.70).

Lemma 13 (Truth Lemma). *Let X be a diamond saturated set of MCSs. Then for any $\Gamma \in X$ and for any formula φ, $M^c|_X, \Gamma \models \varphi$ if and only if $\varphi \in \Gamma$, where $M^c|_X$ is the submodel of the canonical model M^c induced by X.*

The next step in the proof consists in defining the notion of IRR theory.

For every atom p, let

$$name(p) \stackrel{\text{def}}{=} \Box\neg p \wedge \Box(\mathsf{G}p \wedge \mathsf{H}p).$$

The formula $name(p)$ acts as a sort of 'name' for a given world that ensures irreflexivity of the canonical relation $\mathcal{R}^c_{\mathsf{G}}$ for the future tense operator G.

Furthermore, let

$$\Theta = \{\blacklozenge_1(\psi_1 \wedge \blacklozenge_2(\psi_2 \wedge \ldots \wedge \blacklozenge_n\psi_n)\ldots) : \blacklozenge_1, \ldots, \blacklozenge_n \in \{\Diamond, \langle 1\rangle, \ldots, \langle n\rangle, \langle Agt\rangle, \mathsf{F}, \mathsf{P}\} \text{ and } \psi_1, \ldots, \psi_n \in \mathcal{L}_{\text{T-STIT}}(Atm, Agt)\}.$$

Finally, for every $\varphi = \blacklozenge_1(\psi_1 \wedge \blacklozenge_2(\psi_2 \wedge \ldots \wedge \blacklozenge_n\psi_n)\ldots) \in \Theta$ and for every atom p, let

$$\varphi(p) \stackrel{\text{def}}{=} \blacklozenge_1(\psi_1 \wedge \blacklozenge_2(\psi_2 \wedge \ldots \wedge \blacklozenge_n(\psi_n \wedge name(p)))\ldots).$$

A formula of the form $\varphi(p)$ is also used as a 'name' for a given world. In particular, formulas of the form $\varphi(p)$ provide 'names' for worlds that are reachable by any zig-zagging sequence of $\Diamond$s, $\langle i\rangle$s, $\langle Agt\rangle$s, Fs and Ps.

Definition 14 (IRR theory). *An IRR theory is a maximal consistent set of T-STIT formulas* Γ *such that:*

- *for some* p, $name(p) \in \Gamma$*;*
- *if* $\varphi \in \Gamma \cap \Theta$ *then, for some atom* p, $\varphi(p) \in \Gamma$.

The set of all IRR theories is denoted by **IrrTh**.

The following lemma highlights that every consistent T-STIT formula is included in at least one IRR theory. The rule of inference **(IRR)** becomes crucial at this point of the proof.

Lemma 15. *Let* φ *be a consistent T-STIT formula. Then, there exists an IRR theory* Γ *such that* $\varphi \in \Gamma$.

(Sketch). The lemma is proved analogously to Lemma 6.2.4 in Gabbay et al. (1994, Chapter 6). Since the set *Atm* of propositional atoms is infinite, there exists an infinite number of atomic formulas p not occurring in φ. Therefore, thanks to the **(IRR)** rule, it is straightforward to build step-by-step an IRR theory Γ containing φ. ■

The following lemma highlights that the set of all IRR theories satisfies the condition of 'diamond'-saturation defined above (Definition 12).

Lemma 16 (Existence Lemma)**.**
Let Γ *be an IRR theory. Then:*

- *if* $\Diamond\varphi \in \Gamma$*, then there is an IRR theory* Δ *such that* $(\Gamma, \Delta) \in \mathcal{R}^c_{\Box}$ *and* $\varphi \in \Delta$*;*
- *if* $\langle i\rangle\varphi \in \Gamma$*, then there is an IRR theory* Δ *such that* $(\Gamma, \Delta) \in \mathcal{R}^c_i$ *and* $\varphi \in \Delta$*;*
- *if* $\langle Agt\rangle\varphi \in \Gamma$*, then there is an IRR theory* Δ *such that* $(\Gamma, \Delta) \in \mathcal{R}^c_{Agt}$ *and* $\varphi \in \Delta$*;*
- *if* $\mathsf{F}\varphi \in \Gamma$*, then there is an IRR theory* Δ *such that* $(\Gamma, \Delta) \in \mathcal{R}^c_{\mathsf{G}}$ *and* $\varphi \in \Delta$*;*
- *if* $\mathsf{P}\varphi \in \Gamma$*, then there is an IRR theory* Δ *such that* $(\Gamma, \Delta) \in \mathcal{R}^c_{\mathsf{H}}$ *and* $\varphi \in \Delta$.

Proof. I only prove the first item. The other items can be proved analogously.

Suppose $\Diamond\varphi \in \Gamma$. Since Γ is an IRR theory, it follows that $\Diamond(\varphi \wedge name(p)) \in \Gamma$ for some atom p. Let $\Delta_0 = \{\varphi \wedge name(p)\} \cup \{\psi | \Box\psi \in \Gamma\}$. I am going to show that Δ_0 is consistent by reductio ad absurdum. Suppose Δ_0 is not consistent. Then, for some $\psi_1, \ldots, \psi_k \in \{\psi | \Box\psi \in \Gamma\}$, we have:

$$\vdash (\psi_1 \wedge \ldots \wedge \psi_k) \rightarrow \neg(\varphi \wedge name(p)),$$

hence

$$\vdash \Box(\psi_1 \wedge \ldots \wedge \psi_k) \rightarrow \Box\neg(\varphi \wedge name(p)),$$

and hence $\Box\neg(\varphi\wedge name(p)) \in \Gamma$ (because $\Box(\psi_1\wedge\ldots\wedge\psi_k) \in \Gamma$). But this is in contradiction with $\Diamond(\varphi \wedge name(p)) \in \Gamma$.

Now, I turn to define an increasing sequence of consistent extensions $(\Delta_n)_{n\geq 0}$ such that $\Delta_n \subseteq \Delta_{n+1}$ for all n.

Assume $\Delta_n \subseteq \Delta_{n+1}$ has been defined and is consistent. Then, either $\Delta_n \cup \{\chi_n\}$ is consistent or $\Delta_n \cup \{\neg\chi_n\}$ is consistent.

Case 1: $\Delta_n \cup \{\neg\chi_n\}$ *consistent*. Take $\Delta_{n+1} = \Delta_n \cup \{\neg\chi_n\}$.

Case 2a: $\Delta_n \cup \{\neg\chi_n\}$ *not consistent*, $\Delta_n \cup \{\chi_n\}$ *consistent* and $\chi_n \notin \Theta$. Take $\Delta_{n+1} = \Delta_n \cup \{\chi_n\}$.

Case 2b: $\Delta_n \cup \{\neg\chi_n\}$ *not consistent*, $\Delta_n \cup \{\chi_n\}$ *consistent* and $\chi_n \in \Theta$. We have that:

$$\Diamond((\varphi \wedge name(p)) \wedge \bigwedge\nolimits_{\psi\in\Delta_n\setminus\Delta_0} \psi \wedge \chi_n) \in \Gamma.$$

For, otherwise,

$$\Box(((\varphi \wedge name(p)) \wedge \bigwedge\nolimits_{\psi\in\Delta_n\setminus\Delta_0} \psi) \rightarrow \neg\chi_n) \in \Gamma$$

and hence, by definition of Δ_0, $((\varphi \wedge name(p)) \wedge \bigwedge_{\psi\in\Delta_n\setminus\Delta_0} \psi) \rightarrow \neg\chi_n \in \Delta_n$.

Hence, since $((\varphi \wedge name(p)) \wedge \bigwedge_{\psi\in\Delta_n\setminus\Delta_0} \psi) \in \Delta_n$, $\neg\chi_n \in \Delta_n$.

This is in contradiction with the fact that $\Delta_n \cup \{\chi_n\}$ is consistent.

Thus, since Γ is an IRR theory, for some atom q, we have

$$\Diamond((\varphi \wedge name(p)) \wedge \bigwedge\nolimits_{\psi\in\Delta_n\setminus\Delta_0} \psi \wedge \chi_n(q)) \in \Gamma.$$

It follows that $\Delta_n \cup \{\chi_n\} \cup \{\chi_n(q)\}$ is consistent. For, otherwise, for some $\psi_1, \ldots, \psi_k \in \{\psi \mid \Box\psi \in \Gamma\}$, we have:

$$\vdash (\psi_1 \wedge \ldots \wedge \psi_k) \rightarrow \neg((\varphi \wedge name(p)) \wedge \bigwedge\nolimits_{\psi\in\Delta_n\setminus\Delta_0} \psi \wedge \chi_n(q)),$$

hence

$$\vdash \Box(\psi_1 \wedge \ldots \wedge \psi_k) \rightarrow \Box\neg((\varphi \wedge name(p)) \wedge \bigwedge\nolimits_{\psi\in\Delta_n\setminus\Delta_0} \psi \wedge \chi_n(q)),$$

and hence $\Box\neg((\varphi\wedge name(p))\wedge\bigwedge_{\psi\in\Delta_n\setminus\Delta_0} \psi\wedge\chi_n(q)) \in \Gamma$ (because $\Box(\psi_1\wedge\ldots\wedge\psi_k) \in \Gamma$).

But this is in contradiction with $\Diamond((\varphi \wedge name(p)) \wedge \bigwedge_{\psi\in\Delta_n\setminus\Delta_0} \psi \wedge \chi_n(q)) \in \Gamma$.

Let $\Delta_{n+1} = \Delta_n \cup \{\chi_n\} \cup \{\chi_n(q)\}$.

Thus, the sequence $(\Delta_n)_{n\geq 0}$ is defined and, clearly, $\Delta = \bigcup_{n\geq 0} \Delta_n$ is the desired IRR theory such that $(\Gamma, \Delta) \in \mathcal{R}^c_{\Box}$ and $\varphi \in \Delta$. ■

The last part of the proof consists in proving that the submodel of the canonical model for T-STIT including all IRR theories, i.e., the model $M^c|_{\mathbf{IrrTh}}$, is indeed a superadditive temporal Kripke STIT model. The following two propositions ensure that $M^c|_{\mathbf{IrrTh}}$ satisfies the Constraints C2 and C6.

Proposition 17. *Let* $\Gamma_1, \ldots, \Gamma_n$ *be IRR theories such that for all* $1 \leq i, j \leq n$, $(\Gamma_i, \Gamma_j) \in \mathcal{R}^c_{\Box}$. *Then, there exists an IRR theory* Δ *such that* $(\Gamma_1, \Delta) \in \mathcal{R}^c_1, \ldots, (\Gamma_n, \Delta) \in \mathcal{R}^c_n$.

Proof. In order to simplify the exposition, let us assume that $n = 2$. The general case for any arbitrary n can be proved analogously.

Suppose Γ_1 and Γ_2 are IRR theories such that for all $1 \leq i, j \leq 2$, $(\Gamma_i, \Gamma_j) \in \mathcal{R}^c_{\Box}$. I am going to show that there exists an IRR theory Δ such that $(\Gamma_1, \Delta) \in \mathcal{R}^c_1$ and $(\Gamma_2, \Delta) \in \mathcal{R}^c_2$.

Since Γ_1 and Γ_2 are IRR theories, there exists $name(p_1) \in \Gamma_1$ and $name(p_2) \in \Gamma_2$ for some atoms p_1, p_2.

Let $\Delta_0 = \{name(p_1)\} \cup \{name(p_2)\} \cup \{\psi | [1]\psi \in \Gamma_1\} \cup \{\psi | [2]\psi \in \Gamma_2\}$. I am going to show that Δ_0 is consistent by reductio ad absurdum.

Suppose Δ_0 is not consistent. Then, for some $\psi_1, \ldots, \psi_k \in \Delta_0$, we have:

$$\vdash (\psi_1 \wedge \ldots \wedge \psi_k) \rightarrow \bot.$$

Let Y denote the set $\{\psi_1, \ldots, \psi_k\}$. Moreover, let $Y_1 = \{\psi \in Y | [1]\psi \in \Gamma_1$ or $\psi = name(p_1)\}$ and $Y_2 = \{\psi \in Y | [2]\psi \in \Gamma_2$ or $\psi = name(p_2)\}$. Clearly, $Y = Y_1 \cup Y_2$ and:

(A) $[1] \bigwedge_{\psi \in Y_1} \psi \in \Gamma_1$.

The previous item (A) follows from these two facts: (i) for every $\psi \in Y_1$ such that $\psi \neq name(p_1)$, $[1]\psi \in \Gamma_1$ (because ψ is of the form $[1]\chi$ and $[1]\chi \rightarrow [1][1]\chi$ is a T-STIT theorem); and (ii) $[1]name(p_1) \in \Gamma_1$ (because $name(p_1) \rightarrow [1]name(p_1)$ is a T-STIT theorem). Likewise, we have:

(B) $[2] \bigwedge_{\psi \in Y_2} \psi \in \Gamma_2$.

From the definition of the accessibility relation $\mathcal{R}^c_{\Box}$ and the fact that $(\Gamma_1, \Gamma_2) \in \mathcal{R}^c_{\Box}$, it follows that $\Diamond[2] \bigwedge_{\psi \in Y_2} \psi \in \Gamma_1$. Moreover, by the T-STIT theorem $\vdash [1]\psi \rightarrow \Diamond[1]\psi$, we have $\Diamond[1] \bigwedge_{\psi \in Y_1} \psi \in \Gamma_1$. By the Axiom **(AIA)**, it follows that $\Diamond([1] \bigwedge_{\psi \in Y_1} \psi \wedge [2] \bigwedge_{\psi \in Y_2} \psi) \in \Gamma_1$. Hence, by Lemma 16, there is an IRR theory Δ such that $(\Gamma_1, \Delta) \in \mathcal{R}^c_{\Box}$ and $([1] \bigwedge_{\psi \in Y_1} \psi \wedge [2] \bigwedge_{\psi \in Y_2} \psi) \in \Delta$. Hence, by Axiom T for $[i]$, there is an IRR theory Δ such that $(\Gamma_1, \Delta) \in \mathcal{R}^c_{\Box}$ and $(\bigwedge_{\psi \in Y_1} \psi \wedge \bigwedge_{\psi \in Y_2} \psi) \in \Delta$. The latter implies that $\bigwedge_{\psi \in Y} \psi$ is consistent. Hence, $\nvdash (\psi_1 \wedge \ldots \wedge \psi_k) \rightarrow \bot$.

The rest of the proof proceeds in the same way as the proof of Lemma 16. Specifically, I turn to define an increasing sequence of consistent extensions $(\Delta_n)_{n \geq 0}$ such that $\Delta_n \subseteq \Delta_{n+1}$ for all n.

Assume $\Delta_n \subseteq \Delta_{n+1}$ has been defined and is consistent. Then, either $\Delta_n \cup \{\chi_n\}$ is consistent or $\Delta_n \cup \{\neg\chi_n\}$ is consistent.

Case 1: $\Delta_n \cup \{\neg\chi_n\}$ *consistent*. Take $\Delta_{n+1} = \Delta_n \cup \{\neg\chi_n\}$.

Case 2a: $\Delta_n \cup \{\neg\chi_n\}$ *not consistent*, $\Delta_n \cup \{\chi_n\}$ *consistent and* $\chi_n \notin \Theta$. Take $\Delta_{n+1} = \Delta_n \cup \{\chi_n\}$.

Case 2b: $\Delta_n \cup \{\neg\chi_n\}$ *not consistent*, $\Delta_n \cup \{\chi_n\}$ *consistent and* $\chi_n \in \Theta$. We have that:

$$\langle 1 \rangle (name(p_1) \wedge \bigwedge\nolimits_{\psi \in \Delta_n \setminus \Delta_0} \psi \wedge \chi_n) \in \Gamma_1.$$

For, otherwise,

$$[1]((name(p_1) \wedge \bigwedge\nolimits_{\psi \in \Delta_n \setminus \Delta_0} \psi) \rightarrow \neg\chi_n) \in \Gamma_1$$

and hence, by definition of Δ_0, $(name(p_1) \wedge \bigwedge_{\psi \in \Delta_n \setminus \Delta_0} \psi) \rightarrow \neg\chi_n \in \Delta_n$. Hence, since $(name(p_1) \wedge \bigwedge_{\psi \in \Delta_n \setminus \Delta_0} \psi) \in \Delta_n$, $\neg\chi_n \in \Delta_n$. This is in contradiction with the fact that $\Delta_n \cup \{\chi_n\}$ is consistent.

Therefore, for some atom q,

$$\langle 1\rangle(name(p_1) \wedge \bigwedge\nolimits_{\psi\in\Delta_n\setminus\Delta_0} \psi \wedge \chi_n(q)) \in \Gamma_1.$$

It follows that $\Delta_n \cup \{\chi_n\} \cup \{\chi_n(q)\}$ is consistent. For, otherwise, for some $\psi_1, \ldots, \psi_k \in \{\psi | [1]\psi \in \Gamma_1\}$, we have:

$$\vdash (\psi_1 \wedge \ldots \wedge \psi_k) \rightarrow \neg(name(p_1) \wedge \bigwedge\nolimits_{\psi\in\Delta_n\setminus\Delta_0} \psi \wedge \chi_n(q)),$$

hence

$$\vdash [1](\psi_1 \wedge \ldots \wedge \psi_k) \rightarrow [1]\neg(name(p) \wedge \bigwedge\nolimits_{\psi\in\Delta_n\setminus\Delta_0} \psi \wedge \chi_n(q)),$$

and hence $[1]\neg(name(p_1)\wedge\bigwedge_{\psi\in\Delta_n\setminus\Delta_0} \psi\wedge\chi_n(q)) \in \Gamma_1$ (because $[1](\psi_1\wedge\ldots\wedge\psi_k) \in \Gamma_1$). But this is in contradiction with $\langle 1\rangle(name(p_1) \wedge \bigwedge_{\psi\in\Delta_n\setminus\Delta_0} \psi \wedge \chi_n(q)) \in \Gamma_1$.

Let $\Delta_{n+1} = \Delta_n \cup \{\chi_n\} \cup \{\chi_n(q)\}$.

Thus, the sequence $(\Delta_n)_{n\geq 0}$ is defined and, clearly, $\Delta = \bigcup_{n\geq 0} \Delta_n$ is the desired IRR theory such that $(\Gamma_1, \Delta) \in \mathcal{R}^c_1$ and $(\Gamma_2, \Delta) \in \mathcal{R}^c_2$. ■

Proposition 18. *Let* Δ, Γ *and* Γ' *be IRR theories such that* $(\Delta, \Gamma) \in \mathcal{R}^c_{\mathsf{G}}$ *and* $(\Gamma, \Gamma') \in \mathcal{R}^c_{\Box}$. *Then, there exists an IRR theory* Δ' *such that* $(\Delta, \Delta') \in \mathcal{R}^c_{Agt}$ *and* $(\Delta', \Gamma') \in \mathcal{R}^c_{\mathsf{G}}$.

Proof. Suppose Δ, Γ and Γ' are IRR theories such that $(\Delta, \Gamma) \in \mathcal{R}^c_{\mathsf{G}}$ and $(\Gamma, \Gamma') \in \mathcal{R}^c_{\Box}$. It follows that $name(p) \in \Delta$ for some atom p. Hence, by definition of $\mathcal{R}^c_{\mathsf{H}}$, $\mathsf{P}name(p) \in \Gamma$. By definition of $\mathcal{R}^c_{\Box}$, it follows that $\Diamond\mathsf{P}name(p) \in \Gamma'$. We have the following T-STIT theorem:

$$\vdash \Diamond\mathsf{P}\varphi \rightarrow \mathsf{P}\langle Agt\rangle\varphi. \tag{1}$$

The following is the Hilbert-style proof:

1. $\vdash \Diamond\mathsf{H}[Agt]\varphi \rightarrow \mathsf{HF}\Diamond\mathsf{H}[Agt]\varphi$.
   By Axiom($\mathbf{Conv}_{\mathsf{H},\mathsf{G}}$).
2. $\vdash \mathsf{HF}\Diamond\mathsf{H}[Agt]\varphi \rightarrow \mathsf{H}\langle Agt\rangle\mathsf{FH}[Agt]\varphi$.
   By Axiom(**NCUH**), Axiom K and necessitation for H.
3. $\vdash \mathsf{H}\langle Agt\rangle\mathsf{FH}[Agt]\varphi \rightarrow \mathsf{H}\langle Agt\rangle[Agt]\varphi$.
   By Axiom($\mathbf{Conv}_{\mathsf{H},\mathsf{G}}$), Axiom K and necessitation for H and agt.
4. $\vdash \mathsf{H}\langle Agt\rangle[Agt]\varphi \rightarrow \mathsf{H}\varphi$.
   By T-STIT theorem $\vdash \langle Agt\rangle[Agt]\varphi \rightarrow \varphi$, Axiom K and necessitation for H.
5. $\vdash \mathsf{P}\varphi \rightarrow \Box\mathsf{P}\langle Agt\rangle\varphi$.
   From 1–4.
6. $\vdash \Diamond\mathsf{P}\varphi \rightarrow \Diamond\Box\mathsf{P}\langle Agt\rangle\varphi$.
   By 5, Axiom K and necessitation for $\Box$.
7. $\vdash \Diamond\Box\mathsf{P}\langle Agt\rangle\varphi \rightarrow \Box\mathsf{P}\langle Agt\rangle\varphi$.
   By T-STIT theorem $\vdash \Diamond\Box\varphi \rightarrow \Box\varphi$.
8. $\vdash \Box\mathsf{P}\langle Agt\rangle\varphi \rightarrow \mathsf{P}\langle Agt\rangle\varphi$.
   By Axiom T for $\Box$.
9. $\vdash \Diamond\mathsf{P}\varphi \rightarrow \mathsf{P}\langle Agt\rangle\varphi$.
   From 6–8.

Thus, by the previous T-STIT theorem 1, $\mathsf{P}\langle Agt\rangle name(p) \in \Gamma'$. Hence, by Lemma 16, there exists an IRR theory Δ' such that $\langle Agt\rangle name(p) \in \Delta'$ and $(\Delta', \Gamma') \in \mathcal{R}^c_{\mathsf{G}}$. By the T-STIT theorem $\vdash \langle Agt\rangle name(p) \rightarrow name(p)$, the latter implies that there exists an IRR theory Δ' such that $name(p) \in \Delta'$ and $(\Delta', \Gamma') \in \mathcal{R}^c_{\mathsf{G}}$. I am going to show that $(\Delta, \Delta') \in \mathcal{R}^c_{Agt}$.

Assume $[Agt]\psi \in \Delta$. We have the following T-STIT theorem:

$$\vdash (name(p) \wedge [Agt]\psi) \rightarrow \mathsf{G}\Box\mathsf{H}(name(p) \rightarrow \psi). \tag{2}$$

The following is the Hilbert-style proof:

1. $\vdash p \rightarrow \neg name(p)$.
   By definition of $name(p)$ and Axiom T for $\Box$.
2. $\vdash name(p) \rightarrow [Agt](\mathsf{G}p \wedge \mathsf{H}p)$.
   By definition of $name(p)$, Axiom $(\Box \rightarrow i)$ and Axiom $(i \rightarrow Agt)$**.**
3. $\vdash [Agt](\mathsf{G}p \wedge \mathsf{H}p) \rightarrow ([Agt]\mathsf{G}p \wedge [Agt]\mathsf{H}p)$.
   By Axiom K for $[Agt]$.
4. $\vdash ([Agt]\mathsf{G}p \wedge [Agt]\mathsf{H}p) \rightarrow ([Agt]\mathsf{G}\neg name(p) \wedge [Agt]\mathsf{H}\neg name(p))$.
   By 1, Axiom K and necessitation for $[Agt]$, G and H.
5. $\vdash (name(p) \wedge [Agt]\psi) \rightarrow ([Agt]\psi \wedge [Agt]\mathsf{G}\neg name(p) \wedge [Agt]\mathsf{H}\neg name(p))$.
   From 2–4.
6. $\vdash ([Agt]\psi \wedge [Agt]\mathsf{G}\neg name(p) \wedge [Agt]\mathsf{H}\neg name(p)) \rightarrow$
   $([Agt](\neg name(p) \vee \psi) \wedge [Agt]\mathsf{G}(\neg name(p) \vee \psi) \wedge [Agt]\mathsf{H}(\neg name(p) \vee \psi))$.
   By Axiom K and necessitation for $[Agt]$, G and H.
7. $\vdash ([Agt](\neg name(p) \vee \psi) \wedge [Agt]\mathsf{G}(\neg name(p) \vee \psi) \wedge [Agt]\mathsf{H}(\neg name(p) \vee \psi)) \rightarrow$
   $[Agt]((\neg name(p) \vee \psi) \wedge \mathsf{G}(\neg name(p) \vee \psi) \wedge \mathsf{H}(\neg name(p) \vee \psi))$.
   By Axiom K for $[Agt]$.
8. $\vdash [Agt]((\neg name(p) \vee \psi) \wedge \mathsf{G}(\neg name(p) \vee \psi) \wedge \mathsf{H}(\neg name(p) \vee \psi)) \rightarrow$
   $[Agt]\mathsf{G}\mathsf{H}(\neg name(p) \vee \psi)$.
   By Axiom $\mathbf{(Connected_H)}$, Axiom K and necessitation for $[Agt]$.
9. $\vdash [Agt]\mathsf{G}\mathsf{H}(\neg name(p) \vee \psi) \rightarrow \mathsf{G}\Box\mathsf{H}(\neg name(p) \vee \psi)$.
   By Axiom **(NCUH)**.
10. $\vdash \mathsf{G}\Box\mathsf{H}(\neg name(p) \vee \psi) \rightarrow \mathsf{G}\Box\mathsf{H}(name(p) \rightarrow \psi)$.
11. $\vdash (name(p) \wedge [Agt]\psi) \rightarrow \mathsf{G}\Box\mathsf{H}(name(p) \rightarrow \psi)$.
    From 5–10.

Since $[Agt]\psi \in \Delta$ and $name(p) \in \Delta$, by the previous T-STIT theorem 2, we have $\mathsf{G}\Box\mathsf{H}(name(p) \rightarrow \psi) \in \Delta$. Thus, by definition of $\mathcal{R}^c_{\mathsf{G}}$ and the fact that $(\Delta, \Gamma) \in \mathcal{R}^c_{\mathsf{G}}$, $\Box\mathsf{H}(name(p) \rightarrow \psi) \in \Gamma$. By definition of $\mathcal{R}^c_{\Box}$ and the fact that $(\Gamma, \Gamma') \in \mathcal{R}^c_{\Box}$, $\mathsf{H}(name(p) \rightarrow \psi) \in \Gamma'$. Finally, by definition of $\mathcal{R}^c_{\mathsf{H}}$ and the fact that $(\Gamma', \Delta') \in \mathcal{R}^c_{\mathsf{H}}$, $name(p) \rightarrow \psi \in \Delta'$. Since $name(p) \in \Delta'$, it follows that $\psi \in \Delta'$. ■

Lemma 19. *$M^c|_{\mathbf{IrrTh}}$ is a superadditive temporal Kripke STIT model.*

Proof. First of all, it is a routine task to prove that: (i) the canonical relations $\mathcal{R}^c_{\Box}$, $\mathcal{R}^c_i$ and $\mathcal{R}^c_{Agt}$ are equivalence relations; (ii) the canonical relation $\mathcal{R}^c_{\mathsf{G}}$ is transitive; (iii) the canonical relation $\mathcal{R}^c_{\mathsf{H}}$ is the inverse of the canonical relation $\mathcal{R}^c_{\mathsf{G}}$; and (iv) the canonical model M^c satisfies the Constraints C1, C3*, C4 and C5.

Indeed, Axioms T, 4 and B for $\Box$, $[i]$ and for $[Agt]$ are canonical for reflexivity, transitivity and symmetry thereby ensuring that the canonical relations $\mathcal{R}^c_{\Box}$, $\mathcal{R}^c_i$ and $\mathcal{R}^c_{Agt}$ are equivalence relations. Axiom 4 for G is canonical for transitivity, thereby ensuring that the canonical relation $\mathcal{R}^c_{\mathsf{G}}$ is transitive. Axiom **(Conv$_{\mathsf{G},\mathsf{H}}$)** is canonical for the condition $\mathcal{R}_{\mathsf{H}} \subseteq \mathcal{R}^{-1}_{\mathsf{G}}$, while Axiom **(Conv$_{\mathsf{H},\mathsf{G}}$)** is canonical for the condition $\mathcal{R}_{\mathsf{G}} \subseteq \mathcal{R}^{-1}_{\mathsf{H}}$, thereby ensuring that $\mathcal{R}^c_{\mathsf{H}}$ is the inverse of the canonical relation $\mathcal{R}^c_{\mathsf{G}}$.

Finally, Axioms $(\Box \to i)$, $(i \to Agt)$, **(Connected$_{\mathsf{G}}$)** and **(Connected$_{\mathsf{H}}$)** are canonical respectively for the Constraints C1, C3*, C4 and C5, thereby ensuring that the canonical model M^c satisfies them.

Since $M^c|_{\mathbf{IrrTh}}$ is a submodel of the canonical model M^c, $M^c|_{\mathbf{IrrTh}}$ inherits from M^c all previous *universal* properties. Specifically, we have that: (i) the relations $\mathcal{R}^c_{\Box}|_{\mathbf{IrrTh}}$, $\mathcal{R}^c_i|_{\mathbf{IrrTh}}$ and $\mathcal{R}^c_{Agt}|_{\mathbf{IrrTh}}$ are equivalence relations; (ii) the relation $\mathcal{R}^c_{\mathsf{G}}|_{\mathbf{IrrTh}}$ is transitive; (iii) the relation $\mathcal{R}^c_{\mathsf{H}}|_{\mathbf{IrrTh}}$ is the inverse of the relation $\mathcal{R}^c_{\mathsf{G}}|_{\mathbf{IrrTh}}$; and (iv) the model $M^c|_{\mathbf{IrrTh}}$ satisfies the Constraints C1, C3*, C4 and C5.

By Propositions 17 and 18, $M^c|_{\mathbf{IrrTh}}$ also satisfies Constraints C2 and C6.

It remains to prove that the relation $\mathcal{R}^c_{\mathsf{G}}|_{\mathbf{IrrTh}}$ is serial and that the model $M^c|_{\mathbf{IrrTh}}$ satisfies the Constraint C7.

Since $\mathsf{G}\top$ is a T-STIT theorem, every IRR theory contains it. Thus, by Lemma 16, for every IRR theory Γ there exists an IRR theory Δ such that $(\Gamma, \Delta) \in \mathcal{R}^c_{\mathsf{G}}$. This guarantees that the relation $\mathcal{R}^c_{\mathsf{G}}|_{\mathbf{IrrTh}}$ is serial.

Now, suppose that $(\Gamma, \Delta) \in \mathcal{R}^c_{\Box}|_{\mathbf{IrrTh}}$. Since Δ is an IRR theory, $name(p) \in \Delta$ for some atom p. Hence, $p \in \Delta$. Moreover, $\Box\mathsf{G}\neg p \in \Delta$. By definition of $\mathcal{R}^c_{\Box}$, it follows that $\Diamond\Box\mathsf{G}\neg p \in \Gamma$. Thus, by the T-STIT theorem $\vdash \Diamond\Box\varphi \to \Box\varphi$, $\Box\mathsf{G}\neg p \in \Gamma$. By Axiom T for $\Box$, it follows that $\mathsf{G}\neg p \in \Gamma$. The latter implies that $\mathsf{F}p \notin \Gamma$. Since $p \in \Delta$, this guarantees that $(\Gamma, \Delta) \notin \mathcal{R}^c_{\mathsf{G}}|_{\mathbf{IrrTh}}$. Thus, $M^c|_{\mathbf{IrrTh}}$ satisfies the Constraint C7. ■

Lemma 10 follows from Lemma 13, Lemma 15, Lemma 16 and Lemma 19. Indeed, suppose that φ is a T-STIT consistent formula. Then, by Lemma 15, there exists an IRR theory Γ in the model $M^c|_{\mathbf{IrrTh}}$ such that $\varphi \in \Gamma$. By Lemma 19, model $M^c|_{\mathbf{IrrTh}}$ is a superadditive temporal Kripke STIT model. Moreover, by Lemma 16, the set of all IRR theories is diamond saturated. Thus, by Lemma 13, $M^c|_{\mathbf{IrrTh}}, \Gamma \models \varphi$. It follows that φ is T-STIT satisfiable. ■

Theorem 7 follows from Lemma 9 and Lemma 10.

3. Application to normative reasoning

Many normative concepts such as achievement obligation, obligation with deadline and social commitment have an intrinsic agentive and temporal nature – that is to say, they cannot be properly understood without considering their relationships with the concepts of action and time. The aim of this section is to show that the logic T-STIT is expressive enough to capture some of these relationships. I focus on the notion of social commitment and on its relationship with the notion of achievement obligation, postponing the logical analysis of obligations with deadline in STIT to future work.

According to Singh (1999) and Castelfranchi (1995), a social commitment is a kind of normative relationship between a *debtor* and a *creditor*. The contexts in which commitments

are undertaken and established are often institutional contexts. For instance, after signing a contract in the presence of a public notary, a person becomes committed in front of the State to carry out her part of the contract. In this article, I only consider pragmatic commitments and I leave aside propositional commitments (also called dialectical commitments). Pragmatic commitments are about what is to be done whereas propositional commitments are about what is true. Pragmatic commitments concern promises from a debtor to a creditor to perform a given action, while propositional commitments are about positions taken during a dialogue. For example, if i tells to j: 'I will lend you my car for the weekend!' then, he makes a pragmatic commitment to j. On the contrary, if i tells to j: 'Tomorrow, will be sunny. I am sure!' then, he makes a propositional commitment to j.

In order to be able to define social commitment, let us suppose that the set of propositional atoms *Atm* contains special atoms of the form $\mathrm{v}_{i,j}$, one for every $i, j \in Agt$ such that $i \neq j$. These are similar to the special atoms for *violation* that are used in the Anderson's reduction of deontic logic to alethic logic (Anderson, 1958; Lindahl, 1994). In the semantics, the special atom $\mathrm{v}_{i,j}$ is used to identify those worlds in which agent i does not fulfil her commitments to agent j. The atom $\mathrm{v}_{i,j}$ has to be read 'in the actual world, agent i does not fulfil her commitment to agent j'. Special atoms $\mathrm{v}_{i,j}$ can be combined with the tense operator G^* in order to identify those histories in which agent i will never fulfil her commitment to agent j. In particular, formula $\mathsf{G}^*\mathrm{v}_{i,j}$ has to be read 'in the actual history, agent i will never fulfil her commitment to agent j' or also 'in the actual history, agent j is wronged by agent i'.

I say that agent i is committed to agent j to ensure φ (denoted by $\mathbf{C}_{i:j}\varphi$) if and only if: (i) all historic alternatives in which i will never see to it that φ, are histories in which agent i will never fulfil her commitment to agent j; and (ii) i does not see to it that φ. In other words, i will not fulfil her commitment to j unless i will see to it that φ at some point in the future. For every $i, j \in Agt$, I define:

$$\mathbf{C}_{i:j}\varphi \stackrel{\text{def}}{=} \Box(\neg\mathsf{F}^*[i]\varphi \rightarrow \mathsf{G}^*\mathrm{v}_{i,j}) \wedge \neg[i]\varphi.$$

Let me make three observations about the preceding definition of commitment. First of all, it is worth noting that it is related to the notion of achievement obligation (Broersen et al., 2003; Governatori & Rotolo, 2010). The idea is that an agent has an achievement obligation to bring about φ if and only if she has the obligation to bring about φ at some point in the future.[4] Thus, commitment can be seen as a specific kind of achievement obligation. Specifically, it can be conceived as a *directed achievement obligation* from a bearer to a counterparty (for some analysis of the notion of directed obligation in deontic logic, see, e.g., Dignum, 1999; Herrestad & Krogh, 1995; Kanger & Kanger, 1966; Lindahl, 1994; Makinson, 1986).[5] That is to say, agent i is committed to agent j to bring about φ if and only if i has the obligation *towards* j to bring about φ at some point in the future. Secondly, the component $\neg[i]\varphi$ expresses that i is committed to j to ensure φ only if i has not yet fulfilled her commitment to j by bringing about φ. In this sense, the formula $[i]\varphi$ can be conceived as the *discharge* condition for commitment. Suppose agent i is committed to agent j to ensure φ. Then, if i sees to it that φ, then her commitment to j is discharged and is no longer active. Finally, according to the preceding definition, the fact that i brought about φ in the past (before being committed to j) is irrelevant. Indeed, the general intuition is that being committed *now* to do some action means being obliged to do the action either in the present or at some point in the strict future. Therefore, while the fact of performing the action in the present is a sufficient condition for discharging an actual commitment, the fact of having performed the action in the past is not.

The following example illustrates the preceding definition of commitment in a concrete scenario.

Example 20. *Agent* 2 *is the programme chair of a given conference. Agent* 2 *asks agent* 1*, a member of the programme committee, to review some articles submitted for the conference. Agent* 1 *accepts agent* 2*'s request by sending a confirmation e-mail (we suppose that the communication between* 1 *and* 2 *is made through the Easychair system). Consequently, according to the programme committee of the conference, agent* 1 *is committed to* 2 *to review the articles:* $\mathbf{C}_{1:2}$ *review. This means that (i) all historic alternatives in which* 1 *will never review the articles are histories in which* 2 *is wronged by* 1*; and (ii)* 1 *has not yet fulfilled her commitment to* 2 *by reviewing the articles:*

$$\Box(\neg \mathsf{F}^*[1]\mathit{review} \rightarrow \mathsf{G}^*\mathrm{v}_{1,2}) \wedge \neg[1]\mathit{review}.$$

Now, let us consider some logical properties of the commitment operator $\mathbf{C}_{i:j}$. According to the following T-STIT theorem 3, if i is committed to j to ensure $\varphi \wedge \psi$ then i is committed to j to ensure φ and to ensure ψ. For every $i, j \in \mathit{Agt}$ we have:

$$\vdash_{\text{T-STIT}} \mathbf{C}_{i:j}(\varphi \wedge \psi) \rightarrow (\mathbf{C}_{i:j}\varphi \wedge \mathbf{C}_{i:j}\psi). \tag{3}$$

Note that the converse of T-STIT theorem 3 is not valid. Indeed, the fact that i is committed to j to ensure φ and to ensure ψ (i.e., $\mathbf{C}_{i:j}\varphi \wedge \mathbf{C}_{i:j}\psi$) does not necessarily imply that i is committed to j to ensure that φ and ψ are true at the same point in the future (i.e., $\mathbf{C}_{i:j}(\varphi \wedge \psi)$). For example, i may be committed to j to lend her a car and to lend her a motorbike at different points in the future without being committed to lending her a car and a motorbike at the same point in the future.

The following theorem 4 highlights that an agent cannot be committed to bring about tautologies:

$$\vdash_{\text{T-STIT}} \neg\mathbf{C}_{i:j}\top. \tag{4}$$

Indeed, since $[i]\top$ is a T-STIT theorem, agent i's commitment to bring about $\top$ is always discharged.

The following T-STIT theorem 5 is a *weakening* principle for commitment. A similar property of commitment has been isolated by Singh (2008). For every $i, j \in \mathit{Agt}$ we have:

$$\vdash_{\text{T-STIT}} (\mathbf{C}_{i:j}(\varphi \wedge \psi) \wedge [i]\varphi) \rightarrow \mathbf{C}_{i:j}\psi. \tag{5}$$

This means that if agent i is committed to agent j to ensure $\varphi \wedge \psi$ and agent i sees to it that φ, then i is committed to j to ensure ψ. So, if an agent is committed to ensuring two states of affairs φ and ψ and her commitment to ensure φ is discharged, then the agent is committed to ensuring ψ.

The following T-STIT theorem 6 clarifies how the previous definition of commitment behaves in the case of Moore-like sentences of the form $\varphi \wedge \neg[i]\varphi$: an agent is committed to ensure that φ is true and that she does not see to it that φ if and only if the agent is committed to do something inconsistent:

$$\vdash_{\text{T-STIT}} \mathbf{C}_{i:j}(\varphi \wedge \neg[i]\varphi) \leftrightarrow \mathbf{C}_{i:j}\bot. \tag{6}$$

The following T-STIT theorem 7 characterises the relationship between commitments about logical equivalent formulas:

$$\vdash_{\text{T-STIT}} \Box\mathsf{G}^*(\varphi \leftrightarrow \psi) \rightarrow (\mathbf{C}_{i:j}\varphi \leftrightarrow \mathbf{C}_{i:j}\psi). \tag{7}$$

This means that if φ and ψ are equivalent in all future worlds of every history passing through the current moment, then agent i is committed to ensuring φ if and only if agent i is committed to ensuring ψ. Note that the formula $\Box \mathsf{G}^* \varphi$ captures a form of *future necessity* or, even better, *future inevitability*. Therefore, theorem 7 captures the idea that if in the future it is inevitable that φ and ψ have the same truth value, then one cannot be committed to ensuring φ without being committed to ensuring ψ, and one cannot be committed to ensuring ψ without being committed to ensuring φ.

The last T-STIT theorem considered here is the following one about the relationship between time and commitment:

$$\vdash_{\text{T-STIT}} \mathbf{C}_{i:j}\varphi \rightarrow \mathsf{F}(\neg[i]\varphi \rightarrow \mathbf{C}_{i:j}\varphi). \tag{8}$$

Since the Hilbert-style derivation of this T-STIT theorem is not trivial, I give it here:

1. $\vdash \mathbf{C}_{i:j}\varphi \stackrel{\text{def}}{=} \Box(\neg \mathsf{F}^*[i]\varphi \rightarrow \mathsf{G}^* \mathrm{v}_{i,j}) \wedge \neg[i]\varphi$.
2. $\vdash (\Box(\neg \mathsf{F}^*[i]\varphi \rightarrow \mathsf{G}^* \mathrm{v}_{i,j}) \wedge \neg[i]\varphi) \rightarrow$
   $([Agt](\neg \mathsf{F}^*[i]\varphi \rightarrow \mathsf{G}^* \mathrm{v}_{i,j}) \wedge [Agt]\neg[i]\varphi)$.
   By Axiom $(\Box \rightarrow i)$, Axiom $(i \rightarrow Agt)$ and T-STIT theorem $\neg[i]\varphi \rightarrow [i]\neg[i]\varphi$.
3. $\vdash ([Agt](\neg \mathsf{F}^*[i]\varphi \rightarrow \mathsf{G}^* \mathrm{v}_{i,j}) \wedge [Agt]\neg[i]\varphi) \rightarrow$
   $[Agt]((\neg \mathsf{F}^*[i]\varphi \rightarrow \mathsf{G}^* \mathrm{v}_{i,j}) \wedge \neg[i]\varphi)$.
   By Axiom K for $[Agt]$.
4. $\vdash [Agt]((\neg \mathsf{F}^*[i]\varphi \rightarrow \mathsf{G}^* \mathrm{v}_{i,j}) \wedge \neg[i]\varphi) \leftrightarrow$
   $[Agt](([i]\varphi \vee \mathsf{F}[i]\varphi \vee \mathsf{G}^* \mathrm{v}_{i,j}) \wedge \neg[i]\varphi)$.
5. $\vdash [Agt](([i]\varphi \vee \mathsf{F}[i]\varphi \vee \mathsf{G}^* \mathrm{v}_{i,j}) \wedge \neg[i]\varphi) \rightarrow$
   $[Agt](\mathsf{F}[i]\varphi \vee \mathsf{G}^* \mathrm{v}_{i,j})$.
   By Axiom K for $[Agt]$.
6. $\vdash [Agt](\mathsf{F}[i]\varphi \vee \mathsf{G}^* \mathrm{v}_{i,j}) \rightarrow [Agt](\mathsf{F}[i]\varphi \vee \mathsf{G} \mathrm{v}_{i,j})$.
   By Axiom K for $[Agt]$.
7. $\vdash [Agt](\mathsf{F}[i]\varphi \vee \mathsf{G} \mathrm{v}_{i,j}) \rightarrow \mathsf{F}\Box(\mathsf{F}^*[i]\varphi \vee \mathsf{G}^* \mathrm{v}_{i,j})$.
   By the second T-STIT validity in Proposition 6.
8. $\vdash \mathsf{F}\Box(\mathsf{F}^*[i]\varphi \vee \mathsf{G}^* \mathrm{v}_{i,j}) \rightarrow \mathsf{F}\Box(\neg \mathsf{F}^*[i]\varphi \rightarrow \mathsf{G}^* \mathrm{v}_{i,j})$.
9. $\vdash \mathsf{F}\Box(\neg \mathsf{F}^*[i]\varphi \rightarrow \mathsf{G}^* \mathrm{v}_{i,j}) \rightarrow$
   $\mathsf{F}(([i]\varphi \vee \neg[i]\varphi) \wedge \Box(\neg \mathsf{F}^*[i]\varphi \rightarrow \mathsf{G}^* \mathrm{v}_{i,j}))$.
   By T-STIT theorem $\mathsf{F}\varphi \rightarrow \mathsf{F}(\top \wedge \varphi)$.
10. $\vdash \mathsf{F}(([i]\varphi \vee \neg[i]\varphi) \wedge \Box(\neg \mathsf{F}^*[i]\varphi \rightarrow \mathsf{G}^* \mathrm{v}_{i,j})) \rightarrow$
    $\mathsf{F}([i]\varphi \vee \mathbf{C}_{i:j}\varphi)$.
11. $\vdash \mathbf{C}_{i:j}\varphi \rightarrow \mathsf{F}(\neg[i]\varphi \rightarrow \mathbf{C}_{i:j}\varphi)$.
    From 1–10.

According to the previous T-STIT theorem 8, if agent i is committed to agent j to ensure φ then, at some point in the future, either i will see to it that φ or i will remain committed to j to ensure φ. This captures a kind of *persistence* of commitment over time.[6] That is, a commitment will persist as long as the committed agent does not perform the action to

which she is committed. Note that this property of commitment relies on the assumption that there are no 'external' actions or processes that may cancel a pre-existent commitment independently from the fact that the committed agent performs the action for which she is committed. For instance, getting back to the above example, agent 1's commitment to agent 2 to review the articles could be dropped not only because 1 performs the action for which is committed but also because 2 decides to cancel 1's commitment by assigning to a different programme committee member the reviews that were initially assigned to 1.

4. Related work

The logic T-STIT presented in Section 2 differs in several aspects from the system proposed by Wölf (2002). First of all, the language of T-STIT contains the operator $[Agt]$ for the joint action of the grand coalition, while the language of Wölf's logic does not. I have shown that the operator $[Agt]$ is fundamental in order to characterise a basic property relating action and time studied in STIT: the so-called property of *no choice between undivided histories*. Moreover, different to the system of Wölf, T-STIT does not have extra modal operators such as the *difference* operator $\Box^{\neq}$ taken from de Rijke (1992), where $\Box^{\neq}\varphi$ means that 'φ is true in all possible histories that are different from the actual history', and the modal operator $\boxminus$ taken from Maio and Zanardo (1998), where $\boxminus\varphi$ means that 'φ is true in all possible histories that belong to the current instant'. In this sense, the language of T-STIT is more *minimalistic* than the language of Wölf's logic, as it contains only those operators which are strictly necessary for talking about tense and action in an interesting way, while Wölf's logic contains the preceding two extra modal operators $\Box^{\neq}$ and $\boxminus$, which are not necessary for this. Wölf's system and T-STIT are also different at the semantic level. Wölf gives two semantics for his temporal variant of STIT: one based on so-called $T \times W$-based agent frames, which are extensions of $T \times W$-frames (Thomason, 1984), and the other based on so-called tree-based agent-frames, which are similar to the **BT**+**AC** structures of Belnap et al. (2001). Wölf proves that the two semantics are equivalent. The advantage of the T-STIT semantics given in Section 2.2, compared to Wölf's two semantics, is that the former is closer to the standard semantics of modal logic (Blackburn et al., 2001), as it is based on the standard notion of Kripke model with accessibility relations. A comparison between the semantics of T-STIT in terms of temporal Kripke STIT models and Wölf's two semantics is deferred to future work.

T-STIT also differs from the system proposed by Broersen (Broersen, 2008a,b) under several aspects. First of all, Broersen presents a variant of STIT called XSTIT in which the temporal dimension and the agency dimension are fused to make up a single modal operator. In particular, in Broersen's logic there are primitive operators describing the effects of an agent's action in 'next' states, where 'next' refers to immediate successors of the present state. On the contrary, in T-STIT, the temporal dimension and the agency dimension are kept separate. Secondly, different to T-STIT, Broersen's logic XSTIT does not have a future tense operator. Thirdly, while the notion of group action used in T-STIT corresponds to the notion of group action given by Horty (2001), the notion of group action in Broersen's XSTIT logic does not. According to Horty's definition, the set of outcomes that is forced by the joint action of a coalition is equal to the pointwise intersection of the sets of outcomes that are forced by the individual actions of the agents in the coalition. In Broersen's logic only the left-to-right direction of Horty's (2001) definition holds (i.e., the set of outcomes

that is forced by the individual action of an agent in a coalition is included in the set of outcomes that is forced by the joint action of the coalition).

Another logic that is related to the logic T-STIT presented in Section 2 is Schwarzentruber's (2012) variant of STIT with discrete time extended with the tense operator 'next' of linear temporal logic (LTL). Schwarzentruber provides complexity results for this logic without considering the issue of axiomatisation. However, the expressive power of Schwarzentruber's temporal variant of STIT is too limited for modelling normative concepts such as achievement obligation and commitment. Indeed, the tense operator 'next' is insufficient to express temporal properties about commitments such as the fact that an agent is committed to doing something *in the future*. On the contrary, as I have shown in Section 3, the tense operator G (henceforth) allows us to express such properties.

5. Conclusion

I have presented in this work a temporal variant of STIT which supports reasoning about the temporal properties of normative concepts such as achievement obligation and commitment. A sound and complete axiomatisation for this logic has been given.

Directions of future work are manifold. I defer to future work an extension of the logic T-STIT by the tense operators *until* and *before* of linear temporal logic (LTL; Gabbay et al., 1980). With these operators it will be possible to provide a definition of commitment based on the notion of *deadline* of the following form: agent i is committed to agent j to ensure φ before the deadline ψ if and only if, if i does not see to it that φ *before* ψ becomes true, then j will be wronged by i.

Another issue that I plan to investigate in future research is a comparison between the Kripke-style semantics of T-STIT given in Section 2 and an alternative semantics for T-STIT based on the original notion of the **BT**+**AC** structures of Belnap et al. (2001). Indeed, while the two semantics are clearly equivalent in the case of atemporal individual STIT and atemporal group STIT, one might wonder whether they lead to two different sets of validities in the case of T-STIT and, if so, whether the validities differentiating the two semantics capture interesting and intuitive properties.

Finally, on the technical side, I intend to provide an alternative axiomatisation of T-STIT which does not make use of the Gabbay-style irreflexivity rule **(IRR)**. Indeed, following Zanardo (1996, Theorem 6.12), it seems possible to find such an alternative axiomatisation in which the rule **(IRR)** is replaced with a set of more complex axiom schemas.

Another issue for future work is decidability of the satisfiability problem of T-STIT. It has been proved by Herzig and Schwarzentruber (2008) that Horty's group STIT with group agency operators for all coalitions is undecidable. However, the logic T-STIT only has the group agency operator for the grand coalition. Because of this limitation in the expressive power of T-STIT, I believe that its satisfiability problem is decidable.

Acknowledgements

The author acknowledges the support of the LabEx project CIMI. Moreover, he is grateful to Valentin Goranko for his helpful comments on the content of this work.

Notes

1. The distinction between the 'Ockhamist' view and the 'Peircean' view of branching time was proposed by Prior's (1967) seminal work on the logic of time. According to the 'Peircean' view, the truth of a temporal formula should be evaluated with respect either to some history or all

histories passing through a given moment. The 'Ockhamist' view considers a notion of *actual* course of events. In particular, according to the 'Ockhamist' view, the truth of a temporal formula should be evaluated with respect to a particular *actual* history passing through a given moment.

2. The function f is an order-isomorphism between $\mathcal{R}_{\mathsf{H}}(w)$ and $\mathcal{R}_{\mathsf{H}}(v)$ if and only if f is a bijective function $f : \mathcal{R}_{\mathsf{H}}(w) \longrightarrow \mathcal{R}_{\mathsf{H}}(v)$ with the property that for all $u, u' \in \mathcal{R}_{\mathsf{H}}(w)$, $(u, u') \in \mathcal{R}_{\mathsf{H}}$ if and only if $(f(u), f(u')) \in \mathcal{R}_{\mathsf{H}}$.
3. Xu (1998) considers a family of axiom schemas **(AIA$_k$)** for independence of agents of the form $(\Diamond[1]\varphi_1 \wedge \ldots \wedge \Diamond[k]\varphi_k) \rightarrow \Diamond([1]\varphi_1 \wedge \ldots \wedge [k]\varphi_k)$ that is parameterised by the integer k. As pointed out by Belnap et al. (2001), **(AIA$_{k+1}$)** implies **(AIA$_k$)**. Therefore, as *Agt* is finite, in T-STIT the family of axiom schemas can be replaced by the single axiom **(AIA)**.
4. The notion of achievement obligation is traditionally opposed to the concept of maintenance obligation (i.e., the obligation to maintain a given state of affairs φ).
5. Starting from Hohfeld (1917), in legal theory it is assumed that a *right* for j towards i that φ is brought about by i typically correlates with a directed obligation for i towards j to bring about φ (i.e., a *duty* of i towards j to bring about φ).
6. Note that a similar property of persistence characterises the notion of intention (Bratman, 1987) and has been formally characterised in some logical theories of intention (Cohen & Levesque, 1990; Lorini & Herzig, 2006).

References

Anderson, A. (1958). A reduction of deontic logic to alethic modal logic. *Mind*, *22*, 100–103.

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

Belnap, N., Perloff, M., & Xu, M. (2001). *Facing the future: Agents and choices in our indeterminist world*. Oxford: Oxford University Press.

Bentahar, J., Moulin, B., Ch. Meyer, J.-J., & Chaib-draa, B. (2004). A logical model for commitment and argument network for agent communication. In N. Jennings & M. Tambe (Eds.), *3rd International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2004), 19–23 August 2004, New York, NY, USA, proceedings* (Vol. 2, pp. 792–799). Washington, DC: IEEE Computer Society.

Blackburn, P., de Rijke, M., & Venema, Y. (2001). *Modal logic*. Cambridge: Cambridge University Press.

Bratman, M. (1987). *Intentions, plans, and practical reason*. Cambridge, MA: Harvard University Press.

Broersen, J. (2008a). A complete STIT logic for knowledge and action, and some of its applications. In M. Baldoni, T. C. Son, M. B. van Riemsdijk, & M. Winikoff (Eds.), *Declarative Agent Languages and Technologies VI, 6th International Workshop, DALT 2008, Estoril, Portugal, May 12, 2008, revised selected and invited papers* (pp. 47–59). Berlin: Springer.

Broersen, J. (2008b). A logical analysis of the interaction between 'obligation-to-do' and 'knowingly doing'. In R. van der Meyden & L. van der Torre (Eds.), *Deontic Logic in Computer Science, 9th International Conference, DEON 2008, Luxembourg, July 15–18, 2008. Proceedings* (pp. 140–154). Berlin: Springer.

Broersen, J., Dastani, M., & van der Torre, L. (2003). BDIO-CTL: Obligations and the specification of agent behavior. In G. Gottlob & T. Walsh (Eds.), *IJCAI-03, Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence, Acapulco, Mexico, August 9–15, 2003* (pp. 1389–1390). San Francisco, CA: Morgan Kaufmann.

Castelfranchi, C. (1995). Commitment: From individual intentions to groups and organizations. In V. R. Lesser & L. Gasser (Eds.), *Proceedings of the First International Conference on Multiagent Systems, June 12–14, 1995, San Francisco, California, USA* (pp. 528–535). Cambridge, MA: MIT Press.

Chellas, B. J. (1992). Time and modality in the logic of agency. *Studia Logica*, *51*, 485–517.

Cohen, P. R., & Levesque, H. J. (1990). Intention is choice with commitment. *Artificial Intelligence*, *42*, 213–261.

de Rijke, M. (1992). The modal logic of inequality. *The Journal of Symbolic Logic*, *57*, 566–584.

Desai, N., Narendra, N. C., & Singh, M. P. (2008). Checking correctness of business contracts via commitments. In L. Padgham, D. C. Parkes, J. P. Müller, & S. Parsons (Eds.), *7th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2008), Estoril, Portugal, May 12–16, 2008* (Vol. 2, pp. 787–794). New York, NY: ACM Press.

Dignum, F. (1999). Autonomous agents with norms. *Artificial Intelligence and Law*, *7*, 69–79.

Gabbay, D. M., Hodkinson, I. M., & Reynolds, M. A. (1994). *Temporal logic: Mathematical foundations and computational aspects* (Vol. 1). Oxford: Clarendon Press.

Gabbay, D. M., Pnueli, A., Shelah, S., & Stavi, J. (1980). On the temporal analysis of fairness. In P. W. Abrahams, R. J. Lipton, & S. R. Bourne (Eds.), *Conference Record of the Seventh Annual ACM Symposium on Principles of Programming Languages, Las Vegas, Nevada, USA, January 1980* (pp. 163–173). New York, NY: ACM Press.

Goldblatt, R. (1992). *Logics of time and computation*. (2nd ed.). Stanford, CA: Center for the Study of Language and Information.

Governatori, G., & Rotolo, A. (2010). Norm compliance in business process modeling. In M. Dean, J. Hall, A. Rotolo, & S. Tabet (Eds.), *Semantic Web Rules - International Symposium, RuleML 2010, Washington, DC, USA, October 21–23, 2010. Proceedings* (pp. 194–209). Berlin: Springer.

Herrestad, H., & Krogh, C. (1995). Obligations directed from bearers to counterparties. *Proceedings of the Fifth International Conference on Artificial Intelligence and Law, ICAIL '95, May 21–24, 1995, College Park, Maryland, USA* (pp. 210–218). New York, NY: ACM Press.

Herzig, A., & Schwarzentruber, F. (2008). Properties of logics of individual and group agency. In C. Areces & R. Goldblatt (Eds.), *Advances in Modal Logic* (Vol. 7, pp. 133–149). London: King's College.

Hohfeld, W. (1917). Some fundamental legal conceptions as applied in legal reasoning. *The Yale Law Journal*, *26*, 710–770.

Horty, J. F. (2001). *Agency and deontic logic*. Oxford: Oxford University Press.

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

Kanger, S., & Kanger, H. (1966). Rights and parliamentarism. *Theoria*, *6*, 85–115.

Kooi, B., & Tamminga, A. (2008). Moral conflicts between groups of agents. *Journal of Philosophical Logic*, *37*, 1–21.

Lindahl, L. (1994). Stig Kanger's theory of rights. In D. Prawitz, B. Skyrms, & D. Westeråhl (Eds.), *Logic, methodology and philosophy of science IX*. Philadelphia, PA: Elsevier.

Lorini, E., & Herzig, A. (2006). A logic of intention and attempt. *Synthese*, *163*, 45–77.

Lorini, E., & Schwarzentruber, F. (2011). A logic for reasoning about counterfactual emotions. *Artificial Intelligence*, *175*, 814–847.

Maio, M. C. D., & Zanardo, A. (1998). A Gabbay-rule free axiomatization of T x W validity. *Journal of Philosophical Logic*, *27*, 435–487.

Makinson, D. (1986). On the formal representation of rights relations: Remarks on the work of Stig Kanger and Lars Lindahl. *Journal of Philosophical Logic*, *15*, 403–425.

Prior, A. (1967). *Past, present, and future*. Oxford: Clarendon Press.

Reynolds, M. A. (2003). An axiomatization of Prior's Ockhamist logic of historical necessity. In P. Balbiani, N.-Y. Suzuki, F. Wolter, & M. Zakharyaschev (Eds.), *Advances in Modal Logic* (Vol. 4, pp. 355–370). London: King's College.

Schwarzentruber, F. (2012). Complexity results of STIT fragments. *Studia Logica*, *100*, 1001–1045.

Singh, M. P. (1999). An ontology for commitments in multiagent systems. *Artificial Intelligence and Law*, *7*, 97–113.

Singh, M. P. (2008). Semantical considerations on dialectical and practical commitments. In D. Fox & C. P. Gomes (Eds.), *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, AAAI 2008, Chicago, Illinois, USA, July 13–17, 2008* (pp. 176–181). Menlo Park, CA: AAAI Press.

Thomason, R. (1984). Combinations of tense and modality. In D. M. Gabbay & F. Guenthner (Eds.), *Handbook of Philosophical Logic* (2nd ed., Vol. 2, pp. 135–165). Berlin: Springer.

Vakarelov, D. (1992). A modal theory of arrows: Arrow logics I. In D. Pearce & G. Wagner (Eds.), *Logics in AI, European Workshop, JELIA '92, Berlin, Germany, September 7–10, 1992, proceedings* (pp. 1–24). Berlin: Springer.

von Kutschera, F. (1997). T x W completeness. *Journal of Philosophical Logic*, *26*, 241–250.

Wölf, S. (2002). Propositional Q-logic. *Journal of Philosophical Logic*, *31*, 387–414.

Xu, M. (1998). Axioms for deliberative STIT. *Journal of Philosophical Logic*, *27*, 505–552.

Zanardo, A. (1996). Branching-time logic with quantification over branches: The point of view of modal logic. *The Journal of Symbolic Logic*, *61*, 143–166.

A STIT Logic for Reasoning about Social Influence

Emiliano Lorini[1] and Giovanni Sartor[2,3]
[1] IRIT-CNRS, Toulouse University, France
[2] European University Institute, Fiesole, Italy
[3] University of Bologna, Italy

Abstract

In this paper we propose a method for modeling social influence within the STIT approach to action. Our proposal consists in extending the STIT language with special operators that allows us to represent the consequences of an agent's choices over the rational choices of another agent.

1 Introduction

Both human and artificial societies are based on mutual influence. Agents are dependent on others for the realization of their goals, and only by influencing others, and obtaining their cooperation, they can adapt the physical and social world to their needs. Influence may take place through speech acts, as when one issues a request, an order or an advice, promises a reward to or threatens a sanction. It may also result from non-communicative behavior that, intentionally or unintentionally, obstacles or facilitates the performance of an action by another, as when one consumes a resource or blocks or limits access to a resource.

There have been a number of significant contributions to the logic of social influence (see in particular [32]) and to the cognitive aspects and computational aspects involved in it [10, 30]. However, at the current stage, there is no logical theory of social influence that is capable of: (i) capturing the temporal aspect of influence, namely the fact that the influencer's choice must precede the influencee's action,[1] and (ii) addressing the strategic aspect of influencing relationships through extensive-form games. STIT logic (the logic of "seeing to it that") by Belnap et al. [3] presents some features which make it most promising for analyzing these two essential aspects of social influence. In this paper we aim at showing that STIT can provide indeed a useful framework for modeling influence relationships. For this purpose, however STIT needs to be integrated with appropriate constructs, which make the influencee's agency consistent with the fact that the influencer determines the influencee's choices.

The paper is organized as follows. In Section 2 we recall the general semantics of STIT, while in Section 3 we discuss the concept of influence from an informal perspective. Section 4 introduces a variant of STIT logic that will be used in Section 6 to formalize the concept of social influence informally discussed in Section 3. Our logic, called DR-STIT (STIT with Deterministic time and Rational choices), extends the basic STIT language with special operators

[1] The term 'influencer' refers to the agent who exerts influence, whereas the term 'influencee' refers to the agent being influenced.

that allow us to represent the consequences of an agent's *rational* choice. An axiomatization of DR-STIT is given in Section 5. Section 7 is about related work, while Section 9 discusses some perspectives of future research. The proof of Theorem 1 is included in the technical annex at the end of the paper.

2 Background on STIT semantics

STIT logic (the logic of *seeing to it that*) by Belnap et al. [3] is one of the most prominent formal accounts of agency. It is the logic of sentences of the form "the agent i sees to it that φ is true". In [18] Horty extends Belnap et al.'s STIT framework with operators of group agency in order to express sentences of the form "the group of agents J sees to it that φ is true". Though also [3] approaches collective ('joint') agency, Horty's variant of group STIT is the most established today, and it provides the standard combination of agency operators for the individuals and agency operators for the groups. Different semantics for STIT have been proposed in the literature (see, e.g., [3, 6, 39, 26, 24, 37]). The original semantics of STIT by Belnap et al. [3] is defined in terms of $\mathbf{BT}{+}\mathbf{AC}$ structures: branching-time structures ($\mathbf{BT}$) augmented by agent choice functions ($\mathbf{AC}$). A $\mathbf{BT}$ structure is made of a set of moments and a tree-like ordering over them. An $\mathbf{AC}$ for an agent i is a function mapping each moment m into a partition of the set of histories passing through that moment, a history h being a maximal set of linearly ordered moments and the equivalence classes of the partition being the possible choices for agent i at moment m.

Following [24], here we adopt a Kripke-style semantics for STIT. On the conceptual side, the main difference between a Kripke semantics for STIT and Belnap et al.'s $\mathbf{BT}{+}\mathbf{AC}$ semantics is that the former takes the concept of *world* as a primitive instead of the concept of *moment* and defines: (i) a *moment* as an equivalence class induced by a certain equivalence relation over the set of worlds, (ii) a *history* as a linearly ordered set of worlds induced by a certain partial order over the set of worlds, and (iii) an agent i's set of *choices* at a moment as a partition of that moment. The main advantage of the Kripke semantics for STIT over Belnap et al.'s original semantics in terms of $\mathbf{BT}{+}\mathbf{AC}$ structures is that the former is a standard multi-relational semantics commonly used in the area of modal logic [5], whereas the latter is non-standard. By using the Kripke semantics for STIT we can use methods and techniques developed in this area to prove results about axiomatics (see Section 5) and decidability (postponed to future work). We do not have space here to discuss the differences and similarities between the two semantics. However, it is worth noticing that the two semantics are equivalent when considering either atemporal STIT languages (STIT languages without temporal modal operators) such as the one studied in [1] or a temporal STIT language with temporal operators 'next' and 'yesterday' such as the one we will introduce in Section 4. On the contrary, the two semantics are different when considering richer temporal STIT languages such the one studied in [24] which includes the future tense operator $\mathbf{G}$ (where $\mathbf{G}\varphi$ stands for "φ will always be true in the future") and the past tense operator $\mathbf{H}$ (where $\mathbf{H}\varphi$ stands for "φ has always been true in the past") of basic temporal logic [34]. In particular, as clearly shown in [25], when considering such richer temporal languages, Belnap et al.'s original semantics corresponds to a full tree semantics for branching time temporal logics while the Kripke semantics corresponds to a bundled tree semantics. As already observed in the area of branching time temporal logics (see, e.g., [9, 35]), the full tree semantics has more validities than the bundled tree semantics.

The Kripke semantics of STIT is illustrated in Figure 1, where each moment m_1, m_2 and

m_3 consists of a set of worlds represented by points. For example, moment m_1 consists of the set of worlds $\{w_1, w_2, w_3, w_4\}$. Moreover, for every moment there exists a set of histories passing through it, where a history is defined as a linearly ordered set of worlds. For example, the set of histories passing through moment m_1 is $\{h_1, h_2, h_3, h_4\}$. Finally, for every moment, there exists a partition which characterizes the set of available choices of agent 1 in this moment. For example, at moment m_1, agent 1 has two choices, namely $\{w_1, w_2\}$ and $\{w_3, w_4\}$. Note that an agent's set of choices at a certain moment can also be seen as a partition of the set of histories passing through this moment. For example, we can identify the choices available to agent 1's at m_1 with the two sets of histories $\{h_1, h_2\}$ and $\{h_3, h_4\}$.

Clearly, for every moment m in a Kripke semantics for STIT, one can identify the set of histories passing through it by considering all histories that contain at least one world in the moment m. Moreover, an agent i's set of choices available at m can also be seen as as a partition of the set of histories passing through m. At first glance, an important difference between Belnap et al.'s semantics and Kripke semantics for STIT seems to be that in the former the truth of a formula is relative to a moment-history pair m/h, also called *index*, whereas in the latter it is relative to a world w. However, this difference is only apparent, because in the Kripke semantics for STIT there is a one-to-one correspondence between worlds and indexes, in the sense that: (i) for every index m/h there exists a unique world w at the intersection between m and h, (ii) and for every world w there exists a unique index m/h such that the intersection between m and h includes w. (This point will become clearer in Section 4.2 in which a Kripke semantics for our variant of STIT will be specified.)

In the Kripke semantics for STIT the concept of world should be understood as a 'time point' and the equivalence class defining a moment as a set of alternative concomitant 'time points'. In this sense, the concept of moment captures a first aspect of indeterminism, as it represents the alternative ways the *present* could be. A second aspect of indeterminism is given by the fact that moments are related in a (tree-like) branching time structure. In this sense, the *future* could evolve in different ways from a given moment. In the Kripke semantics for STIT these two aspects of indeterminism are related, as illustrated in Figure 1. Indeed, if two distinct moments m_2 and m_3 are in the future of moment m_1, then there are two distinct worlds in m_1 (w_1 and w_3) such that a successor of the former (w_5) is included in m_2 and a successor of the latter (w_7) is included in m_3.

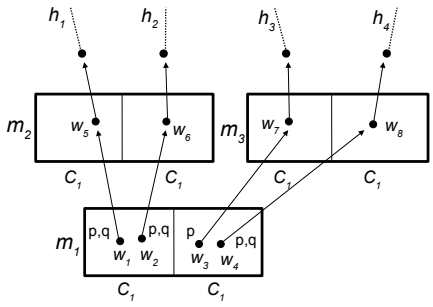


Figure 1: Illustration of Kripke semantics of STIT

The STIT semantics provides for different concepts of agency, all characterized by the fact that an agent acts only if she sees to it that a certain state of affairs is the case. In this paper we

consider two different notions of agency, namely the so-called Chellas STIT, named after its proponent [11], and the deliberative STIT [19].[2]

An agent i Chellas-sees-to-it that φ, denoted by formula $[i\ \mathbf{stit}]\varphi$, at a certain world w if and only if, for every world v, if w and v belong to the same choice of agent i then φ is true at world v. For example, in Figure 1, agent 1 Chellas-sees-to-it that p at world w_1 because p is true both at world w_1 and at at world w_2.

Deliberative STIT satisfies the same positive condition as Chellas STIT *plus* a negative condition: an agent i deliberately-sees-to-it that φ, denoted by formula $[i\ \mathbf{dstit}]\varphi$, at a certain world w if and only if: (i) agent i Chellas-sees-to-it that φ at w, that is to say, agent i's current choice at w ensures φ, and (ii) at w agent i could make a choice that does not necessarily ensure φ. Notice that the latter is equivalent to say that there exists a world v such that w and v belong to the same moment and φ is false at v. For example, in Figure 1, agent 1 deliberately sees to it that q at world w_1 because q is true both at world w_1 and at world w_2, while being false at world w_3. In other terms, while the truth of $[i\ \mathbf{stit}]\varphi$ only requires that i's choice ensures that φ, the truth of $[i\ \mathbf{dstit}]\varphi$ also requires that i had the opportunity of making an alternative choice that would not guarantee that φ would be the case. Deliberative STIT, we would argue, captures a fundamental aspect of the concept of action, namely, the idea that for a state of affairs to be the consequence of an action (or for an action to be the cause of a state of affairs), it is not sufficient that the action is a sufficient condition for that state of affairs to hold, it is also required that, without the action, the state of affairs possibly would not hold (a similar idea is also included in the logic of "bringing it about" by Pörn, see in particular [33]). In this sense, while $[i\ \mathbf{stit}]\varphi$ at w is consistent with (and is indeed entailed by) the necessity of φ at w, $[i\ \mathbf{dstit}]\varphi$ at w is incompatible with the necessity of φ at w, since it requires that at w also $\neg\varphi$ was an open possibility. Consequently, the deliberative STIT is more appropriate than the Chellas STIT to describe the consequences of an agent's action, as *incompatibility with necessity* is a requirement for any reasonable concept of action.[3]

3 The concept of influence

As emphasized by [4], the first prominent philosopher of the twentieth century to provide a compatibilist solution to the free will problem based on a conditional or hypothetical analysis is G. E. Moore. According to Moore's famous analysis [29], one is free in performing an action if one "could have done otherwise", but the latter expression has to be understood in a particular way, namely, as the requirement that "should one have done otherwise if one had chosen to do so". In this sense, according to Moore, determinism is compatible with free will insofar as an agent might have several choices or alternatives *available* at a given moment defining her *choice set*, even though what the agent does is determined by her *actual* choice, which is in turn determined by the agent's *choice context* including her preferences and beliefs and the composition of her choice set.

Our analysis of social influence expands Moore's view by assuming that the agent's choice context determining the agent's actual choice might be determined by external causes. Specifically, the external conditions in which an agent finds herself or the other agents with whom the

[2] We shall not consider achievement STIT of [2].

[3] The classical argument against the use of Chellas STIT for modeling action is that, according to Chellas STIT, an agent brings about all tautologies and that it is counterintuitive to say that a tautology is a consequence of an agent's action.

agent interacts may provide an input to the agent's decision-making process in such a way that a determinate action should follow. Indeed, as Leibniz [22, 383] observed, "precepts, armed with power to punish and to recompense, are very often of use and are included in the order of causes which make an action exist".

We argue that genuine influence consists in *determining* the voluntary action of an agent by modifying her *choice context*, so that a different choice becomes preferable to the influencee on comparison to what would be her preferred option without this modification. This may happen:

- by expanding the set of the available choices (influence via choice set expansion), or
- by restricting the set of the available choices (influence via choice set restriction),[4] or
- by changing the payoffs associated to such choices, as when rewards or punishments are established (influence via payoff change).

There exists a fourth kind of social influence that cannot be modelled by means of the formal machinery developed in this paper, namely influence via belief change or persuasion. Indeed, as classical decision and game theory highlight, an agent's choice depends not only on the payoffs associated with the different outcomes but also on what the agent believes the others will do. A typical example of influence via belief change is the situation of a coordination game, as the one depicted in Figure 2, in which two rational players, the row player and the column player, have to choose the same action (either α or β) in order to get a reward. Assume there is a communication phase before the game is played. During the preplay communication phase, the row player can induce the column player to believe that she will play action α. Consequently, since he is rational and believes that the row player will choose action α, the column player will choose action α. An obvious extension of the present logical theory of influence, to be

	α	β
α	1,1	0,0
β	0,0	1,1

Figure 2: Coordination game

developed in the future, will consist in adding to our logic the beliefs of agents, and ways of influencing their action by changing their beliefs. By extending our logical theory with beliefs of agents, we will also be able to represent the logical relationship between (rational) choices

[4]With 'choice set' we mean all options that an agent takes into consideration when deciding what to do. This typically coincides with the set of alternatives that the agent believes to be possible means for achieving her goals. Therefore, an agent can *indirectly* affect the composition of another agent's choice set by modifying her beliefs about what is a possible means for achieving her goals. However, this intermediate step is not explicitly modeled in our logical analysis of social influence.

of agents and their beliefs about the choices of the others.[5]

To illustrate the concept of social influence, let us consider two examples, the first about influence via choice set restriction and the second one about influence via payoff change. We do not give an example about influence via choice set expansion, as it is very similar to influence via choice set restriction. The example of influence via choice restriction is illustrated in Figure 3. The example represents a situation where there are three fruits on a table, an apple, a banana and a pear. The actions at issue consist in bringing about that the apple is eaten (ap), the banana is eaten (ba) or the pear is eaten (pe). Let us assume that agent 2 has certain preferences that remain constant along the tree structure. In particular, at all moments agent 2 prefers eating apples to bananas to pears. Let us also assume that 2 is rational, in the sense that she acts in such a way as to achieve the outcome she prefers. Rational choices of agent 2 are depicted in grey. By choosing to eat the apple at w_1, 1 generates a situation where, given her preferences, 2 will necessarily eat the banana, rather than the pear. Indeed, although at moment m_2, 2 has two choices available, namely the choice of eating the banana and the choice of eating the pear, only the former is rational, in the sense of being compatible with 2's preferences. In this sense, by deciding to eat the apple at w_1 and removing this option from 2's choice set, 1 influences 2 to decide to eat the banana at w_7.[6]

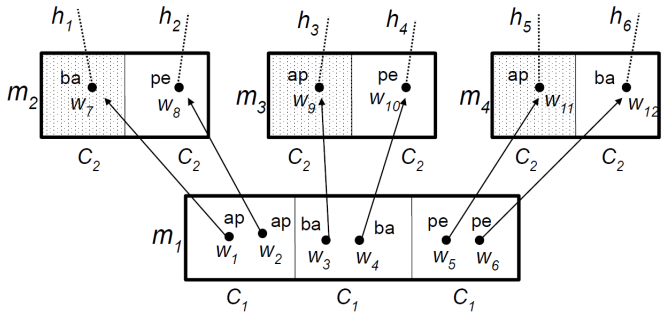


Figure 3: Example of influence via choice restriction

The example of influence via payoff change is illustrated in Figure 4. The example represents a situation where agent 1 has to decide whether to give a monetary incentive to agent 2

[5]For an extension of a STIT-like framework by agents' epistemic attitudes (e.g., beliefs and knowledge) and the formalization of rationality principles connecting choices of agents and their beliefs about choices of the others, see [27].

[6]Notice that this sounds like regimentation which, according to [20], consists in forcing or precluding certain actions. However there is a subtle difference between our concept of 'influence via choice set modification' and the concept of regimentation. On the one hand, regimentation operates on the executability preconditions of an action (i.e., whether an action is *practically* feasible or not). On the other hand, influence operates on an agent's choice set (i.e., whether an action is *choosable* or not).

for the purchase of an electric car rather than a diesel car. Let us assume that agent 2 prefers a diesel car to an electric car without monetary incentive and prefers an electric car to a diesel car in the presence of a monetary incentive of 3000 euros. At moment m_1, agent 2 can decide either to give the incentive of 3000 euros (in) or not to give it (no). It turns out that if 1 gives the incentive then in the next moment m_2 the only rational choice for agent 2 is to buy an electric car (el), whereas if 2 does not give the incentive then in the next moment m_3 the only rational choice for agent 2 is to buy a diesel car (ds). In this sense, by deciding to give an incentive at w_1 and changing the payoffs associated to 2's choice of buying an electric car, 1 influences 2 to buy an electric car at w_5.

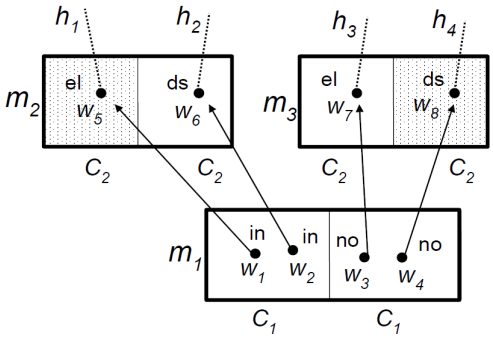


Figure 4: Example of influence via payoff change

These examples lead us to the following informal definition of social influence:

> An agent i influences another agent j to perform a certain (voluntary) action if and only if, i sees to it that that every rational choice of j will lead j to perform the action.

Before concluding this section, let us emphasize an important aspect about the compatibility between the STIT semantics and the concept of influence. An important feature of STIT logic is its being characterized by the notion of independence of choices (see Section 4.2 below for more details). That is, in STIT choices are taken by each agent independently from the *concomitant* choices of the others. From this perspective, STIT semantics and the concept of influence would seem incompatible. However, our approach is consistent with the STIT semantics, since we assume that the influencer's move *temporally precedes* the influencee's move, determining the range of choices available to the latter. In other words, the influencer's choice at a given moment determines the influencee's range of choices at a *later* moment.

In the next section we present a logic which enables us to formalize the previous concept of social influence. Specifically, it enables us to represent both aspects of social influence relations: the influencee's freedom to select the action she prefers in her choice set, and the influencer's ability of determining the influencee's choice by modifying the influencee's choice set.

4 DR-STIT logic

Our logic is a variant of STIT with discrete time and rational choices interpreted in Kripke semantics. The name of our logic is DR-STIT, which stands for "STIT with Deterministic time and Rational choices". On the syntactic level, DR-STIT is nothing but the extension of atemporal individual STIT by: (i) the temporal operators 'next' (tomorrow) and 'previous' (yesterday) of linear temporal logic, (ii) the operator of group agency for the *grand coalition* (the coalition of all agents), and (iii) special operators of agency describing the consequences of an agent's rational choice. In Section 6 the logic will be used to define the concept of social influence that we have informally discussed in Section 3.

In DR-STIT the so-called Chellas STIT operators are taken as primitive operators of agency. As pointed out by [19], deliberative STIT operators and Chellas STIT operators are interdefinable and just differ in the choice of primitive operators. The language of DR-STIT has some similarities with the language of the logic XSTIT studied by Broersen [7], as they both limit themselves to the 'next' time operator without considering more complex tense operators. However, there is an important difference between Broersen's XSTIT and our logic DR-STIT. In XSTIT the temporal dimension and the agency dimension are fused to make up a single modal operator. In particular, in XSTIT there are primitive operators describing the effects of an agent's action in 'next' states, where 'next' refers to immediate successors of the present state. On the contrary, in DR-STIT the temporal dimension and the agency dimension are kept separated, as agency operators are distinguished from temporal operators.

The following two sections present the syntax and a Kripke-style semantics for DR-STIT (Subsections 4.1 and 4.2).

4.1 Syntax

Assume a countable (possibly infinite) set of atomic propositions denoting facts $Atm = \{p, q, \ldots\}$ and a finite set of agents $Agt = \{1, \ldots, n\}$.

The language $\mathcal{L}_{\text{DR-STIT}}(Atm, Agt)$ of the logic DR-STIT is the set of formulae defined by the following BNF:

$$\begin{aligned} \varphi \quad ::= \quad & p \mid \neg\varphi \mid \varphi \wedge \varphi \mid \Box\varphi \mid \mathbf{X}\varphi \mid \mathbf{Y}\varphi \mid \\ & [i\ \mathbf{stit}]\varphi \mid [Agt\ \mathbf{stit}]\varphi \mid [i\ \mathbf{rstit}]\varphi \end{aligned}$$

where p ranges over Atm and i ranges over Agt. The other Boolean constructions $\top$, $\bot$, $\vee$, $\rightarrow$ and $\leftrightarrow$ are defined from $\neg$ and $\wedge$ in the standard way.

Operators of the form $[i\ \mathbf{stit}]$ are Chellas STIT operators that have been informally discussed in Section 2. Formula $[i\ \mathbf{stit}]\varphi$ captures the fact that φ is guaranteed by a present action of agent i, and has to be read 'agent i sees to it that φ regardless of what the other agents do'. We shorten the reading of $[i\ \mathbf{stit}]\varphi$ to 'agent i sees to it that φ'. $[Agt\ \mathbf{stit}]$ is a group STIT operator which captures the fact that φ is guaranteed by a present choice of all agents, and has to be read 'all agents see to it that φ by acting together'. The dual operators of $[i\ \mathbf{stit}]$ and $[Agt\ \mathbf{stit}]$ are defined in the usual way:

$$\begin{aligned} \langle i\ \mathbf{stit}\rangle\varphi &\stackrel{\text{def}}{=} \neg[i\ \mathbf{stit}]\neg\varphi, \\ \langle Agt\ \mathbf{stit}\rangle\varphi &\stackrel{\text{def}}{=} \neg[Agt\ \mathbf{stit}]\neg\varphi. \end{aligned}$$

Operators of the form $[i\ \mathbf{rstit}]$ describe the effects of a rational (or preferred) choice of agent i. Following one of the arguments of Section 3, namely the idea that a choice is rational if it is consistent with the agent's preference over her alternative choices, we here conceive the terms 'rational choice' and 'preferred choice' as synonyms. Specifically, formula $[i\ \mathbf{rstit}]\varphi$ has to be read either "if agent i's current action is the result of a rational choice of i, then i sees to it that φ" or, "if agent i's current choice is a preferred choice of i, then i sees to it that φ". So, $[i\ \mathbf{rstit}]\varphi$ is true for every formula φ when agent i's current choice is not rational.[7] The dual of the operator $[i\ \mathbf{rstit}]$ is defined as follows: $\langle i\ \mathbf{rstit}\rangle\varphi \stackrel{\text{def}}{=} \neg[i\ \mathbf{rstit}]\neg\varphi$. The formula $\langle i\ \mathbf{rstit}\rangle\top$ has to be read either "agent i's current choice is rational" or, "agent i's current choice is a preferred choice of i".[8] The formula $[i\ \mathbf{stit}]\varphi \wedge \langle i\ \mathbf{rstit}\rangle\top$, which is logically equivalent to $[i\ \mathbf{rstit}]\varphi \wedge \langle i\ \mathbf{rstit}\rangle\top$, has to be read either "agent i sees to it that φ as a result of her rational choice" or, "agent i sees to it that φ as a result of her preferred choice". Notice that this concept of preference is binary in the sense that a choice is either preferred or not and there is no ordering over choices involved.

$\Box\varphi$ stands for 'φ is true regardless of what every agent does' or 'φ is true no matter what the agents do' or simply 'φ is necessarily true'. We define the dual of $\Box$ as follows: $\Diamond\varphi \stackrel{\text{def}}{=} \neg\Box\neg\varphi$. Note that the operators $[i\ \mathbf{stit}]$ and $\Box$ can be combined in order define the deliberative STIT operator $[i\ \mathbf{dstit}]$ we have discussed in Section 2:

$$[i\ \mathbf{dstit}]\varphi \stackrel{\text{def}}{=} [i\ \mathbf{stit}]\varphi \wedge \neg\Box\varphi.$$

Moreover, the operators $[i\ \mathbf{rstit}]$ and $\Box$ can be combined in order define a special kind of deliberative STIT operator for rational choices:

$$[i\ \mathbf{rdstit}]\varphi \stackrel{\text{def}}{=} \langle i\ \mathbf{rstit}\rangle\top \rightarrow [i\ \mathbf{dstit}]\varphi.$$

$[i\ \mathbf{rdstit}]\varphi$ has to be read either "if agent i's current action is the result of a rational choice of i, then i *deliberately* sees to it that φ" or, "if agent i's current choice is a preferred choice of i, then i *deliberately* sees to it that φ". The operator $[i\ \mathbf{rdstit}]$ is a conditional form of deliberative STIT operator which represents the consequences of an agent's action *under the condition that* this action is the result of a rational choice of the agent.

Finally, $\mathbf{X}$ and $\mathbf{Y}$ are the standard operators 'next' (tomorrow) and 'previous' (yesterday) of linear temporal logic. $\mathbf{X}\varphi$ and $\mathbf{Y}\varphi$ has to be read respectively 'φ is going to be true in the next world' and 'φ was true in the previous world'.

Before concluding this section, let us justify the use of the different operators.

In the light of the results presented in [1] the historical necessity operator $\Box$ could be omitted, as the formula $\Box\varphi$ and $[i\ \mathbf{stit}][j\ \mathbf{stit}]\varphi$ are logically equivalent, when $i \neq j$. So, the only reason why we decided to add it to the language of the logic DR-STIT is conceptual. We believe that it is important to have the historical necessity operator as a primitive object in the formal language, as it captures something conceptually different from 'agency', as captured by the operator $[i\ \mathbf{stit}]$.

[7] Our notion of "doing something as the result of a rational (or preferred) choice" is synonym of List & Rabinowicz's notion of "doing an action with endorsement" [23].

[8] The formula $\langle i\ \mathbf{rstit}\rangle\top$ looks similar to the formula $\langle\varphi\rangle\top$ of public announcement logic (PAL) [31] meaning that "the public announcement of φ is executable".

As for the grand coalition operator $[Agt\ \mathbf{stit}]$, the reason for having it in the language is mainly technical. As we will show in Section 5, by means of this operator we can syntactically express a fundamental property of STIT models, the so-called property of *no choice between undivided histories*, capturing the basic connection between action and time (see the next section for further details).

Finally, the temporal operators $\mathbf{X}$ and $\mathbf{Y}$ are used here for several reasons. As for the operator $\mathbf{X}$, we use it in order to capture the temporal aspect of social influence, namely the fact that the influencer's action temporally precedes the influencee's choice. As for the operator $\mathbf{Y}$, we have decided to include it both for technical and for conceptual reasons. This operator allows to syntactically express some basic temporal properties of STIT models such as the fact that the temporal relation for the past is assumed to be deterministic, i.e., every world has at most one temporal predecessor (see the next section for more details). A second reason for having the operator $\mathbf{Y}$ in the object language is that we want to stay as close as possible to the STIT tradition which typically considers temporal STIT languages including both 'forward looking' and 'backward looking' operators. For instance, the temporal STIT language studied by Horty [18] includes both the future tense operator $\mathbf{G}$ (with $\mathbf{G}\varphi$ meaning "φ will always be true in the future") and the past tense operator $\mathbf{H}$ (with $\mathbf{H}\varphi$ meaning "φ has always been true in the past"). Finally, the presence of the operator $\mathbf{Y}$ in the object language yields to interesting perspectives of future research such as the possibility of representing a variant of social influence based on the concept of *achievement* STIT à la Belnap & Perloff [2] whose formal characterization requires a 'backward looking' operator such as the operator $\mathbf{Y}$.

4.2 Kripke semantics for DR-STIT

The basic notion in the semantics is the notion of Kripke STIT model with discrete time and rational choices. For notational convenience, in what follows we are going to use the following abbreviations. Given a set of elements W, an arbitrary binary relation $\mathcal{R}$ on W and an element w in W, let $\mathcal{R}(w) = \{v \in W | w\mathcal{R}v\}$. Moreover, given two binary relations $\mathcal{R}_1$ and $\mathcal{R}_2$ on W let $\mathcal{R}_1 \circ \mathcal{R}_2$ be the standard operation of composition between binary relations. If we abstract away from the notion of rational choice, Kripke STIT models with discrete time and rational choices are nothing but a subclass of the general class of temporal Kripke STIT models studied by [24]. The latter can be seen as extensions of Zanardo's Ockhamist models [41] by a choice component, i.e., by accessibility relations for individual choices and an accessibility relation for the collective choice of the grand coalition Agt.

Definition 1 *A Kripke STIT model with discrete time and rational choices is a tuple* $M = (W, \equiv, \rightarrow, \leftarrow, \{\sim_i\}_{i \in Agt}, \sim_{Agt}, \{RC_i\}_{i \in Agt}, \mathcal{V})$ *where:*

- W *is a nonempty set of possible worlds;*
- $\equiv$ *is an equivalence relation on* W*;*
- $\rightarrow$ *is a serial and deterministic relation on* W*;*
- $\leftarrow$ *is the inverse relation of* $\rightarrow$ *(i.e.,* $\leftarrow = \{(w, v) | v \rightarrow w\}$*) and is supposed to be deterministic;*
- $\sim_{Agt}$ *and every* $\sim_i$ *are equivalence relations on* W*;*

- *for all $i \in Agt$, $RC_i \subseteq C_i$ with C_i being the partition of W induced by the equivalence relation $\sim_i$;*
- *$\mathcal{V} : W \longrightarrow 2^{Atm}$ is a valuation function for atomic propositions;*

and that satisfies the following six constraints:

(C1) *for all $w, v \in W$: if $w \mathcal{F} v$ then $w \not\equiv v$, with $\mathcal{F}$ denoting the transitive closure of the binary relation $\rightarrow\!\!\triangleright$;*

(C2) *for all $i \in Agt$: $\sim_i \subseteq \equiv$;*

(C3) *for all $u_1, \ldots, u_n \in W$: if $u_i \equiv u_j$ for all $i, j \in \{1, \ldots, n\}$ then $\bigcap_{1 \leq i \leq n} \sim_i (u_i) \neq \emptyset$;*

(C4) *for all $w \in W$ and for all $i \in Agt$: there exists $v \in W$ such that $w \equiv v$ and $\sim_i (v) \in RC_i$;*

(C5) *for all $w \in W$: $\sim_{Agt} (w) = \bigcap_{i \in Agt} \sim_i (w)$;*

(C6) $\rightarrow\!\!\triangleright \circ \equiv \subseteq \sim_{Agt} \circ \rightarrow\!\!\triangleright$.

Let us explain in detail each component of the preceding definition.

$\equiv (w)$ is the set of worlds that are alternative to the world w. Following the Ockhamist's view of time [34, 41], we call the equivalence classes induced by the equivalence relation $\equiv$ *moments*. The set of all moments in the model M is denoted by Mom and the elements in Mom are denoted by $m, m', \ldots$

$\rightarrow\!\!\triangleright (w)$ is the set of direct temporal successors of world w, that is to say, $w \rightarrow\!\!\triangleright v$ means that v is in the future of w and there is no third world that is in the future of w and in the past of v. The fact that $\rightarrow\!\!\triangleright$ is serial and deterministic means that every world has *exactly one* direct temporal successor. $\triangleleft\!\!\leftarrow (w)$ defines the set of direct temporal predecessors of world w, that is to say, $v \triangleleft\!\!\leftarrow w$ means that v is in the past of w and there is no third world that is in the past of w and in the future of v. The fact that $\triangleleft\!\!\leftarrow$ is deterministic means that every world has *at most one* direct temporal predecessor. We do not assume $\triangleleft\!\!\leftarrow$ to be serial because past is not necessarily endless.

The Constraint C1 in Definition 1 ensures that if two worlds belong to the same moment then one of them cannot be in the future of the other. (Note that $\mathcal{F}(w)$ is the set of worlds that are in the future of w.) Since the relation $\equiv$ is reflexive, the Constraint C1 implies that the relations $\rightarrow\!\!\triangleright$, $\triangleleft\!\!\leftarrow$ and $\mathcal{F}$ are all irreflexive.

Let $\mathcal{T}(w) = \mathcal{P}(w) \cup \{w\} \cup \mathcal{F}(w)$ be the set of worlds that are temporally related with world w, where $\mathcal{P} = \{(w, v) | v \mathcal{F} w\}$ is the inverse of the relation $\mathcal{F}$ and $\mathcal{P}(w)$ is the set of worlds that are in the past of w. The fact that the relation $\mathcal{F}$ is irreflexive and transitive ensures that $\mathcal{F}$ is a strict linear (or total) order on the set $\mathcal{T}(w)$. For every world w in W, we call the linearly ordered set $(\mathcal{T}(w), \mathcal{F})$ the history going through w. Because of the seriality of the relation $\rightarrow\!\!\triangleright$, every history is infinite. Note that the concept of history and the assumption that histories are infinite allow us to formally characterize chains of influence of any arbitrary length, namely the fact an agent i_1 influences an agent i_2 to influence an agent i_3 to influence...an agent i_m to make φ true. The formal properties of chains of influence of length 2 will be studied in Section 6.

For notational convenience, let $Hist$ denote the set of all histories in the model M and let the elements of $Hist$ be denoted by $h, h', \ldots$ Moreover, for every moment $m \in Mom$, let

$$Hist_m = \{h \in Hist | \exists w \in W \text{ such that } w \in m \cap h\}$$

be the set of all histories passing through the moment m and let

$$Ind = \{m/h | m \in Mom \text{ and } h \in Hist_m\}$$

be the set of all indexes in the model M.

Clearly, our Kripke semantics for DR-STIT allows us to interchangeably use the term 'world' and 'history going through a certain world' without lost of generality. Indeed, for every world w there exists a unique history going through it. This point is highlighted by the following proposition.

Proposition 1 *Let $w \in W$. Then, there exists a unique $h \in Hist$ such that $w \in h$.*

As every world in a model is identified with a unique history going through it, the equivalence relation $\equiv$ can also be understood as an equivalence relation between historic alternatives: $w \equiv v$ means that the history going through v is alternative to the history going through w, or the history going through w and the history going through v pass through the same moment.

Furthermore, in the semantics for DR-STIT there is a one-to-one correspondence between worlds and indexes, as for every index m/h in Ind there exists a unique world w at the intersection between m and h, and for every world w there exists a unique index m/h such that the intersection between m and h includes w. This fact is highlighted by the following two propositions.

Proposition 2 *Let $m \in Mom$ and let $h \in Hist_m$. Then, $m \cap h$ is a singleton.*

Proposition 3 *Let $w \in W$. Then, there exists a unique $m/h \in Ind$ such that $w \in m \cap h$.*

For every world w, the set $\sim_i (w)$ identifies agent i's *actual* choice at w, that is to say, the set of worlds that can be obtained by agent i's actual choice at w. Because of the one-to-one correspondence between worlds and histories, one can also identify i's *actual* choice at w with the set $\{h \in Hist | \exists v \in \sim_i (w) \text{ such that } v \in h\}$. In other words, in DR-STIT an agent chooses among different sets of histories.

Constraint C2 in Definition 1 just means that an agent can only choose among possible alternatives. This constraint ensures that, for every world w, the equivalence relation $\sim_i$ induces a partition of the set $\equiv(w)$. An element of this partition is a choice that is *possible (or available)* for agent i at w.

Constraint C3 expresses the so-called assumption of *independence of agents* or *independence of choices*: if $\sim_1 (u_1)$ is a possible choice for agent 1 at w, $\sim_2 (u_2)$ is a possible choice for agent 2 at w,..., $\sim_n (u_n)$ is a possible choice for agent n at w, then their intersection is non-empty. More intuitively, this means that agents can never be deprived of choices due to the *concomitant* choices made by other agents. As emphasized in Section 3, this property of the STIT semantics does not negate the whole possibility of influence, as a choice of an agent i at a given moment m might determine the range of choices available to an agent j at a *later* moment m.

Let C_i denote the partition of the set of worlds W induced by the equivalence relation $\sim_i$. This partition characterizes agent i's set of choices in the model M. The set $RC_i \subseteq C_i$ characterizes agent i's set of *rational* choices or also, the set of *preferred* choices of agent i (given the assumption that a choice is rational if it is consistent with the agent's preference over

her alternative choices). The Constraint C4 just means that, at each moment, an agent has at least one rational choice available.

For every world w, the set $\sim_{Agt}(w)$ identifies the *actual* choice of group Agt at w, that is to say, the set of worlds that can be obtained by the collective choice of all agents at w. Constraint C5 just says that the collective choice of the grand coalition Agt is equal to the intersection of the choices of all individuals. This corresponds to the notion of joint action proposed by Horty in [18], where the joint action of a group is described in terms of the result that the agents in the group bring about by acting together.

The Constraint C6 expresses a basic relation between action and time: if v is in the future of w and u and v are in the same moment, then there exists an alternative z in the collective choice of all agents at w such that u is in the future of z. This constraint corresponds to the property of *no choice between undivided histories* given in STIT logic [3, Chap. 7]. It captures the idea that if two histories come together in some future moment then, in the present, each agent does not have a choice between these two histories. This implies that if an agent can choose between two histories at a later stage, then she does not have a choice between them in the present.

A formula φ of the logic DR-STIT is evaluated with respect to a a Kripke STIT model with discrete time and rational choices $M = (W, \equiv, \rightarrow, \leftarrow, \{\sim_i\}_{i\in Agt}, \sim_{Agt}, \{RC_i\}_{i\in Agt}, \mathcal{V})$ and a world w in M. We write $M, w \models \varphi$ to mean that φ is true at world w in M. The truth conditions of DR-STIT formulae are then defined as follows:

$$
\begin{array}{rcl}
M, w \models p & \iff & p \in \mathcal{V}(w) \\
M, w \models \neg\varphi & \iff & M, w \not\models \varphi \\
M, w \models \varphi \wedge \psi & \iff & M, w \models \varphi \text{ AND } M, w \models \psi \\
M, w \models \Box\varphi & \iff & \forall v \in \equiv(w) : M, v \models \varphi \\
M, w \models \mathbf{X}\varphi & \iff & \forall v \in \rightarrow(w) : M, v \models \varphi \\
M, w \models \mathbf{Y}\varphi & \iff & \forall v \in \leftarrow(w) : M, v \models \varphi \\
M, w \models [i\ \mathbf{stit}]\varphi & \iff & \forall v \in \sim_i(w) : M, v \models \varphi \\
M, w \models [Agt\ \mathbf{stit}]\varphi & \iff & \forall v \in \sim_{Agt}(w) : M, v \models \varphi \\
M, w \models [i\ \mathbf{rstit}]\varphi & \iff & \text{IF } \sim_i(w) \in RC_i \text{ THEN} \\
 & & \forall v \in \sim_i(w) : M, v \models \varphi
\end{array}
$$

Notice that the operators $[i\ \mathbf{stit}]$, $[Agt\ \mathbf{stit}]$, and $[i\ \mathbf{rstit}]$ are instantaneous action operators as they describe the consequences of choices at the actual moment.

Since the relation $\rightarrow$ is serial and deterministic, we can specify a total (successor) function $succ : W \rightarrow W$ such that $succ(w) = v$ if and only if $v \in \rightarrow(w)$. Moreover, since $\leftarrow$ is deterministic we can specify a partial (predecessor) function $pred : W \nrightarrow W$ such that $pred(w) = v$ if and only if $v \in \leftarrow(w)$. This allows to simplify the truth conditions of the operators $\mathbf{X}$ and $\mathbf{Y}$ as follows:

$$
\begin{array}{rcl}
M, w \models \mathbf{X}\varphi & \iff & M, succ(w) \models \varphi \\
M, w \models \mathbf{Y}\varphi & \iff & \text{IF } pred(w) \text{ is defined THEN} \\
 & & M, pred(w) \models \varphi
\end{array}
$$

For any formula φ of the language $\mathcal{L}_{\text{DR-STIT}}(Atm, Agt)$, we write $\models_{\text{DR-STIT}} \varphi$ if φ is DR-STIT *valid*, i.e., for all Kripke STIT models with discrete time and rational choices M

and for all worlds w in M, we have $M, w \models \varphi$. We say that φ is DR-STIT *satisfiable* if $\neg\varphi$ is not DR-STIT valid.

5 Axiomatization

Figure 5 contains a complete axiomatization of the logic DR-STIT with respect to the class of Kripke STIT models with discrete time and rational choices.

PC	Tautologies of classical propositional calculus
S5($\Box$)	All S5-principles for the operator $\Box$
KD(X)	All KD-principles for the operator $\mathbf{X}$
K(Y)	All K-principles for the operator $\mathbf{Y}$
S5($[i\ \mathbf{stit}]$)	All S5-principles for the operators $[i\ \mathbf{stit}]$
S5($[Agt\ \mathbf{stit}]$)	All S5-principles for the operator $[Agt\ \mathbf{stit}]$
(Alt$_\mathbf{X}$)	$\neg\mathbf{X}\varphi \rightarrow \mathbf{X}\neg\varphi$
(Alt$_\mathbf{Y}$)	$\neg\mathbf{Y}\varphi \rightarrow \mathbf{Y}\neg\varphi$
(Conv$_{\mathbf{X},\mathbf{Y}}$)	$\varphi \rightarrow \mathbf{XY}\varphi$
(Conv$_{\mathbf{Y},\mathbf{X}}$)	$\varphi \rightarrow \mathbf{YX}\varphi$
(Rel$_{\Box,[i\ \mathbf{stit}]}$)	$\Box\varphi \rightarrow [i\ \mathbf{stit}]\varphi$
(AIA)	$(\Diamond[1\ \mathbf{stit}]\varphi_1 \wedge \ldots \wedge \Diamond[n\ \mathbf{stit}]\varphi_n) \rightarrow$ $\Diamond([1\ \mathbf{stit}]\varphi_1 \wedge \ldots \wedge [n\ \mathbf{stit}]\varphi_n)$
(Rel$_{[i\ \mathbf{rstit}],[i\ \mathbf{stit}]}$)	$\langle i\ \mathbf{rstit}\rangle\top \rightarrow ([i\ \mathbf{rstit}]\varphi \leftrightarrow [i\ \mathbf{stit}]\varphi)$
(RatCh)	$\langle i\ \mathbf{rstit}\rangle\top \rightarrow [i\ \mathbf{stit}]\langle i\ \mathbf{rstit}\rangle\top$
(OneRat)	$\Diamond\langle i\ \mathbf{rstit}\rangle\top$
(Rel$_{[i\ \mathbf{stit}],[Agt\ \mathbf{stit}]}$)	$([1\ \mathbf{stit}]\varphi_1 \wedge \ldots \wedge [n\ \mathbf{stit}]\varphi_n) \rightarrow$ $[Agt\ \mathbf{stit}](\varphi_1 \wedge \ldots \wedge \varphi_n)$
(NCUH)	$\mathbf{X}\Diamond\varphi \rightarrow \langle Agt\ \mathbf{stit}\rangle\mathbf{X}\varphi$
(MP)	$\dfrac{\varphi,\ \varphi \rightarrow \psi}{\psi}$

Figure 5: Sound and complete axiomatization of DR-STIT

This includes all tautologies of classical propositional calculus **(PC)** as well as modus ponens **(MP)**. Moreover, we have all principles of the normal modal logic S5 for every operator $[i\ \mathbf{stit}]$, for the operator $[Agt\ \mathbf{stit}]$ and for the operator $\Box$, all principles of the normal modal logic KD for the temporal operator $\mathbf{X}$ and all principles of the normal modal logic K for the temporal operator $\mathbf{Y}$. That is, we have Axiom K for each operator: $(\blacksquare\varphi \wedge \blacksquare(\varphi \rightarrow \psi)) \rightarrow \blacksquare\psi$ with $\blacksquare \in \{\Box, \mathbf{X}, \mathbf{Y}, [Agt\ \mathbf{stit}]\} \cup \{[i\ \mathbf{stit}] | i \in Agt\}$. We have Axiom D for the temporal operator $\mathbf{X}$: $\neg(\mathbf{X}\varphi \wedge \mathbf{X}\neg\varphi)$. We have Axiom 4 for $\Box$, $[Agt\ \mathbf{stit}]$ and for every $[i\ \mathbf{stit}]$: $\blacksquare\varphi \rightarrow \blacksquare\blacksquare\varphi$ with $\blacksquare \in \{\Box, [Agt\ \mathbf{stit}]\} \cup \{[i\ \mathbf{stit}] | i \in Agt\}$. Furthermore, we have Axiom T for $\Box$, $[Agt\ \mathbf{stit}]$ and for every $[i\ \mathbf{stit}]$: $\blacksquare\varphi \rightarrow \varphi$ with $\blacksquare \in \{\Box, [Agt\ \mathbf{stit}]\} \cup \{[i\ \mathbf{stit}] | i \in$

$Agt\}$. We have Axiom B for $\Box$, $[Agt\ \mathbf{stit}]$ and for every $[i\ \mathbf{stit}]$: $\varphi \rightarrow \blacksquare\neg\blacksquare\neg\varphi$ with $\blacksquare \in \{\Box, [Agt\ \mathbf{stit}]\} \cup \{[i\ \mathbf{stit}] | i \in Agt\}$. Finally we have the rule of necessitation for each modal operator: $\frac{\varphi}{\blacksquare\varphi}$ with $\blacksquare \in \{\Box, [Agt\ \mathbf{stit}], \mathbf{X}, \mathbf{Y}\} \cup \{[i\ \mathbf{stit}] | i \in Agt\}$.

We have principles for temporal operators and for the relationship between time and action. **($\mathbf{Alt_X}$)** and **($\mathbf{Alt_Y}$)** are the basic axioms for the determinism of 'tomorrow' and 'yesterday'. **($\mathbf{Conv_{X,Y}}$)** and **($\mathbf{Conv_{Y,X}}$)** are the basic interaction axioms between 'tomorrow' and 'yesterday' according to which "if φ is true in the present, then tomorrow is going to be true that yesterday φ has been true" and "if φ is true in the present, then yesterday has been true that tomorrow φ is going to be true".

($\mathbf{Rel}_{\Box,[i\ \mathbf{stit}]}$) and **(AIA)** are the two central principles in Xu's axiomatization of the Chellas STIT operators $[i\ \mathbf{stit}]$ [40]. According to Axiom **($\mathbf{Rel}_{\Box,[i\ \mathbf{stit}]}$)**, if φ is true regardless of what every agent does, then every agent sees to it that φ. In other words, an agent brings about those facts that are inevitable. According to Axiom **($\mathbf{Rel}_{[i\ \mathbf{stit}],[Agt\ \mathbf{stit}]}$)**, all agents bring about together what each of them brings about individually. The principles **S5**($\Box$) and **S5**($[i\ \mathbf{stit}]$) together with the Axioms **($\mathbf{Rel}_{\Box,[i\ \mathbf{stit}]}$)** and **(AIA)** and the rule of inference **(MP)** constitute the axiomatics of individual STIT logic as given in [1].

Axiom **(NCUH)** establishes a fundamental relationship between action and time and corresponds to the semantic constraint of 'no choice between undivided histories' (Constraint C6 in Definition 1): if in the next world φ is going to be possible then the actual collective choice of all agents will possibly result in a world in which φ is true. As emphasized in Section 4.1, this axiom requires the grand coalition operator $[Agt\ \mathbf{stit}]$.

Axiom **($\mathbf{Rel}_{[i\ \mathbf{rstit}],[i\ \mathbf{stit}]}$)** is the basic interaction principle between the rational choice operator $[i\ \mathbf{rstit}]$ and the Chellas STIT operator $[i\ \mathbf{stit}]$. It highlights that the Chellas STIT operator and the rational choice STIT operator coincide under the condition that the agent's current choice is rational, expressed by the formula $\langle i\ \mathbf{rstit}\rangle\top$.

Axiom **(RatCh)** means that the rationality of an agent just depends on her actual choice: if agent i is rational then i sees to it that she is rational. Given Axiom **($\mathbf{Rel}_{[i\ \mathbf{rstit}],[i\ \mathbf{stit}]}$)**, Axiom **(RatCh)** could be replaced by the equivalent formula $\langle i\ \mathbf{rstit}\rangle\top \rightarrow [i\ \mathbf{rstit}]\langle i\ \mathbf{rstit}\rangle\top$.

Finally, according to Axiom **(OneRat)**, an agent always has a rational choice in her repertoire.

Theorem 1 *The set of DR-STIT validities is completely axiomatized by the principles given in Figure 5.*

The proof of Theorem 1 is given in the Annex A at the end of the paper.

6 Formalization of influence

We can now get back to the main issue of the paper, namely the problem of modeling the concept of social influence in STIT. Let us consider the following definition of social influence:

> An agent i influences another agent j to perform a certain (voluntary) action if and only if, i sees to it that j will perform the action.

This definition of social influence is problematic for two reasons: (i) if an agent i sees to it that some state of affairs φ will be true, then φ will *necessarily* be true after i's choice, and (ii) necessity is incompatible with action. Indeed, as emphasized in Section 2, the performance of

an action by agent j producing the state of affairs φ, presupposes that this action could have not taken place and that φ possibly would not hold, which means that the action of agent j was not necessary. This point is made clear by the following validity of our logic DR-STIT, that is also a validity of STIT in general. For all $i, j \in Agt$, we have:

$$\models_{\text{DR-STIT}} \neg[i\ \mathbf{stit}]\mathbf{X}[j\ \mathbf{dstit}]\varphi.$$

Proof. Let us provide the Hilbert-style proof of this DR-STIT validity by means of the proof calculus given in Section 5:

1. $\vdash$ $[i\ \mathbf{stit}]\mathbf{X}[j\ \mathbf{dstit}]\varphi \leftrightarrow [i\ \mathbf{stit}]\mathbf{X}([j\ \mathbf{stit}]\varphi \wedge \Diamond\neg\varphi)$
2. $\vdash$ $[i\ \mathbf{stit}]\mathbf{X}([j\ \mathbf{stit}]\varphi \wedge \Diamond\neg\varphi) \rightarrow [Agt\ \mathbf{stit}]\mathbf{X}([j\ \mathbf{stit}]\varphi \wedge \Diamond\neg\varphi)$
   By Axiom $\mathbf{Rel}_{[i\ \mathbf{stit}],[Agt\ \mathbf{stit}]}$
3. $\vdash$ $[Agt\ \mathbf{stit}]\mathbf{X}([j\ \mathbf{stit}]\varphi \wedge \Diamond\neg\varphi) \rightarrow \mathbf{X}\Box([j\ \mathbf{stit}]\varphi \wedge \Diamond\neg\varphi)$
   By Axiom **NCUH**
4. $\vdash$ $\mathbf{X}\Box([j\ \mathbf{stit}]\varphi \wedge \Diamond\neg\varphi) \rightarrow \mathbf{X}\Box(\varphi \wedge \Diamond\neg\varphi)$
   By Axiom T for $[j\ \mathbf{stit}]$, Axiom K and necessitation for $\mathbf{X}$ and $\Box$
5. $\vdash$ $\mathbf{X}\Box(\varphi \wedge \Diamond\neg\varphi) \rightarrow \mathbf{X}(\Box\varphi \wedge \Box\Diamond\neg\varphi)$
   By Axiom K for $\Box$, and Axiom K and necessitation for $\mathbf{X}$
6. $\vdash$ $\mathbf{X}(\Box\varphi \wedge \Box\Diamond\neg\varphi) \rightarrow \mathbf{X}(\Box\varphi \wedge \Diamond\neg\varphi)$
   By Axiom T for $\Box$, and Axiom K and necessitation for $\mathbf{X}$
7. $\vdash$ $\mathbf{X}(\Box\varphi \wedge \Diamond\neg\varphi) \rightarrow \bot$
   By Axiom D for $\mathbf{X}$
8. $\vdash$ $\neg[i\ \mathbf{stit}]\mathbf{X}[j\ \mathbf{dstit}]\varphi$
   From 1-7

This means that agent i cannot see to it that in the next world agent j deliberately sees to it that some state of affairs φ is true. As the preceding proof highlights, this is a consequence of no choice between undivided histories (Constraint C6 in Definition 1) and of the corresponding Axiom **NCUH**. As a side note, we observe that $\neg[i\ \mathbf{dstit}]\mathbf{X}[j\ \mathbf{dstit}]\varphi$ is valid because $[i\ \mathbf{dstit}]\mathbf{X}[j\ \mathbf{dstit}]\varphi$ implies $[i\ \mathbf{stit}]\mathbf{X}[j\ \mathbf{dstit}]\varphi$. Moreover, formulae $\neg[i\ \mathbf{stit}][j\ \mathbf{dstit}]\varphi$ and $\neg[i\ \mathbf{dstit}][j\ \mathbf{dstit}]\varphi$ are valid as well, when $i \neq j$. These two validities are consequences of independence of choices (Constraint C3 in Definition 1).

However, a non-problematic notion of social influence can be expressed in our logic DR-STIT by means of the special operators $[i\ \mathbf{rdstit}]$. Indeed, these operators allow us to formally represent the idea we have discussed in Section 3, namely that the influencer induces the influencee to perform a certain action by constraining her choice set in such a way that the choice that the influencer wants to be chosen is exactly the one that the influencee would choose, given her preferences. Specifically, we shall say that an agent i influences another agent j to make φ true, denoted by $[i\ \mathbf{infl}\ j]\varphi$, if and only if i sees to it that if agent i's current choice is rational then i is going to deliberately see to it that φ. That is, for all $i, j \in Agt$ such that $i \neq j$, we define:

$$[i\ \mathbf{infl}\ j]\varphi \stackrel{\text{def}}{=} [i\ \mathbf{stit}]\mathbf{X}[j\ \mathbf{rdstit}]\varphi.$$

In the previous definition of influence, we use the operator $[j\ \mathbf{rdstit}]$ rather than $[j\ \mathbf{rstit}]$ since, as emphasized in Section 2, we assume that influence consists in determining the (voluntary)

action of a certain agent and the deliberative STIT is more appropriate than the Chellas STIT to capture the concept of action. Moreover, we use $[i\ \mathbf{stit}]$ rather than $[i\ \mathbf{dstit}]$ because we assume that the minimal condition for agent i to influence agent j to perform a certain action is that i's choice is a *sufficient* condition for j's action to occur. Finally, the reason why the operator $[i\ \mathbf{stit}]$ is followed by the temporal operator $\mathbf{X}$ is that influence requires that the influencer's choice precedes the influencee's action. On the contrary, we do not require that $[j\ \mathbf{rdstit}]$ is followed by $\mathbf{X}$ since, as observed in Section 2, in STIT the concept of action is simply captured by the deliberative STIT operator which does not necessarily need to be followed by temporal modalities.

Note that, differently from formula $[i\ \mathbf{stit}]\mathbf{X}[j\ \mathbf{dstit}]\varphi$, formula $[i\ \mathbf{infl}\ j]\varphi$ is satisfiable. In order to illustrate this, let us go back to the examples of Figures 3 and 4 in Section 3. Since agent 2 prefers eating bananas to pears, her only rational choice at moment m_2 is $\{w_7\}$. That is, we assume that $\{w_7\} \in RC_2$ while $\{w_8\} \notin RC_2$. From this assumption, it follows that formula $[1\ \mathbf{infl}\ 2]ba$ is true at world w_1. Indeed, $[1\ \mathbf{stit}]\mathbf{X}[2\ \mathbf{rdstit}]ba$ is clearly true at w_1 because for all $v \in \sim_1 \circ \twoheadrightarrow (w_1) = \{w_7, w_8\}$ we have: (i) if $\sim_2 (v) \in RC_2$ then $M, u \models ba$ for all $u \in \sim_2 (v)$, and (ii) $M, u \models \neg ba$ for some $u \in \equiv (v)$. Similarly, Since agent 2 prefers buying an electric car to a diesel car in the presence of a monetary incentive, her only rational choice at moment m_2 is $\{w_5\}$. That is, we assume that $\{w_5\} \in RC_2$ while $\{w_6\} \notin RC_2$. From this assumption, it follows that formula $[1\ \mathbf{infl}\ 2]el$ is true at world w_1. Indeed, $[1\ \mathbf{stit}]\mathbf{X}[2\ \mathbf{rdstit}]el$ is clearly true at w_1 because for all $v \in \sim_1 \circ \twoheadrightarrow (w_1) = \{w_5, w_6\}$ we have: (i) if $\sim_2 (v) \in RC_2$ then $M, u \models el$ for all $u \in \sim_2 (v)$, and (ii) $M, u \models \neg el$ for some $u \in \equiv (v)$.

The following proposition highlights some interesting properties of the influence operator $[i\ \mathbf{infl}\ j]$.

Proposition 4 *For all $i, j, k \in Agt$ such that $i \neq j$, $i \neq k$ and $j \neq k$ we have:*

$$\models_{\text{DR-STIT}} [i\ \mathbf{infl}\ j]\varphi \rightarrow [i\ \mathbf{stit}]\mathbf{X}\Diamond\neg\varphi \tag{1}$$

$$\models_{\text{DR-STIT}} ([i\ \mathbf{infl}\ j]\varphi \wedge [i\ \mathbf{infl}\ j]\psi) \rightarrow [i\ \mathbf{infl}\ j](\varphi \wedge \psi) \tag{2}$$

$$\models_{\text{DR-STIT}} \neg[i\ \mathbf{infl}\ j]\top \tag{3}$$

$$\models_{\text{DR-STIT}} \neg[i\ \mathbf{infl}\ j]\bot \tag{4}$$

$$\models_{\text{DR-STIT}} \neg([i\ \mathbf{infl}\ j]\varphi \wedge [i\ \mathbf{infl}\ k]\neg\varphi) \tag{5}$$

$$\models_{\text{DR-STIT}} [i\ \mathbf{infl}\ j][j\ \mathbf{infl}\ k]\varphi \leftrightarrow [i\ \mathbf{infl}\ j]\mathbf{X}[k\ \mathbf{rdstit}]\varphi \tag{6}$$

Validity 1 captures the idea that agent i influences agent j to perform a certain voluntary action only if the result of j's action *is not necessary*. Indeed, as emphasized above, action is incompatible with necessity. Validity (2) characterizes the behavior of the operator $[i\ \mathbf{infl}\ j]$ with conjunction. Note that its converse (i.e., $[i\ \mathbf{infl}\ j](\varphi \wedge \psi) \rightarrow ([i\ \mathbf{infl}\ j]\varphi \wedge [i\ \mathbf{infl}\ j]\psi)$) is not DR-STIT valid. Indeed, the fact that, after i's action, j has a choice available which could possibly make $\varphi \wedge \psi$ false, does not imply that after i's action, j has a choice available which could possibly make φ false and a choice available which could possibly make ψ false. Validities (3) and (4) just say that an agent cannot influence another agent to bring about tautologies or contradictions. These two validities follow from Axiom **(OneRat)**. Indeed, Axiom **(OneRat)** guarantees that, after i's action, j has at least one rational choice. Since, it is never the case that an agent deliberately brings about tautologies or contradictions (i.e., $\neg[i\ \mathbf{dstit}]\top$ and $\neg[i\ \mathbf{dstit}]\bot$ are both valid formulae), it follows that i cannot influence j to bring about

tautologies or contradictions. According to validity (5), an agent cannot influence two different agents to bring about conflicting results. Finally, validity (6) provides a characterization of chain of influences: the fact 'i influences j to influence k to make φ true' just means that i influences j to ensure that in the next state if k's current choice is rational, then k deliberately sees to it that φ.

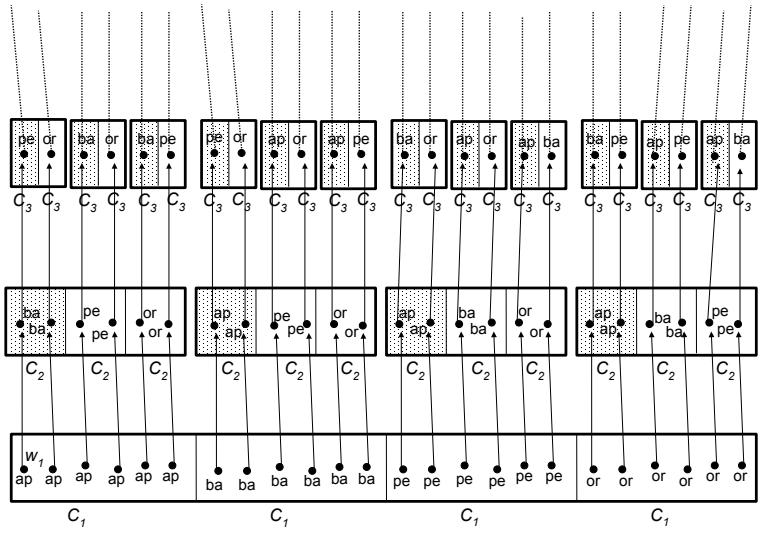


Figure 6: Example of chain of influence of length 2

The concept of chain of influence is clearly illustrated in Figure 6 which is nothing but the three-agent variant of the fruit example given in Section 3 (Figure 3). There are now four fruits on the table, an apple (ap), a banana (ba), a pear (pe), and an orange (or). Agent 1 is the first one to choose by picking one of the four fruits. Agent 2 is the second one to choose by picking one of the remaining three fruits. Agent 3 is the last one to choose by picking one of the two remaining alternatives. Let us assume that agent 2 and agent 3 have the same preferences over the fruits that remain constant along the tree structure, namely, they prefer apples to bananas, bananas to pears, and pears to oranges. Let us also assume that agent 2 and agent 3 are rational, in the sense that they choose the option they prefer among the available alternatives. As in Figure 3, rational choices of agent 2 and agent 3 are depicted in grey. It is easy to check that the formula $[1\ \mathbf{infl}\ 2][2\ \mathbf{infl}\ 3]pe$ is true at world w_1, at the bottom left corner of the figure. Indeed, by picking the apple at w_1, agent 1 induces agent 2 to pick the banana at the second stage of the game who, in turn, induces agent 3 to pick the pear at the third stage of the game. In other words, at world w_1, a chain of influence of length 2 occurs leading from agent 1's choice of picking the apple to agent 3's choice of picking the pear via agent 2's choice of picking the banana. As emphasized in Section 4.2, the fact that, in the DR-STIT semantics, histories are infinite allow us to formally characterize chains of influence of any arbitrary length. Consequently, the example of Figure 6 can be generalized to any arbitrary set $\{i_1, \ldots, i_k\}$ of agents and chain of influence of length $k-1$. In this case, we have (i) to start with a set $\{f_1, \ldots, f_{k+1}\}$ of fruits, (ii) to assume that the agents choose sequentially

starting with agent i_1's choice over the complete set of alternatives and finishing with agent i_k's choice over two remaining alternatives, and (iii) to assume that every agent in $\{i_2, \ldots, i_k\}$ has the following preference over the fruits $f_1 > f_2 > \ldots > f_{k+1}$. In the initial moment of the model corresponding to the example, where agent i_1 chooses to pick fruit f_1, the following formula would hold $[i_1 \ \mathbf{infl} \ i_2] \ldots [i_{k-1} \ \mathbf{infl} \ i_k] f_k$. This means that, by picking fruit f_1, agent i_1 induces agent i_2 to pick fruit f_2 who, in turn, induces agent i_3 to pick fruit f_3, and so on, until agent i_k is induced by agent i_{k-1} to pick fruit f_k.

As emphasized above, the operator $[i \ \mathbf{infl} \ j]$ captures the minimal condition for agent i to influence agent j to perform a certain action, namely the fact that i's choice is sufficient for j's action to occur. A stronger notion of influence also requires that j's action would have not occurred had i made a different choice. This stronger notion of influence is captured by the following abbreviation in which the Chellas STIT operator $[i \ \mathbf{stit}]$ is replaced by the deliberative STIT operator $[i \ \mathbf{dstit}]$, with $i \neq j$:

$$[i \ \mathbf{strinfl} \ j]\varphi \stackrel{\text{def}}{=} [i \ \mathbf{dstit}]\mathbf{X}[j \ \mathbf{rdstit}]\varphi.$$

where $[i \ \mathbf{strinfl} \ j]\varphi$ has to be read "agent i *strongly* influences agent j to make φ true". As the following validity highlights, the previous definition guarantees that agent i does not *strongly* influence agent j to make φ true, when j's action of bringing about φ is inevitable, in the sense that, *necessarily*, if in the next state j makes a rational choice then j will deliberately see to it that φ:

$$\models_{\text{DR-STIT}} [i \ \mathbf{strinfl} \ j]\varphi \rightarrow \neg\Box\mathbf{X}[j \ \mathbf{rdstit}]\varphi$$

It is worth noting that the six validities in the preceding Proposition 4 are preserved by replacing the influence operator $[i \ \mathbf{infl} \ j]$ with the strong influence operator $[i \ \mathbf{strinfl} \ j]$.

Before concluding this section, let us point that our concept and corresponding operator of influence only entails the performance of the influencee's action if the influencee chooses rationally, that is to say, if the influencee chooses what she prefers. In other words, *successful* influence also requires that in the next world along the actual history, the influencee chooses rationally, as specified by the following abbreviation, with $i \neq j$:

$$[i \ \mathbf{succinfl} \ j]\varphi \stackrel{\text{def}}{=} [i \ \mathbf{infl} \ j]\varphi \wedge \mathbf{X}\langle j \ \mathbf{rstit}\rangle\top.$$

where $[i \ \mathbf{succinfl} \ j]\varphi$ has to be read "agent i *successfully* influences agent j to make φ true".[9] This operator of successful influence clearly implies that in the next world along the actual history the influencee performs the action for which she has been influenced. This is expressed by the following validity:

$$\models_{\text{DR-STIT}} [i \ \mathbf{succinfl} \ j]\varphi \rightarrow \mathbf{X}[j \ \mathbf{dstit}]\varphi.$$

7 Related work

The concept of influence has been modeled by Ingmar Pörn [32, 33] whose logic of action builds upon [21]. However, in this weaker logic, contrary to STIT, it is not contradictory to

[9] It is also reasonable to define the following concept of *successful strong* influence:

$$[i \ \mathbf{succstrinfl} \ j]\varphi \stackrel{\text{def}}{=} [i \ \mathbf{strinfl} \ j]\varphi \wedge \mathbf{X}\langle j \ \mathbf{rstit}\rangle\top.$$

where $[i \ \mathbf{succstrinfl} \ j]\varphi$ has to be read "agent i *successfully strongly* influences agent j to make φ true".

affirm that an agent i brings it about that another agent j brings it about that φ (i.e., $\mathbf{E}_i\mathbf{E}_j\varphi$ is consistent). The same holds in the definition of the 'bringing it about' operator $\mathbf{E}_i$ proposed by [15], which was extended in [36] to distinctively address the production of outcomes by influencing others. As we observed above, it seems to us that STIT's inconsistency of i's seeing to it that j deliberately sees to it that φ ($[i\ \mathbf{stit}]\mathbf{X}[j\ \mathbf{dstit}]\varphi$) correctly reflects the so-called negative condition of agency, namely, the fact that an outcome can properly be attributed to an action only when the outcome might not have obtained, had this action not taken place (a different understanding of such a negative condition, however, is assumed by [15]). Moreover neither formalization of the 'bringing it about' operator $\mathbf{E}_i$ includes a way to deal with time, and aspect that is essential, we believe, for capturing how the influencer's action constrains the subsequent behavior of the influencee.

Our idea of the rational choice of an agent, while formally similar to the idea of an obligatory choice in Horty's semantics for deontic logic (see [18, Chapter 4]), has a completely different purpose, being complementary, rather than alternative to a deontic model. In particular, Horty's utilitarian semantics provides a foundation for obligations governing a community of agents, by assigning to each history a single social utility. This social utility determines individual obligations, under the assumption that individuals ought to choose histories having the highest social utility (this is the concept of 'ought' of utilitarian morality), or rather to make a choice that is not dominated by some other choice, according to the social utility of the histories it includes. We, on the contrary, through our notion of a rational choice only want to model how each influencee would act in the context resulting from the influencer's action, if she were to act rationally, where by 'rationally' we only mean, 'according to her individual preferences over the set of choices that are available to her'. Thus, we must allow in principle for as many preference orderings over possible choices as there are individuals, and distinguish the notion of rational action from the idea of a morally (or legally) obligatory action. In this way that we can cover both inducements to behave in a socially beneficial ways (e.g. through sanctions or incentives), and inducements to behave antisocially (e.g., through threats or bribes). By combining our logic of influence with a deontic logic, we can then distinguish deontically permissible and impermissible influence patterns, namely cases when the influencer or the influencee violate deontic constraints in exercising the influence or in conforming to it.

8 Model checking and applications

STIT logic has attracted an increasing attention, being used both in theoretical analyses of action, and in specifications for agent-based computer applications. We think that by enriching STIT with the capability to model influence we enable it to address a vast set of contexts and applications, i.e., whenever interactions between agents, humans and artificial ones, are at stake. In fact social agents achieve their goals by obtaining the cooperation of other agents, namely, by influencing them so that they act in the desired way. This motivates inducements, rewards and sanctions.

The first problem that our logic DR-STIT might help to solve is in the area of organizational design. One might want to verify whether a certain agent in a team can induce another agent to perform a given action. For example, one might want influencing power to be equally distributed between the agents in a team so that there is no concentration of influencing power in a single agent. By using our logic, we could check, within a team of agents, what agents have the power to influence what other agents to perform what actions. This would enable us

to determine whether an appropriate system of checks and balances is in place, so as to warrant the stability of the organization.

The second problem that our logic DR-STIT might help to solve is in the law area. Indeed, the notion of influence is crucial for the allocation of responsibilities according to normative systems. We need the notion of influence to be able to direct blame and sanctions not only against those who directly perform damaging acts, but also against those who have induced the authors to perform such acts. This is crucial in criminal law, where accomplices are responsible for criminal acts performed by others, exactly on the basis of the fact that they influenced the author of the criminal act into performing that act (an act that might not have taken place, had the accomplice not facilitated, ordered or instigated its performance). Similarly, in tort law, secondary liability, or indirect infringement, arises when a party contributes to directly infringing acts carried out by another party. We believe that in the regulation of societies of intelligent artificial agents, able to engage in cooperation, these responsibilities must also be introduced, to effectively target cooperation aimed at socially obnoxious activities.

These two problems can be addressed as instances of the model checking problem for our logic DR-STIT. For instance, let us consider the second problem of verifying whether an agent 1 has induced another agent 2 to perform a certain damaging act. Suppose agent 1 and agent 2 are driving into opposite directions in the middle of a city. Suddenly, agent 1 invades agent 2's track (in). If agent 2 goes straight (gs), he will crash against agent's 1's car (cc). In order to avoid a car accident, agent 2 can either swerve to the left (sl) or swerve to the right (sr). If he decides to swerve to the left, he will crash against the wall of a house and damage it (dw), whereas if he decides to swerve to the right, he will invade the sidewalk and will probably hurt someone walking on it (hs). Suppose that agent 2 has individual preferences which are aligned with social preferences, in the sense that he prefers those options that are considered to be the best for the society. Specifically, the best situation for him is the one in which the wall is not damaged, no passerby is hurt and no crash against agent 1's car occurs. Moreover, he prefers the situation in which the wall is damaged, no passerby is hurt and no crash against agent 1's occurs to the situation in which the wall is not damaged, no passerby is hurt and a crash against agent 1's car occurs. Finally, the worst situation for agent 2 is the one in which the wall is not damaged, a passerby is hurt and no crash against agent 1's occurs. (We assume that at most one of the atoms cc, dw and hs can be true at a given moment.) Thus, after agent 1 has invaded agent 2's track, agent 1's only rational choice is to swerve to the left with the consequence of damaging the wall. On the contrary, in the counterfactual situation in which agent 1 stays on her track and does not invade agent 2's track, agent 1's only rational choice is to go straight. All this story is represented in Figure 7.

Suppose now we want to verify whether, by invading agent 2's track, agent 1 successfully influences agent 2 to damage the wall of the house, in the sense of *successful influence* defined in Section 6. Evaluating whether agent 2's damaging action has been *successfully* determined by a previous action of agent 1 might be useful in order to decide whether responsibility for the damage should be also attributed to agent 1 as she influenced agent 2 to cause the damage. In formal terms, this corresponds to the task of verifying that formula $[1\ \mathbf{succinfl}\ 2]\mathbf{X}dw$ is true at world w_1 in the model depicted in Figure 7. This is nothing but an instance of the model checking problem for the logic DR-STIT.

Let us briefly comment how model checking can be concretely performed in our logic DR-STIT. As observed in Section 4.2, because of Constraint C1 and of the seriality of the temporal relation $\rightarrow$, Kripke STIT models with discrete time and rational choices of Definition

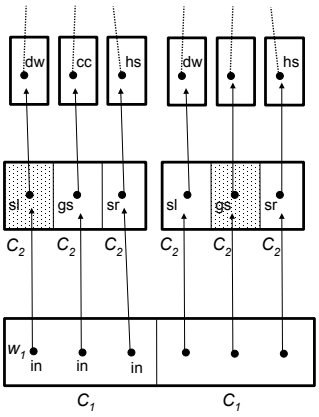


Figure 7: Example of model checking for DR-STIT

1 are infinite structures. As the input of a model checking problem is usually a finite model, we need to define a finite version of models for the logic DR-STIT. We call *finite* Kripke STIT models with discrete time and rational choices such a kind of models. Specifically, by a *finite* Kripke STIT model with discrete time and rational choices (or *finite* model, for short) we mean a tuple $M = (W, \equiv, \twoheadrightarrow, \twoheadleftarrow, \{\sim_i\}_{i\in Agt}, \sim_{Agt}, \{RC_i\}_{i\in Agt}, \mathcal{V})$ where W, $\equiv$, $\twoheadrightarrow$, $\twoheadleftarrow$, $\sim_i$, $\sim_{Agt}$, RC_i and $\mathcal{V}$ are as in Definition 1 except that: (1) W is finite, (2) $\twoheadrightarrow$ is not serial, and (3) $\mathcal{V}$ associates to every world $w \in W$ a *finite* set of atomic formulas in Atm. The truth conditions of DR-STIT formulas relative to *finite* Kripke STIT models with discrete time and rational choices are the same as the truth conditions relative to Kripke STIT models with discrete time and rational choices. The model checking problem for DR-STIT is the following decision problem: given a *finite* model M, a world w_1 in M (the origin) and a DR-STIT formula φ, is it the case that $M, w_1 \models \varphi$?

The model checking algorithm for DR-STIT is just an adaptation of the well-known modal checking algorithm for multimodal logics and CTL [12]. The idea of the algorithm is to label the worlds of the finite model step-by-step with sub-formulas of the formula φ to be checked, starting from the smallest ones, the atomic propositions appearing in φ. At each step, a formula should be added as a label to just those worlds of the model at which it is true. Specifically, at each step, the 'labelling' procedure goes as follows. Let φ' be a sub-formulas of the original formula φ to be checked at world w_1 of the finite model $M = (W, \equiv, \twoheadrightarrow, \twoheadleftarrow, \{\sim_i\}_{i\in Agt}, \sim_{Agt}$ $, \{RC_i\}_{i\in Agt}, \mathcal{V})$. Suppose that all sub-formulas of φ' have been checked. Then, for each world $w \in W$:

- if $\varphi' = p$ then label w with φ' iff $p \in \mathcal{V}(w)$;
- if $\varphi' = \neg\psi$ then label w with φ' iff w was not labeled with ψ;

- if $\varphi' = \psi_1 \wedge \psi_1$ then label w with φ' iff w was labeled both with ψ_1 and with ψ_2;
- if $\varphi' = \Box\psi$ then label w with φ' iff all worlds $v \in W$ such that $w \equiv v$ were labeled with ψ;
- if $\varphi' = \mathbf{X}\psi$ then label w with φ' iff all worlds $v \in W$ such that $w \twoheadrightarrow v$ were labeled with ψ;
- if $\varphi' = \mathbf{Y}\psi$ then label w with φ' iff all worlds $v \in W$ such that $w \twoheadleftarrow v$ were labeled with ψ;
- if $\varphi' = [i\ \mathbf{stit}]\psi$ then label w with φ' iff all worlds $v \in W$ such that $w \sim_i v$ were labeled with ψ;
- if $\varphi' = [Agt\ \mathbf{stit}]\psi$ then label w with φ' iff all worlds $v \in W$ such that $w \sim_{Agt} v$ were labeled with ψ;
- if $\varphi' = [i\ \mathbf{rstit}]\psi$ then label w with φ' iff if $w \in RC_i$ then all worlds $v \in W$ such that $w \sim_i v$ were labeled with ψ.

The labelling procedure terminates, as the size of the sub-formulas of the original formula φ to be checked increases and the formula has a finite size. Let $W' \subseteq W$ be the set of worlds at which formula φ is added as a label, at termination of the labelling procedure. If $w_1 \in W'$, then the model checking algorithm returns "yes", otherwise it returns "no". The overall complexity of this model checking algorithm for DR-STIT is linear in the number of the different sub-formulas of φ and in the size of the model M. It follows that the model checking problem for DR-STIT is PTIME-complete with respect to $size(\varphi) + size(M)$.

9 Conclusion

Let's take stock. We have started the paper by raising the challenge of modeling the concept of social influence in STIT theory. Then, we have proposed a variant of STIT with special operators describing the consequences of an agent's rational choice and shown that our logic offers a suitable framework for modeling this concept. On the technical side, we have provided a proof calculus for this logic.

Directions of future work are manifold. An important issue that has not been addressed in this paper is the relationship between the concept of rational (or preferred) choice, represented by the set RC_i in the definition of Kripke STIT model with discrete time and rational choices (Definition 1), and the concept of preference over the worlds. Indeed, the fact that the choice of an agent is considered to be rational (or preferred) depends on the fact that, by making this choice, the agent will maximize her preferences over the worlds (or outcomes). This is one of the fundamental aspect of classical decision theory. The logic DR-STIT, as it stands, has nothing to say about this relationship. In order to overcome this limitation, we plan to enrich Definition 1 with a preference relation over the worlds for each agent and to extend the language of DR-STIT by modal operators for preference such as the ones studied in the area of modal logic of preferences (see, e.g., [38]). This will allow us to formally characterize the preference relation over the worlds that is implicit in the two examples given in Section 3. As emphasized in Section 3, we also plan to extend our logic by modal operators for beliefs in

order to capture forms of influence via belief change or persuasion. The extension of DR-STIT by modal operators for preference and modal operators for belief will go into the direction of logics of mental attitudes [13, 28] and of combination of STIT logic with mentalistic concepts such as knowledge and intention [17, 8].

Another interesting direction of future research is an extension of our analysis of social influence to the *achievement* STIT operator of [2]. The interesting aspect of this operator is that it allows for a fine-grained characterization of the *counterfactual* dimension of causality in action. Specifically, the achievement STIT operator is a 'backward-looking' operator of agency. That is, in order to say that agent i is the cause of φ (in the achievement STIT sense), one must look in the past and check whether agent i had the possibility of making a different choice resulting in φ to be false now. We believe that the achievement STIT operator as defined by Belnap & Perloff (or, at least, an approximation of it) can be expressed in our logic DR-STIT, by combining in the appropriate way the Chellas STIT operator $[i\ \mathbf{stit}]$, the operator of historical necessity $\Box$, and the 'forward-looking' and 'backward-looking' temporal operators $\mathbf{X}$ and $\mathbf{Y}$.

On the technical side, we plan to look at the computational properties of our logic DR-STIT starting with decidability of the satisfiability problem and then moving to the analysis of the computational complexities of both model checking and the satisfiability problem.

References

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

[2] N. Belnap and M. Perloff. Seeing to it that: a canonical form for agentives. *Theoria*, 54:175–199, 1988.

[3] N. Belnap, M. Perloff, and M. Xu. *Facing the future: agents and choices in our indeterminist world*. Oxford University Press, New York, 2001.

[4] B. Berofsky. Ifs, cans, and free will: the issues. In R. Kane, editor, *The Oxford Handbook of Free Will*, pages 181–201. Oxford University Press, Oxford, 2002.

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

[6] J. Broersen. Deontic epistemic stit logic distinguishing modes of mens rea. *Journal of Applied Logic*, 9(2):137–152, 2011.

[7] J. Broersen. A complete STIT logic for knowledge and action, and some of its applications. In *Proceedings of the 6th International Workshop on Declarative Agent Languages and Technologies (DALT 2008)*, volume 5397 of *LNCS*, pages 47–59. Springer-Verlag, 2008.

[8] J. Broersen. Making a start with the Stit logic analysis of intentional action. *Journal of Philosophical Logic*, 40:399–420, 2011.

[9] J. Burgess. Logic and time. *Journal of Symbolic Logic*, 44:556–582, 1979.

[10] C. Castelfranchi. Modelling social action for AI agents. *Artificial Intelligence*, 103(1-2):157–182, 1998.

[11] B. J. Chellas. Time and modality in the logic of agency. *Studia Logica*, 51:485–517, 1992.

[12] E. M. Clarke and B. H. Schlingloff. Model checking. In A. Robinson, A. Voronkov, editors, *Handbook of automated reasoning*, pages 1635–1790 . Elsevier, 2001.

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

[14] W. Conradie, V. Goranko and D. Vakarelov. Algorithmic correspondence and completeness in modal logic I: The core algorithm SQEMA. *Logical Methods in Computer Science*, 2(1):1–26, 2006.

[15] D. Elgesem. The modal logic of agency. *Nordic Journal of Philosophical Logic*, 2:1–46, 1997.

[16] A. Herzig and E. Lorini. A dynamic logic of agency I: STIT, abilities and powers. *Journal of Logic, Language and Information*, 19(1):89–121, 2010.

[17] A. Herzig and E. Troquard. Knowing how to play: uniform choices in logics of agency. In *Proceedings of the 5th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2006)*, pages 209–216. ACM Press, 2006.

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

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

[20] A. Jones and M. Sergot. On the characterisation of law and computer systems: The normative systems perspective. In J. J. Ch. Meyer, R. J. Wieringa, editors, *Deontic Logic in Computer Science: Normative System Specification*, pages 275–307. John Wiley and Sons, UK, 1993.

[21] S. Kanger. Law and logic. *Theoria*, 38:105–132, 1972.

[22] G. W. Leibniz. *Theodicy: Essays on the Goodness of God, the Freedom of Man and the Origin of Evil*. Open Court, La Salle, Ill., [1719] 1985. Transl. E.M. Huggard.

[23] C. List and W. Rabinowicz. Two intuitions about free will: Alternative possibilities and endorsement. Technical report, London School of Economics, London, 2013.

[24] E. Lorini. Temporal STIT logic and its application to normative reasoning. *Journal of Applied Non-Classical Logics*, 23(4):372–399, 2013.

[25] E. Lorini and R. Ciuni. Comparing semantics for temporal STIT logic. *IRIT Technical Report*, n. IRIT/RT–2015–01–FR, 2015.

[26] E. Lorini and F. Schwarzentruber. A logic for reasoning about counterfactual emotions. *Artificial Intelligence*, 175(3-4):814–847, 2011.

[27] E. Lorini and F. Schwarzentruber. A modal logic of epistemic games. *Games*, 1(4):478–526, 2010.

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

[29] G. E. Moore. *Ethics: The Nature of Moral Philosophy*. Oxford University Press, Oxford, [1912] 2005.

[30] P. Panzarasa, N. Jennings, and T. J. Norman. Formalising collaborative decision making and practical reasoning in multi-agent systems. *Journal of Logic and Computation*, 12(1):55–117, 2002.

[31] J. Plaza. Logics of public communications. In *Proceedings of the Fourth International Symposium on Methodologies for Intelligent Systems (ISMIS)*, pages 201–216. ACM Press, 1989.

[32] I. Pörn. *The Logic of Power*. Blackwell, Oxford, 1970.

[33] I. Pörn. On the nature of social order. In J.E. Fenstad, I.T. Frolov, and R. Hilpinen, editors, *Logic, Metodology and Philosophy of Science. Vol. 8*, pages 553–67. North Holland, Amsterdam, 1989.

[34] A. Prior. *Past, Present, and Future*. Clarendon Press, Oxford, 1967.

[35] M. Reynolds. Axioms for branching time. *Journal of Logic and Computation*, 4:679–697, 2002.

[36] F. Santos, A. Jones, and J. Carmo. Action concepts for describing organised interaction. In *Proceedings of the Thirtieth Annual Hawaii International Conference on System Sciences*, pages 373–382. IEEE Computer Society, 1997.

[37] F. Schwarzentruber. Complexity results of STIT fragments. *Studia Logica*, 100(5):1001–1045, 2012.

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

[39] S. Wölf. Propositional Q-logic. *Journal of Philosophical Logic*, 31:387–414, 2002.

[40] M. Xu. Axioms for deliberative STIT. *Journal of Philosophical Logic*, 27:505–552, 1998.

[41] A. Zanardo. Branching-time logic with quantification over branches: The point of view of modal logic. *Journal of Symbolic Logic*, 61(1):143–166, 1996.

A Proof of Theorem 1

Proof.

Proving that the principles given in Figure 5 are sound with respect to the class of Kripke STIT models with discrete time and rational choices (KDRs for short) is just a routine task. The proof of completeness requires more work and is divided in four steps.

From KDRs to standard KDRs. The first step consists in providing a DR-STIT semantics in terms of *standard* KDRs (SKDRs for short), that is, tuples of the form $(W, \equiv, \rightarrow, \leftarrow, \{\sim_i\}_{i\in Agt}, \sim_{Agt}, \{\mathcal{RC}_i\}_{i\in Agt}, \mathcal{V})$ where $W, \equiv, \leftarrow, \rightarrow, \{\sim_i\}_{i\in Agt}, \sim_{Agt}$ and $\mathcal{V}$ are exactly as in Definition 1 and $\{\mathcal{RC}_i\}_{i\in Agt}$ is a family of binary relations on W that satisfy the following three constraints for all $i \in Agt$:

(C7) for all $w \in W$: if $\mathcal{RC}_i(w) \neq \emptyset$ then $\mathcal{RC}_i(w) = \sim_i(w)$;

(C8) for all $w, v \in W$: if $v \in \sim_i(w)$ and $\mathcal{RC}_i(w) \neq \emptyset$ then $\mathcal{RC}_i(v) \neq \emptyset$;

(C9) for all $w \in W$: there exists $v \in W$ such that $w \equiv v$ and $\mathcal{RC}_i(v) \neq \emptyset$.

Truth conditions of DR-STIT formulae relative to the class of SKDRs are exactly as the one in Section 4.2, except the truth condition for $[i\ \mathbf{rstit}]\varphi$ which is as follows:

$$M, w \models [i\ \mathbf{rstit}]\varphi \iff \forall v \in \mathcal{RC}_i(w) : M, v \models \varphi$$

We have that:

Lemma 1 *For every formula φ in $\mathcal{L}_{\text{DR-STIT}}(Atm, Agt)$, φ is satisfiable relative to the class of KDRs iff it is satisfiable relative to the class of SKDRs.*

Proof. ($\Rightarrow$) Let $M = (W, \equiv, \rightarrow, \leftarrow, \{\sim_i\}_{i\in Agt}, \sim_{Agt}, \{RC_i\}_{i\in Agt}, \mathcal{V})$ be a KDR and let $\underline{w}$ be a world in M such that $M, \underline{w} \models \varphi$. We define a corresponding structure $M' = (W, \equiv, \rightarrow, \leftarrow, \{\sim_i\}_{i\in Agt}, \sim_{Agt}, \{\mathcal{RC}_i\}_{i\in Agt}, \mathcal{V})$ such that for all $i \in Agt$ and for all $w, v \in W$, $w\mathcal{RC}_i v$ if and only if $w \sim_i v$ and $\sim_i(w) \in RC_i$. It is a routine task to verify that M' satisfies the previous Constraints C7, C8 and C9. By induction on the structure of φ, it is easy to check that $M', \underline{w} \models \varphi$. The only interesting case is $\varphi = [i\ \mathbf{rstit}]\psi$:

$$\begin{aligned}
M, \underline{w} \models [i\ \mathbf{rstit}]\psi &\iff \text{IF } \sim_i(\underline{w}) \in RC_i \text{ THEN} \\
&\quad \forall v \in \sim_i(\underline{w}) : M, v \models \psi \\
&\iff \forall v \in W : \text{IF } \sim_i(\underline{w}) \in RC_i \text{ AND } w \sim_i v \text{ THEN} \\
&\quad M, v \models \psi \\
&\iff \forall v \in \mathcal{RC}_i(w) : M, v \models \varphi \\
&\iff M', w \models [i\ \mathbf{rstit}]\psi
\end{aligned}$$

($\Leftarrow$) Let $M = (W, \equiv, \rightarrow, \leftarrow, \{\sim_i\}_{i\in Agt}, \sim_{Agt}, \{\mathcal{RC}_i\}_{i\in Agt}, \mathcal{V})$ be a SKDR and let w be a world in M such that $M, w \models \varphi$. We define a corresponding structure $M' = (W, \equiv, \rightarrow, \leftarrow, \{\sim_i\}_{i\in Agt}, \sim_{Agt}, \{RC_i\}_{i\in Agt}, \mathcal{V})$ such that for all $i \in Agt$ and for all $X \subseteq W$, $X \in RC_i$ if and only if $X = \sim_i(w)$ for some $w \in W$ and $\sim_i(w) = \mathcal{RC}_i(w)$. It is a routine task to check that RC_i is a subset of the partition of W induced by the equivalence relation $\mathcal{RC}_i$ which satisfies Constraint C4 in Definition 1. Moreover, by induction on the structure of φ, it is easy to check that $M', \underline{w} \models \varphi$. The only interesting case is $\varphi = [i\ \mathbf{rstit}]\psi$:

$$\begin{aligned}
M, \underline{w} \models [i\ \mathbf{rstit}]\psi &\iff \forall v \in \mathcal{RC}_i(\underline{w}) : M, v \models \psi \\
&\iff \text{IF } \sim_i(\underline{w}) = \mathcal{RC}_i(\underline{w}) \text{ THEN} \\
&\quad \forall v \in \sim_i(\underline{w}) : M, v \models \psi \\
&\iff \text{IF } \sim_i(\underline{w}) \in RC_i \text{ THEN} \\
&\quad \forall v \in \sim_i(\underline{w}) : M, v \models \psi \\
&\iff M', \underline{w} \models [i\ \mathbf{rstit}]\psi
\end{aligned}$$

From SKDRs to SKDRs with possible cycles. The second step consists in defining a DR-STIT semantics in terms of standard Kripke STIT models with discrete time, rational choices *and possible cycles* (SKDRCs). The latter are like SKDRs except that they do not necessarily satisfy the Constraint C1 in Definition 1 about the special kind of irreflexivity of the temporal relation $\mathcal{F}$ between moments. We have that:

Lemma 2 *For every formula φ in $\mathcal{L}_{DR\text{-}STIT}(Atm, Agt)$, φ is satisfiable relative to the class of SKDRs iff it is satisfiable relative to the class of SKDRCs.*

Proof. The lemma is provable by showing that, for every SKDRC M, we can build a SKDR M' and define a bounded morphism from M' to M [5, Def. 2.12]. In other words, we show that the Constraint C1 is not modally definable in the logic DR-STIT.

Let us denote sequences of worlds by symbols $\sigma, \sigma', \ldots$ Given a sequence σ and a world w, we write $\sigma.w$ to denote the sequence obtained by adding w at the end of sequence σ. We consider two types of sequences, namely finite sequences of the form $\langle w_1, \ldots, w_n\rangle$ and left-infinite sequences of the form $\langle \ldots, w_{-2}, w_{-1}\rangle$. Let $FSeq$ be the set of all finite sequences and let $LISeq$ be the set of all left-infinite sequences. Given a finite sequence $\sigma \in FSeq$, let $length(\sigma)$ denote the length of σ. Moreover, given a finite sequence $\sigma \in FSeq$ of length n, let $\sigma[k]$ with $1 \leq k \leq n$ denote the element in the k-position of the sequence. Similarly, given a left-infinite sequence $\sigma \in LISeq$, let $\sigma[k]$ with $k \leq -1$ denote the element in the k-position of the sequence. Given a sequence $\sigma \in FSeq \cup LISeq$, let $\sigma[last]$ denote the element in the last position of the sequence. That is, if $\sigma = \langle w_1, \ldots, w_n\rangle \in FSeq$ then $\sigma[last] = w_n$, and if $\sigma = \langle \ldots, w_{-2}, w_{-1}\rangle \in LISeq$ then $\sigma[last] = w_{-1}$.

Let

$$\begin{aligned} Path \;=\; & \{\sigma \in FSeq | \sigma[k] \twoheadrightarrow \sigma[k+1] \text{ for } 1 \leq k < length(\sigma)\} \cup \\ & \{\sigma \in LISeq | \sigma[k] \twoheadrightarrow \sigma[k+1] \text{ for } k < -1\} \end{aligned}$$

be the set of all paths.

Finally, let $MPPath$ be the set of all paths that are maximal in the past, that is, $\sigma \in MPPath$ iff $\sigma \in Path \cap LISeq$ or $\sigma \in Path \cap FSeq$ and there is no v such that $v \twoheadleftarrow \sigma[1]$.

Let $M = (W, \equiv, \twoheadrightarrow, \twoheadleftarrow, \{\sim_i\}_{i\in Agt}, \sim_{Agt}, \{\mathcal{RC}_i\}_{i\in Agt}, \mathcal{V})$ be a SKDRC and let $\underline{w}$ be a world in M such that $M, \underline{w} \models \varphi$.

We transform M into a new model $M' = (W', \equiv', \twoheadrightarrow', \twoheadleftarrow', \{\sim'_i\}_{i\in Agt}, \sim'_{Agt},$

$\{\mathcal{RC}'_i\}_{i\in Agt}, \mathcal{V}')$ where:

$$
\begin{array}{rcl}
W' & = & MPPath \\
\equiv' & = & \{(\sigma,\sigma')|\sigma,\sigma' \in LISeq, \sigma[last] \equiv \sigma'[last] \text{ and } \sigma[k] \sim_{Agt} \sigma'[k] \\
 & & \text{for } k < -1\} \cup \\
 & & \{(\sigma,\sigma')|\sigma,\sigma' \in FSeq, length(\sigma) = length(\sigma'), \sigma[last] \equiv \sigma'[last] \\
 & & \text{and } \sigma[k] \sim_{Agt} \sigma'[k] \text{ for } 1 \leq k < length(\sigma)\} \\
\rightarrow' & = & \{(\sigma,\sigma')|\sigma' = \sigma.w \text{ for some } w \in W\} \\
\leftarrow' & \text{is} & \text{the inverse relation of } \rightarrow' \\
\sim'_i & = & \{(\sigma,\sigma')|\sigma,\sigma' \in LISeq, \sigma[last] \sim_i \sigma'[last] \text{ and } \sigma[k] \sim_{Agt} \sigma'[k] \\
 & & \text{for } k < -1\} \cup \\
 & & \{(\sigma,\sigma')|\sigma,\sigma' \in FSeq, length(\sigma) = length(\sigma'), \sigma[last] \sim_i \sigma'[last] \\
 & & \text{and } \sigma[k] \sim_{Agt} \sigma'[k] \text{ for } 1 \leq k < length(\sigma)\} \\
\sim'_{Agt} & = & \{(\sigma,\sigma')|\sigma,\sigma' \in LISeq \text{ and } \sigma[k] \sim_{Agt} \sigma'[k] \\
 & & \text{for } k \leq -1\} \cup \\
 & & \{(\sigma,\sigma')|\sigma,\sigma' \in FSeq, length(\sigma) = length(\sigma') \text{ and} \\
 & & \sigma[k] \sim_{Agt} \sigma'[k] \text{ for } 1 \leq k \leq length(\sigma)\} \\
\mathcal{RC}'_i & = & \{(\sigma,\sigma')|(\sigma,\sigma') \in \sim'_i \text{ and } \sigma[last]\mathcal{RC}_i\sigma'[last]\} \\
p \in \mathcal{V}'(\sigma) & \text{iff} & p \in \mathcal{V}(\sigma[last])
\end{array}
$$

We can prove that the mapping $f : \sigma \mapsto \sigma[last]$ defines a bounded morphism from M' to M. It follows from the definition of $\equiv', \rightarrow', \leftarrow', \sim'_i, \sim'_{Agt}$ and $\mathcal{RC}'_i$ that $\sigma \equiv' \sigma'$ implies $\sigma[last] \equiv \sigma'[last]$, $\sigma \rightarrow' \sigma'$ implies $\sigma[last] \rightarrow \sigma'[last]$, $\sigma \leftarrow' \sigma'$ implies $\sigma[last] \leftarrow \sigma'[last]$, $\sigma \sim'_i \sigma'$ implies $\sigma[last] \sim_i \sigma'[last]$, $\sigma \sim'_{Agt} \sigma'$ implies $\sigma[last] \sim_{Agt} \sigma'[last]$ and $\sigma\mathcal{RC}'_i\sigma'$ implies $\sigma[last]\mathcal{RC}_i\sigma'[last]$. The other way around, it follows from the definition of $\rightarrow'$ that $f(\sigma) \rightarrow w$ implies that there is σ' (viz. $\sigma' = \sigma.w$) such that $\sigma \rightarrow' \sigma'$ and $f(\sigma') = w$. It remains to prove that $f(\sigma) \equiv w$ implies that there is σ' such that $\sigma \equiv' \sigma'$ and $f(\sigma') = w$. This follows from Constraint C6 in Definition 1. Indeed, we have two possible cases.
CASE 1: $\sigma \in Path \cap LISeq$. By Constraint C6, it is routine to verify that if $\sigma[last] = v$ and $v \equiv w$ then there is $\sigma' \in LISeq$ such that $\sigma'[last] = w$ and $\sigma[k] \sim'_{Agt} \sigma'[k]$ for all $k < -1$.
CASE 2: $\sigma \in Path \cap FSeq$. By Constraint C6, it is routine to verify that if $\sigma[last] = v$ and $v \equiv w$ then there is $\sigma' \in LISeq$ such that $\sigma'[last] = w$, $length(\sigma) = length(\sigma')$ and $\sigma[k] \sim'_{Agt} \sigma'[k]$ for all $1 \leq k < length(\sigma)$.

In a similar way we can prove that:

- $f(\sigma) \sim_i w$ implies that there is σ' such that $\sigma \sim'_i \sigma'$ and $f(\sigma') = w$,
- $f(\sigma) \sim_{Agt} w$ implies that there is σ' such that $\sigma \sim'_{Agt} \sigma'$ and $f(\sigma') = w$,
- $f(\sigma)\mathcal{RC}_i w$ implies that there is σ' such that $\sigma\mathcal{RC}'_i\sigma'$ and $f(\sigma') = w$.

It is also a routine task to verify that M' is a SKDR.

As f is a bounded morphism it holds that $M, \underline{w} \models \varphi$ if and only if $M', \sigma \models \varphi$ when $\sigma[last] = \underline{w}$. Thus, $M', \sigma \models \varphi$ for all σ such that $\sigma[last] = \underline{w}$ since $M, \underline{w} \models \varphi$.

From SKDRCs to superadditive SKDRCs. The third step consists in introducing a DR-STIT semantics in terms of *superadditive* SKDRCs. The only difference between SKDRCs and *superadditive* SKDRCs is that in the latter the Constraint C5 in Definition 1 is replaced by the following weaker Constraint C5*:

(C5*) for all $w \in W$: $\sim_{Agt}(w) \subseteq \bigcap_{i \in Agt} \sim_i (w)$.

We prove that:

Lemma 3 *For every formula φ in $\mathcal{L}_{DR\text{-}STIT}(Atm, Agt)$, φ is satisfiable relative to the class of SKDRCs iff it is satisfiable relative to the class of* superadditive *SKDRCs.*

Proof. The lemma is provable by showing that, for every superadditive SKDRC M, we can build a SKDRC M' and define a bounded morphism from M' to M. In other words, we show that the direction $\sim_{Agt}(w) \supseteq \bigcap_{i \in Agt} \sim_i (w)$ of the Constraint C5 is not modally definable in the logic DR-STIT. A similar technique is used in [24, Lemma 1].

Let $M = (W, \equiv, \twoheadrightarrow, \twoheadleftarrow, \{\sim_i\}_{i \in Agt}, \sim_{Agt}, \{\mathcal{RC}_i\}_{i \in Agt}, \mathcal{V})$ be a superadditive SKDRC and let $\underline{w}$ be a world in M such that $M, \underline{w} \models \varphi$.

For every $i \in Agt$ let

$$\Delta_i = \{\sim_i (w) | w \in W\}$$

be the partition of W induced by the equivalence class $\sim_i$. Elements of Δ_i are called i's choices. Moreover, let the elements of $\prod_{i \in Agt} \Delta_i$ be denoted by $\delta, \delta', \ldots$ and let δ_i denote the element in the vector δ corresponding to the agent i. Let

$$\Delta = \{\delta \in \prod_{i \in Agt} \Delta_i | \bigcap_{i \in Agt} \delta_i \neq \emptyset\}$$

be the set of collective choices.[10] For notational convenience, we write δ^w to denote the collective choice in Δ that includes the world w. That is, δ^w is the collective choice in Δ such that $w \in \bigcap_{i \in Agt} \delta_i^w$.

For every $\delta \in \Delta$, let

$$\Gamma_\delta = \{\sim_{Agt}(w) | w \in \bigcap_{i \in Agt} \delta_i\}$$

be the set of $\sim_{Agt}$-equivalence classes that are included in $\bigcap_{i \in Agt} \delta_i$. For every $\delta \in \Delta$, let

$$k_\delta : [1, \ldots, card(\Gamma_\delta)] \longrightarrow \Gamma_\delta$$

be a *bijection* associating every integer between 1 and $card(\Gamma_\delta)$ to a unique element of Γ_δ. In the sequel we write Γ_δ^n to indicate the element $k_\delta(n)$ for every $1 \leq n \leq card(\Gamma_\delta)$.

For every $w \in W$ let

$$TC_w = \{\delta \in \Delta | \exists v \in \bigcap_{i \in Agt} \delta_i \text{ such that } v \in \mathcal{T}(w)\}$$

[10] Note that this collective choice is unique because every δ_i is an equivalence class.

be the set of collective choices that are temporally related with w, with $\mathcal{T}(w) = \mathcal{P}(w) \cup \{w\} \cup \mathcal{F}(w)$. Moreover, let

$$PC_w = \{\delta \in \Delta | \exists v \in \bigcap_{i \in Agt} \delta_i \text{ such that } v \in \mathcal{P}(w)\}$$

be the set of collective choices that are in the past of w. Finally, let

$$FC_w = \{\delta \in \Delta | \exists v \in \bigcap_{i \in Agt} \delta_i \text{ such that } v \in \mathcal{F}(w)\}$$

be the set of collective choices that are in the future of w.

Let Λ_w be the set of all total functions $f : TC_w \longrightarrow \mathbb{Z}^n$ that satisfy the following constraint $\mathbf{C}_f$:

$$(\mathbf{C}_f) \text{ if } f(\delta_w) = \overrightarrow{x} \text{ then } w \in \Gamma_{\delta_w}^{\Sigma_{x_i \in \{x_1, \ldots, x_n\}} x_i}$$

where $\mathbb{Z}$ is the set of integers and $\overrightarrow{x} = \langle x_1, \ldots, x_n \rangle$. We write $f_i(\delta)$ to denote the i-element in the vector $f(\delta)$, with $1 \leq i \leq n$.

By the Constraint C6 in Definition 1 and the previous Constraint $\mathbf{C}_f$, it is easy to prove that:

(A) if $f \in \Lambda_w$ and $w \equiv v$ then there exists $f' \in \Lambda_v$ such that, for all $\delta \in PC_w$, $f'(\delta) = f(\delta)$.

Moreover, by the previous Constraint $\mathbf{C}_f$, it is easy to prove that:

(B) if $f \in \Lambda_w$ and $v \in \mathcal{T}(w)$ then there exists $f' \in \Lambda_v$ such that, for all $\delta \in TC_w$, $f'(\delta) = f(\delta)$.

We are now able to transform the model M into a new model $M' = (W', \equiv', \rightarrow', \leftarrow', \{\sim'_i\}_{i \in Agt}, \sim'_{Agt}, \{\mathcal{RC}'_i\}_{i \in Agt}, \mathcal{V}')$ where:

$$\begin{array}{rcl}
W' & = & \{w_f | w \in W \text{ and } f \in \Lambda_w\} \\
w_f \equiv' v_{f'} & \text{iff} & w \equiv v \text{ and } f(\delta) = f'(\delta) \text{ for all } \delta \in PC_w \\
w_f \rightarrow' v_{f'} & \text{iff} & w \rightarrow v \text{ and } f(\delta) = f'(\delta) \text{ for all } \delta \in TC_w \\
\leftarrow' & \text{is} & \text{the inverse relation of } \rightarrow' \\
w_f \sim'_i v_{f'} & \text{iff} & \delta_i^w = \delta_i^v, f_i(\delta^w) = f'_i(\delta^v) \text{ and } f(\delta) = f'(\delta) \text{ for all } \delta \in PC_w \\
w_f \sim'_{Agt} v_{f'} & \text{iff} & \delta^w = \delta^v, f(\delta^w) = f'(\delta^v) \text{ and } f(\delta) = f'(\delta) \text{ for all } \delta \in PC_w \\
w_f \mathcal{RC}'_i v_{f'} & \text{iff} & w_f \sim'_i v_{f'} \text{ and } w \mathcal{RC}_i v \\
\mathcal{V}'(w_f) & = & \mathcal{V}(w)
\end{array}$$

It is a routine task to check that the mapping $g : w_f \mapsto w$ defines a *bounded morphism* from M' to M. Indeed, it follows from the definitions of $\equiv'$, $\rightarrow'$, $\leftarrow'$, $\sim'_i$, $\sim'_{Agt}$ and $\mathcal{RC}'_i$ that for all $w_f, v_{f'} \in W'$:

- $w_f \equiv' v_{f'}$ implies $w \equiv v$;
- $w_f \rightarrow' v_{f'}$ implies $w \rightarrow v$;

- $w_f \twoheadleftarrow' v_{f'}$ implies $w \twoheadleftarrow v$.
- $w_f \sim'_i v_{f'}$ implies $w \sim_i v$;
- $w_f \sim'_{Agt} v_{f'}$ implies $w \sim_{Agt} v$;
- $w_f \mathcal{RC}'_i v_{f'}$ implies $w \mathcal{RC}_i v$;

For instance, suppose that $w_f \sim'_i v_{f'}$. By definition of $\sim'_i$ the latter implies that $\delta_i^w = \delta_i^v$, $w \in \delta_i^w$ and $v \in \delta_i^v$. The latter implies that $w \sim_i v$.

Now, suppose that $w_f \sim'_{Agt} v_{f'}$. This implies that $\delta^w = \delta^v$ and $f(\delta^w) = f'(\delta^v)$. By the preceding Constraint C_f, it follows that $w, v \in \Gamma_{\delta^w}^{f_1(\delta^w)+\ldots+f_n(\delta^w)}$. Thus, $w \sim_{Agt} v$.

The other way around we have that:

- if $g(w_f) \equiv v$ then there is $v_{f'}$ such that $w_f \equiv' v_{f'}$;
- if $g(w_f) \twoheadrightarrow v$ then there is $v_{f'}$ such that $w_f \twoheadrightarrow' v_{f'}$;
- if $g(w_f) \twoheadleftarrow v$ then there is $v_{f'}$ such that $w_f \twoheadleftarrow' v_{f'}$;
- if $g(w_f) \sim_i v$ then there is $v_{f'}$ such that $w_f \sim'_i v_{f'}$;
- if $g(w_f) \sim_{Agt} v$ then there is $v_{f'}$ such that $w_f \sim'_{Agt} v_{f'}$;
- if $f(w_f) \mathcal{RC}_i v$ then there is $v_{f'}$ such that $w_f \mathcal{RC}'_i v_{f'}$.

The first item and the last item are trivial and we do not prove it here. Let us prove the second, third, fourth and fifth items.

Suppose that $w_f \in W'$ and $f(w_f) \twoheadrightarrow v$. This implies that $w \twoheadrightarrow v$. It follows that $w_f \twoheadrightarrow' v_{f'}$ where the function f' is such that, for all $\delta \in TC_w$, $f'(\delta) = f(\delta)$. This function f' is guaranteed to exist because of the observation (B) above. Thus, $w_f \twoheadrightarrow' v_{f'}$.

Suppose that $w_f \in W'$ and $f(w_f) \twoheadleftarrow v$. This implies that $w \twoheadleftarrow v$ which is equivalent to $v \twoheadrightarrow w$. It follows that $v_{f'} \twoheadrightarrow' w_f$ where the function f' is such that, for all $\delta \in TC_v$, $f(\delta) = f'(\delta)$. This function f' is guaranteed to exist because of the observation (B) above. Thus, $w_f \twoheadleftarrow' v_{f'}$.

Suppose that $w_f \in W'$ and $f(w_f) \sim_i v$. This implies that $w \sim_i v$. Therefore, $\sim_i(w) = \sim_i (v)$ as $\sim_i$ is an equivalence relation. Hence, $\delta_i^v = \delta_i^w$. It follows that $w_f \sim'_i v_{f'}$ where the function f' is defined as follows: (i) $f'_i(\delta^v) = f_i(\delta^w)$ and all $f'_j(\delta^v)$ with $j \neq i$ are such that $v \in \Gamma_{\delta^v}^{f'_1(\delta^v)+\ldots+f'_n(\delta^v)}$, (ii) for all $\delta \in PC_v$, $f'(\delta) = f(\delta)$, (iii) for all $\delta \in FC_v$, $f'(\delta)$ is arbitrarily defined. This function f' is guaranteed to exist because of the observation (A) above and the fact that $\sim_i \subseteq \equiv$.

Finally, suppose that $w_f \in W'$ and $f(w_f) \sim_{Agt} v$. This implies that $w \sim_{Agt} v$. Therefore, $\sim_{Agt}(w) = \sim_{Agt}(v)$ as $\sim_{Agt}$ is an equivalence relation. Hence, $\delta^v = \delta^w$. It follows that $w_f \sim'_{Agt} v_{f'}$ where the function f' is defined as follows: (i) $f'(\delta^v) = f(\delta^w)$, (ii) for all $\delta \in PC_v$, $f'(\delta) = f(\delta)$, (iii) for all $\delta \in FC_v$, $f'(\delta)$ is arbitrarily defined. This function f' is guaranteed to exist because of the observation (A) above and the fact that $\sim_{Agt} \subseteq \equiv$.

It is also a routine task to verify that M' is a SKDRC.

As g is a bounded morphism it holds that $M, \underline{w} \models \varphi$ if and only if $M', \underline{w}_f \models \varphi$. Thus, $M', \underline{w}_f \models \varphi$ for all $\underline{w}_f \in W'$ since $M, \underline{w} \models \varphi$.

Completeness wrt superadditive SKDRCs. The fourth step consists in proving the following lemma:

Lemma 4 *The set of DR-STIT formulae that are valid in the class of superadditive SKDRCs is completely axiomatized by the principles given in Figure 5.*

Proof. it is a routine task to check that all principles in Figure 5 correspond one-to-one to their semantic counterparts on the class of superadditive SKDRCs. This can for instance easily be found using the algorithm SQEMA [14]. In particular, **S5**($\Box$), **S5**($[i\ \mathbf{stit}]$) and **S5**($[Agt\ \mathbf{stit}]$) correspond to the fact that $\equiv$, $\sim_i$ and $\sim_{Agt}$ are equivalence relations, respectively. **KD(X)** corresponds to the fact that $\rightarrow\!\!\triangleright$ is a serial relation, while **(Alt$_{\mathbf{X}}$)** to the fact that $\rightarrow\!\!\triangleright$ is deterministic. **K(Y)** together with **(Conv$_{\mathbf{X},\mathbf{Y}}$)** and **(Conv$_{\mathbf{Y},\mathbf{X}}$)** correspond to the fact that $\triangleleft\!\!\leftarrow$ is the inverse relation of $\rightarrow\!\!\triangleright$ and **(Alt$_{\mathbf{Y}}$)** to the fact that $\triangleleft\!\!\leftarrow$ is deterministic. Finally, **(Rel$_{\Box,[i\ \mathbf{stit}]}$)**, **(AIA)**, **(Rel$_{[i\ \mathbf{rstit}],[i\ \mathbf{stit}]}$)**, **(RatCh)**, **(OneRat)**, **(Rel$_{[i\ \mathbf{stit}],[Agt\ \mathbf{stit}]}$)** and **(NCUH)**, correspond to the Constraints C2, C3, C7, C8, C9, C5* and C6, respectively.

Moreover, it is routine, too, to check that all principles given in Figure 5 are in the so-called Sahlqvist class. This means that they are complete with respect to the defined model classes, cf. [5, Th. 2.42].

Theorem 1 is a consequence of Lemma 1, Lemma 2, Lemma 3 and Lemma 4.

Artificial Intelligence 175 (2011) 814–847

Contents lists available at ScienceDirect

Artificial Intelligence

www.elsevier.com/locate/artint

A logic for reasoning about counterfactual emotions ☆

Emiliano Lorini *, François Schwarzentruber

Institut de Recherche en Informatique de Toulouse (IRIT), 118 route de Narbonne, 31062 Toulouse Cedex, France

ARTICLE INFO

Article history:
Received 20 December 2009
Received in revised form 22 November 2010
Accepted 22 November 2010
Available online 2 December 2010

Keywords:
Modal logic
Emotions
STIT

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 (Belnap et al. (2001) [9], Horty (2001) [30], Horty and Belnap (1995) [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 *df*STIT our STIT fragment. After having extended *df*STIT 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 present an application of our formalization of counterfactual emotions to a concrete example.

1. Introduction

A major objective of AI is to develop interactive cognitive systems which 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, e.g., [39,40,58,20]). 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.

☆ 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.

* Corresponding author. Tel.: +33 0561556447; fax: +33 561556258.
E-mail addresses: lorini@irit.fr (E. Lorini), schwarze@irit.fr (F. Schwarzentruber).

0004-3702/$ – see front matter
doi:10.1016/j.artint.2010.11.022

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, e.g., [12]). 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 *df*STIT 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 *df*STIT 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 for the formalization of counterfactual emotions given in Section 6.

In Section 5, we extend the STIT fragment *df*STIT 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 *df*STIT. 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.

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

[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.

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, p. 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. 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, p. 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, 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 al. introduce constructions of the form $[i\ stit : \varphi]$ 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

[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.

STIT with groups has been proposed by Horty in [30]: it deals with constructions of the form $[J\ stit:\varphi]$ to be read "group J sees to it that φ". For notational convenience, we write here $[J]\varphi$ instead of $[J\ stit:\varphi]$. Unfortunately, in [29] Horty's STIT logic 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 *df*STIT which is sufficiently expressive to formalize counterfactual emotions. First, in Section 3.1, we recall the syntax of STIT and define the syntactic fragment *df*STIT. In Section 3.2, we recall definition of models of the logic STIT. Then, in Section 3.3, we recall the logic NCL (Normal Coalition Logic) [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 *df*STIT and to give a complete axiomatization of *df*STIT (Section 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 al.'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:

$$\chi ::= \bot \mid p \mid \chi\wedge\chi \mid \neg\chi \quad \text{(propositional formulas)}$$

$$\psi ::= [J]\chi \mid \psi\wedge\psi \quad \text{(see-to-it formulas)}$$

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

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 *df*STIT-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 $J \in 2^{AGT}$, we define $R_J(w) = \{v \in W \mid wR_Jv\}$.

$R_J(w)$ 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 $w_2, \ldots$, 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:

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

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:

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

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 product 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 with agents and groups 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]. As CL, NCL is 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 a 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 non-empty 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 for the set of all agents AGT.

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:

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

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) $[J_1]\varphi \vee [J_2]\varphi \to [J_1 \cup J_2]\varphi$
Elim($\emptyset$) $\langle\emptyset\rangle\varphi \to \langle J\rangle\langle AGT \setminus J\rangle\varphi$
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* $\mathcal{L}_{\mathsf{STIT}}$.

- *If* $card(AGT) \leqslant 2$: φ *is STIT-satisfiable iff* φ *is NCL-satisfiable*;
- *If* $card(AGT) \geqslant 3$: *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 Fig. 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 Appendix A 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*:

[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 subsets of W is $W \times W$ by convention.

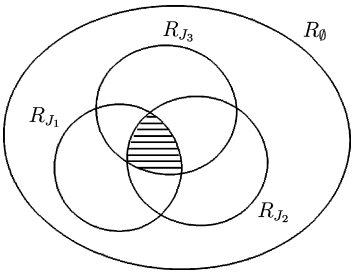


Fig. 1. Independence of agents in NCL.

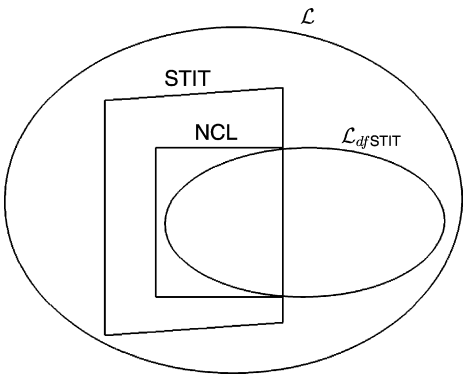


Fig. 2. Overview of the languages $\mathcal{L}$ and $\mathcal{L}_{df\mathsf{STIT}}$ and of the logics STIT and NCL.

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

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 $\varphi \in \mathcal{L}_{df\mathsf{STIT}}$. 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.*

Fig. 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 Theorem 2.

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

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

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

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 $\mathcal{L}_{df\mathsf{STIT}}$.

3.5. Discussion

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

The first reason is practical as the complexity of *df*STIT is lower than the complexity of NCL: the satisfiability problem for NCL is NEXPTIME-complete [5] while it is NP-complete for *df*STIT (Corollary 1). Moreover, as we will show in Section 5 (Theorem 3), the complexity of *df*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*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 nowadays 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 *df*STIT 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 for 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 $\mathtt{CHP}_J\chi$ this formal representation, which is defined as follows:

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

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 $\mathtt{CHP}_i\chi$ instead of $\mathtt{CHP}_{\{i\}}\chi$ which means "agent i could have prevented χ to be true".

The following is the semantic counterpart of the operator $\mathtt{CHP}_J$. We have that $\mathcal{M}, w \models \mathtt{CHP}_J\chi$ 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.

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

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

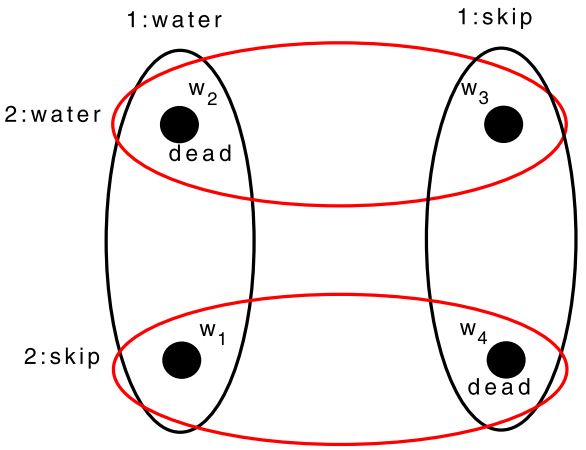


Fig. 3. The four worlds w_1, w_2, w_3 and w_4 are in the equivalence class determined by $R_\emptyset$. Vertical circles represent the actions that agent 1 can choose, whereas horizontal circles represent the actions that agent 2 can choose. For example, w_1 is the world that results from agent 1 choosing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

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 $\mathtt{CHP}_1\,dead$ and $\mathtt{CHP}_2\,dead$ 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 $\mathtt{CHP}_J$. For every J and for every J_1, J_2 such that $J_1 \subseteq J_2$:

$$\models_{\mathsf{STIT}} \mathtt{CHP}_{J_1}(\chi_1 \vee \chi_2) \rightarrow (\mathtt{CHP}_{J_1}\chi_1 \vee \mathtt{CHP}_{J_1}\chi_2), \tag{1}$$

$$\models_{\mathsf{STIT}} \mathtt{CHP}_{J_1}\chi \rightarrow \mathtt{CHP}_{J_2}\chi, \tag{2}$$

$$\models_{\mathsf{STIT}} (\mathtt{CHP}_{J_1}\chi_1 \wedge \mathtt{CHP}_{J_1}\chi_2) \rightarrow \mathtt{CHP}_{J_1}(\chi_1 \wedge \chi_2), \tag{3}$$

$$\models_{\mathsf{STIT}} \neg\mathtt{CHP}_J\top, \tag{4}$$

$$\models_{\mathsf{STIT}} \neg\mathtt{CHP}_J\bot. \tag{5}$$

Proof. We give the proof of Validity 2 as an example. Let $\mathcal{M}$ be a STIT-model and $w \in W$ such that $\mathcal{M}, w \models \mathtt{CHP}_{J_1}\chi$. 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 $\mathcal{M}, w \models \mathtt{CHP}_{J_2}\chi$. □

According to Validity 1, J_1 could have prevented χ_1 or χ_2 to be true implies 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 $\mathtt{CHP}_J$ 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 forces χ" and,
2. "the group of agents J has not a joint strategy that forces 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 forces χ" and,
4. "the group of agents J did not choose a joint strategy that forces 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 $AGT \setminus \{i\}$ 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 $\mathtt{CHP}_i\,\chi$ which expresses a counterfactual dependence between the state of affairs χ and agent i's choice.

It is worth noting that $\mathtt{CHP}_i\,\chi$ 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 $\mathtt{CHP}_1\,Awin$ and $\mathtt{CHP}_2\,Awin$ are true at w_7. In fact, at w_7 candidate A wins the elections and, given what the other voters have 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, $\mathtt{CHP}_i\,Awin$ 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 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 *df*STIT 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.

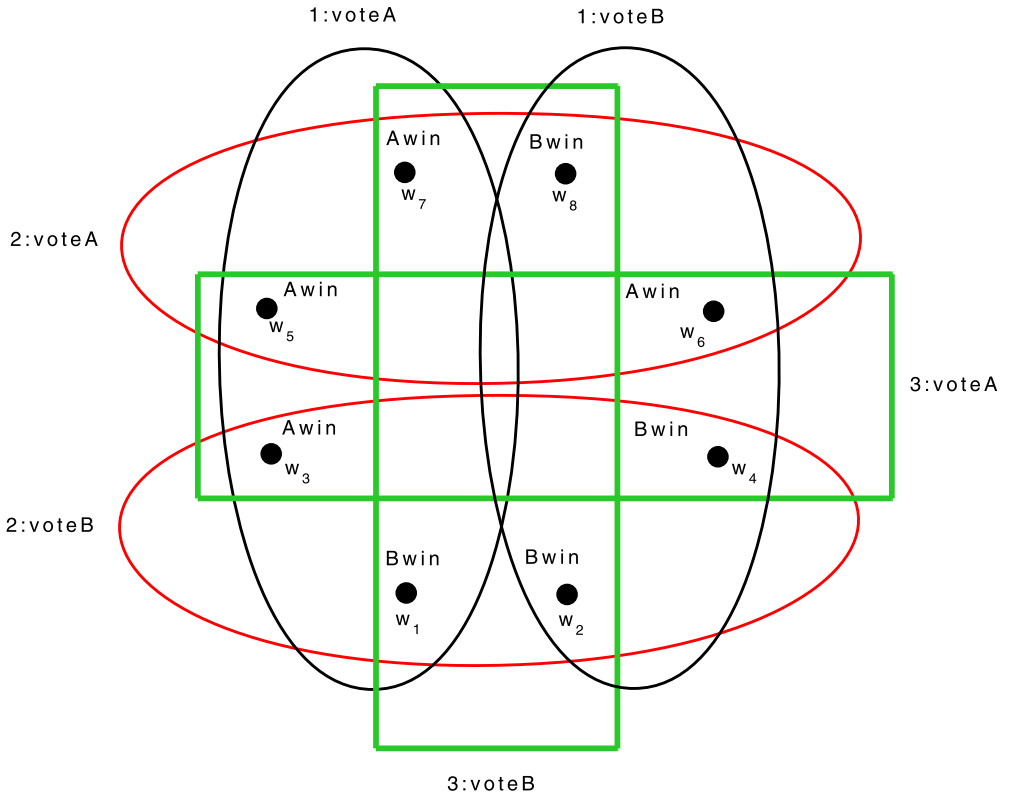


Fig. 4. Vertical circles represent the actions that voter 1 can choose, horizontal circles represent the actions that voter 2 can choose, and rectangles represent the actions that voter 3 can choose. For example, w_1 is the world in which candidate B wins the election and that results from agent 1 voting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

5.1. Definition

First we extend the language $\mathcal{L}_{\text{STIT}}$ of Section 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 $\mathcal{L}_{\text{KSTIT}}$ 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 $\mathcal{L}_{\text{KSTIT}}$. 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 *df*KSTIT.

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

$$\chi ::= \bot \mid p \mid \chi \wedge \chi \mid \neg\chi \quad \text{(propositional formulas)},$$

$$\psi ::= [J]\chi \mid \psi \wedge \psi \quad \text{(see-to-it formulas)},$$

$$\varphi ::= \chi \mid \psi \mid \varphi \wedge \varphi \mid \neg\varphi \mid \langle\emptyset\rangle\psi \mid \mathtt{K}_i\varphi \quad \text{(see-to-it, “can”, knowledge formulas)},$$

where p ranges over *ATM*, i ranges over *AGT* and J over $2^{AGT} \setminus \{\emptyset\}$. For instance, $\mathtt{K}_1\langle\emptyset\rangle[\{1,2\}]p \in \mathcal{L}_{df\text{KSTIT}}$. But $\langle\emptyset\rangle\mathtt{K}_1[\{1,2\}]p \notin \mathcal{L}_{df\text{KSTIT}}$. Let us give the semantics of the logic *df*KSTIT.

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:

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

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 df*KSTIT *is NP-complete if card*$(AGT) = 1$ *and PSPACE-complete if card*$(AGT) \geqslant 2$.

5.3. Axiomatization

The study of an axiomatization for *df*KSTIT 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 Corollary 2 about a complete axiomatization of the logic *df*KSTIT.

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 *df*KSTIT, 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:

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

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

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

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

$$\models_{\mathsf{KSTIT}} \big(\mathtt{DES}_i\,\chi_1 \wedge \mathtt{DES}_i(\chi_1 \rightarrow \chi_2)\big) \rightarrow \mathtt{DES}_i\,\chi_2, \tag{6}$$

$$\models_{\mathsf{KSTIT}} \mathtt{DES}_i\,\chi \leftrightarrow \mathtt{K}_i\mathtt{DES}_i\,\chi, \tag{7}$$

$$\models_{\mathsf{KSTIT}} \neg\mathtt{DES}_i\,\chi \leftrightarrow \mathtt{K}_i\neg\mathtt{DES}_i\,\chi. \tag{8}$$

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:

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

The following is the semantic counterpart of the previous syntactic definition of regret. We have that $\mathcal{M}, w \models \mathtt{REGRET}_i\,\chi$ 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.

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 desire:

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

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:

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

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

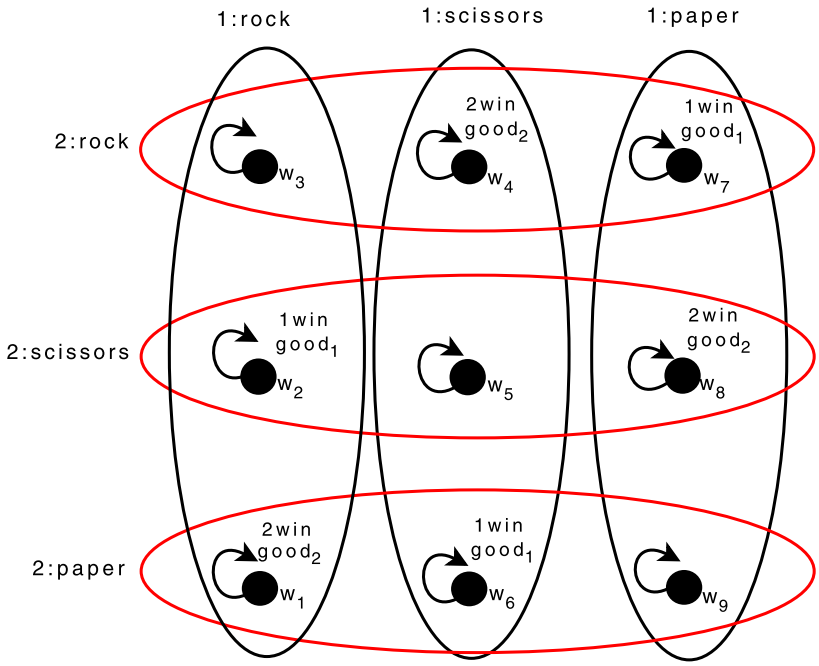


Fig. 5. Vertical circles represent the actions that player 1 can choose, whereas horizontal circles represent the actions that player 2 can choose. For the sake of simplicity, we suppose that players 1 and 2 do not have uncertainty: everywhere in the model players 1 and 2 only consider possible the world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

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 choosing a different action:

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

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

- for all $u \in R_{\emptyset}(v)$, if $\mathcal{M}, u \models \mathit{good}_i$ 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 rejoicing, 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 ($1\mathit{Win}$) and it knows that (by playing *paper* or *scissors*) it could have prevented $1\mathit{Win}$ to be true (i.e. $\texttt{K}_1\texttt{CHP}_1\,1\mathit{Win}$ is true at w_2). Since $\texttt{DES}_1\,1\mathit{Win}$ holds at w_2, it follows that at w_2 player 1 rejoices for having won the game, that is, $\texttt{REJOICE}_1\,1\mathit{Win}$ is true at w_2.

The following validity highlights that rejoicing implies desire satisfaction:

$$\models_{\textsf{KSTIT}} \texttt{REJOICE}_i\,\chi \rightarrow (\texttt{K}_i\chi \wedge \texttt{DES}_i\,\chi). \tag{12}$$

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 introspec-

tion:

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

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

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:

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

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. $\mathtt{REGRET}_1\neg 1\mathit{Win}$), but also he feels disappointed for this (i.e. $\mathtt{DISAPPOINTMENT}_1\neg 1\mathit{Win}$). In fact, at w_1, 1 knows that (by playing *scissors*) the others (i.e. player 2) could have prevented $\neg 1\mathit{Win}$ to be true (i.e. $\mathrm{K}_1\mathtt{CHP}_{AGT\setminus\{1\}}\neg 1\mathit{Win}$ is true at w_1).

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

$$\models_{\mathsf{KSTIT}} \mathtt{DISAPPOINTMENT}_i\,\chi \leftrightarrow \mathrm{K}_i\mathtt{DISAPPOINTMENT}_i\,\chi, \tag{15}$$

$$\models_{\mathsf{KSTIT}} \neg\mathtt{DISAPPOINTMENT}_i\,\chi \leftrightarrow \mathrm{K}_i\neg\mathtt{DISAPPOINTMENT}_i\,\chi. \tag{16}$$

Moreover, like regret, disappointment implies desire frustration:

$$\models_{\mathsf{KSTIT}} \mathtt{DISAPPOINTMENT}_i\,\chi \rightarrow (\mathrm{K}_i\chi \wedge \mathtt{DES}_i\neg\chi). \tag{17}$$

It is worth noting that regret and disappointment do not necessarily occur in parallel, i.e. the formulas $\mathtt{REGRET}_i\,\chi \wedge \neg\mathtt{DISAPPOINTMENT}_i\,\chi$ and $\neg\mathtt{REGRET}_i\,\chi \wedge \mathtt{DISAPPOINTMENT}_i\,\chi$ 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: $\mathtt{GOOD}_1\,\mathit{dinnerTogether}$ and $\mathtt{DES}_1\,\mathit{dinnerTogether}$ 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 ($\neg\mathit{dinnerTogether}$), and 1 knows that (by going to the restaurant) the others (i.e. agent 2) could have prevented $\neg\mathit{dinnerTogether}$ to be true (i.e. $\mathrm{K}_1\mathtt{CHP}_{AGT\setminus\{1\}}\neg\mathit{dinnerTogether}$ is true at w_1). It follows that at w_1 agent 1 feels disappointed for not having dinner with 2, that is, $\mathtt{DISAPPOINTMENT}_1\neg\mathit{dinnerTogether}$ 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. $\mathtt{REGRET}_1\neg\mathit{dinnerTogether}$ is false at w_1). In fact, at w_1, 1 knows that $\neg\mathit{dinnerTogether}$ only depends on what 2 has decided to do. Therefore, at w_1, 1 does not think that he could have prevented $\neg\mathit{dinnerTogether}$ to be true (i.e. $\neg\mathrm{K}_1\mathtt{CHP}_1\neg\mathit{dinnerTogether}$ is true at w_1).

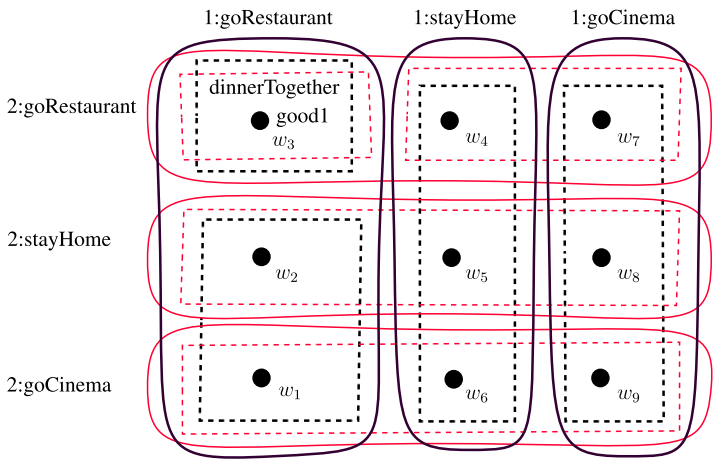


Fig. 6. Again, vertical circles represent the actions that agent 1 can choose, whereas horizontal circles represent the actions that agent 2 can choose. In this example, we suppose that agents 1 and 2 can only have uncertainty about the current choice of the other (vertical dotted rectangles represent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

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

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

$$\models_{\mathsf{KSTIT}} \neg\texttt{ELATION}_i\,\chi \leftrightarrow \texttt{K}_i\neg\texttt{ELATION}_i\,\chi. \tag{19}$$

Moreover, like rejoicing, elation implies desire satisfaction:

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

Finally, like regret and disappointment, elation and rejoicing do not necessarily occur in parallel, i.e. the formulas $\texttt{REJOICE}_i\,\chi \wedge \neg\texttt{ELATION}_i\,\chi$ and $\neg\texttt{REJOICE}_i\,\chi \wedge \texttt{ELATION}_i\,\chi$ 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 $\texttt{REGRET}_i\,\chi$ and $\texttt{REJOICE}_i\,\chi$ 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 al.'s individual STIT logic [9] extended by knowledge operators would be sufficient to write the formulas $\texttt{DISAPPOINTMENT}_i\,\chi$ and $\texttt{ELATION}_i\,\chi$. Let us recall that the fusion of two decidable modal logics is decidable [24]. As the satisfiability problem of Belnap et al.'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 health care, 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:

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

Note that all the four formulas are in $\mathcal{L}_{df\mathsf{KSTIT}}$, even the third one which is equivalent to $\mathrm{K}_t\mathrm{K}_u\neg\langle\emptyset\rangle([u]studied_u \wedge \neg pass_u)$. 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

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

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:

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

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:

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

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 $df\mathsf{KSTIT}^+$, which is nothing else than the dynamic extension of the syntactic fragment dfKSTIT of the logic KSTIT studied in Section 5.

$$\chi ::= \bot \mid p \mid \chi \wedge \chi \mid \neg\chi \quad \text{(propositional formulas)},$$

$$\psi ::= [J]\chi \mid \psi \wedge \psi \quad \text{(see-to-it formulas)},$$

$$\varphi ::= \chi \mid \psi \mid \varphi \wedge \varphi \mid \neg\varphi \mid \langle\emptyset\rangle\psi \mid \mathtt{K}_i\varphi \mid \big[\|\varphi\|\big]\varphi \quad \text{(see-to-it, "can", knowledge, update formulas)},$$

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:

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

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 $i \in AGT$, $E_i^{\|\theta\|} = \{(w,v) \mid (w,v) \in E_i \text{ and } (M, w \models \theta \text{ iff } M, v \models \theta)\}$;
- $V^{\|\theta\|} = V$.

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 $[\|\theta\|]$:

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

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*:

$$(Red_p) \quad \big[\|\theta\|\big]p \leftrightarrow p,$$
$$(Red_\neg) \quad \big[\|\theta\|\big]\neg\varphi \leftrightarrow \neg\big[\|\theta\|\big]\varphi,$$
$$(Red_\wedge) \quad \big[\|\theta\|\big](\varphi_1 \wedge \varphi_2) \leftrightarrow \big(\big[\|\theta\|\big]\varphi_1 \wedge \big[\|\theta\|\big]\varphi_2\big),$$
$$(Red_{[J]}) \quad \big[\|\theta\|\big][J]\varphi \leftrightarrow [J]\big[\|\theta\|\big]\varphi,$$
$$(Red_{\mathrm{K}_i}) \quad \big[\|\theta\|\big]\mathrm{K}_i\varphi \leftrightarrow \big(\big(\theta \rightarrow \mathrm{K}_i\big(\theta \rightarrow \big[\|\theta\|\big]\varphi\big)\big) \wedge \big(\neg\theta \rightarrow \mathrm{K}_i\big(\neg\theta \rightarrow \big[\|\theta\|\big]\varphi\big)\big)\big).$$

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

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

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. $red(\mathrm{K}_i\varphi) = \mathrm{K}_i red(\varphi)$
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. $red([\|\theta\|]\mathrm{K}_i\varphi) = red((\theta \rightarrow \mathrm{K}_i(\theta \rightarrow [\|\theta\|]\varphi)) \wedge (\neg\theta \rightarrow \mathrm{K}_i(\neg\theta \rightarrow [\|\theta\|]\varphi)))$
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*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 KSTIT$^+$ are completely axiomatized by the axioms and inference rules of df 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*KSTIT$^+$ follows straightforwardly from the known decidability of the base logic *df*KSTIT (Theorem 3) and from Propositions 3 and 4. Indeed, *red* provides an effective procedure for reducing a *df*KSTIT$^+$-formula φ into an equivalent *df*KSTIT-formula $red(\varphi)$.

Corollary 5. *The satisfiability problem of df KSTIT$^+$ is decidable.*

7.3. Adapting behavior during a dialogue with the user

It has to be noted that in PAL 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 $\mathrm{K}_i\varphi$. Thus, we write $say(i,\varphi)$ instead of $!\mathrm{K}_i\varphi$, and $[say(i,\varphi)]\psi$ abbreviates $[!\mathrm{K}_i\varphi]\psi$. 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 $\mathrm{K}_i\varphi$ (i.e. $!\mathrm{K}_i\varphi$), $\langle\!\langle say(i,\varphi)\rangle\!\rangle\psi$ abbreviates $\langle\!\langle !\mathrm{K}_i\varphi\rangle\!\rangle\psi$.

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. $\mathrm{K}_i\varphi$). 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:

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

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:

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

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:

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

- $\mathrm{K}_t\mathrm{K}_u\langle\emptyset\rangle[u]\neg studied_u$
- $\mathrm{K}_t\, pass_u$.

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 emotion 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). Similarly 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 article. This framework is based on a decidable and finitely axiomatizable fragment of STIT logic called *df*STIT. We have shown that an epistemic extension of *df*STIT called *df*KSTIT 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 *df*KSTIT is NP-complete if $card(AGT) = 1$ and PSPACE-complete if $card(AGT) \geqslant 2$. This first result is fundamental in order to claim that we can write down algorithms in *df*KSTIT to reason about counterfactual emotions such as regret, rejoicing, disappointment and elation. Moreover, we have provided a complete axiomatization of *df*KSTIT logic. This second result is also important because it shows that we can perform syntactic reasoning in *df*KSTIT 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 *df*STIT (NP-complete) and the complexity of the satisfiability problem of a formula in *df*KSTIT (PSPACE-complete). Of course, the complexity for *df*KSTIT cannot 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 *df*KSTIT called *df*KSTIT$^+$ 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*.

Appendix A. Annex

A.1. Proof of Proposition 1

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

- If $card(AGT) \leqslant 2$: φ is STIT-satisfiable iff φ is NCL-satisfiable;
- If $card(AGT) \geqslant 3$: 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 $w R_\emptyset v$, 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 $w R_J \circ R_{\overline{J}} v$.

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 $card(AGT) \geqslant 3$, 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 $card(AGT) \geqslant 3$, 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. □

A.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 $J_1, \ldots, J_r \subseteq AGT$ 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 $J, K \subseteq AGT$ be two coalitions such that $J \cap K = \emptyset$. As $uR_\emptyset w$, we have $uR_J \circ R_{\bar{J}} w$. 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 $J_1, \ldots, J_r, J_{r+1} \subseteq AGT$ 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 $r \geqslant 1$, $\mathcal{P}(r)$ is true. □

A.3. Proof of Theorem 2

Let $\varphi \in \mathcal{L}_{df\text{STIT}}$. 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 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:

- $Pos = \{\psi \mid \langle\emptyset\rangle\psi \in SF(\varphi) \text{ and } \mathcal{M}, z_0 \models \langle\emptyset\rangle\psi\} \cup \{Z_0\}$ 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 $W' = C^n = C \times C \times \cdots \times C$ where $C = Pos \cup \{0, \ldots, card(Neg) - 1\}$. 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:

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

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

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

  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 dfSTIT). 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 $i : \{0, \ldots, card(Neg) - 1\} \to Neg$.
  We need such a bijection between integers in $\{0, \ldots, card(Neg) - 1\}$ 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:

$$i(\vec{x}) = i\left(\sum_{j \in \{1,\ldots,n\} \mid x_j \in \{0,\ldots,card(Neg)-1\}} x_j \bmod card(Neg)\right)$$

  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 $y \in \bigcap_{P \in Pos} R_{Coord^{\vec{x}}_{=P}}(y_P)$.

[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 $P = Q$.

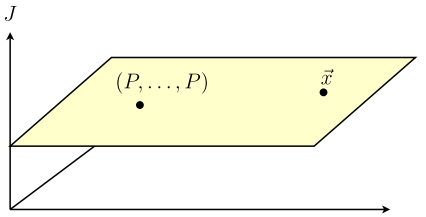


Fig. 7. Model $\mathcal{M}'$: a point $(P,\ldots,P)$ and the subspace of all points $\vec{x}$ such that $\vec{x}_J = (P,\ldots,P)_J$, that is, the subspace of worlds of $\mathcal{M}'$ where every agent in J performs action "P".

2. We have defined $y \in W$. Now let us prove that $\mathcal{M}, y \models \mathit{Boxes}_{\vec{x}}$. In other words, we are going to prove that for all $\chi \in \mathit{Pos}_{\vec{x}}$, $\mathcal{M}, y \models \chi$.
   Let $\chi \in \mathit{Pos}_{\vec{x}}$. By definition of $\mathit{Pos}_{\vec{x}}$, there exists $P \in \mathit{Pos}$ and $[J]\chi \in \mathit{SF}(P)$ such that $J \subseteq \mathit{Coord}^{\vec{x}}_{=P}$. 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 $y_P R_{\mathit{Coord}^{\vec{x}}_{=P}} y$. But as $J \subseteq \mathit{Coord}^{\vec{x}}_{=P}$, we have $R_{\mathit{Coord}^{\vec{x}}_{=P}} \subseteq R_J$. So, we have $y_P R_J y$ and, consequently, we have $\mathcal{M}, y \models \chi$. So we have $\mathcal{M}, y \models \mathit{Boxes}_{\vec{x}}$. □

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 $[J]\chi \in \mathit{Neg}$.
  - If there exists $y \in W$ such that $\mathcal{M}, y \models \neg\chi \wedge \mathit{Boxes}_{\vec{x}}$ then $f(\vec{x}) \stackrel{\text{def}}{=} y$.
  - Else, we choose a world y in W such that $\mathcal{M}, y \models \mathit{Boxes}_{\vec{x}}$ (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.

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

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

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 $[J]\chi \in \mathit{SF}_1(\varphi)$, 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 $P \in \mathit{Pos}$ we have $\mathcal{M}', (P,\ldots P) \models P$. Let $P \in \mathit{Pos}$. P is a conjunction of formulas of the form $[J]\chi$ where χ is a Boolean formula. Let $[J]\chi \in \mathit{SF}(P)$. 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 $J \subseteq \mathit{Coord}^{\vec{x}}_{=P}$. So $\chi \in \mathit{Pos}_{\vec{x}}$ implies that $\models \mathit{Boxes}_{\vec{x}} \rightarrow \chi$. But, by Lemma 3, we have $\mathcal{M}', \vec{x} \models \mathit{Boxes}_{\vec{x}}$ 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 $\langle\emptyset\rangle N \in \mathit{SF}(\varphi)$ 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 Fig. 8 during this part.

Consider $y_0 = f(\vec{x}_0) \in W$. By definition of f, we have $\mathcal{M}, y_0 \models \mathit{Boxes}_{\vec{x}_0}$. 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*.

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 Fig. 8, we want that J_i performs the same joint action both in $\vec{x}_0$ and in $\vec{x}$.

The case where $J_i = \mathit{AGT}$ 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 $j \in \overline{J_i} \setminus \{j_0\}$;
- $x_{j_0} = i^{-1}([J_i]\neg\chi_i) - \sum_{j \in \mathit{AGT} \mid j \neq j_0 \text{ and } x_j \in \{0,\ldots,\mathit{card}(\mathit{Neg})-1\}} x_j \bmod N$.

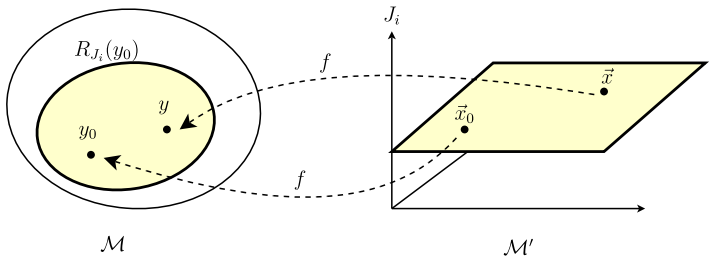


Fig. 8. Proof of $\Leftarrow$ of Fact 1.

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 $\mathcal{M}, y \models \mathit{Boxes}_{\vec{x}}$. Indeed, $\mathit{Boxes}_{\vec{x}}$ 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 $\models \mathit{Boxes}_{\vec{x}_0} \rightarrow \mathit{Boxes}_{\vec{x}}$. Hence $\mathcal{M}, y \models \mathit{Boxes}_{\vec{x}}$. To sum up, we have $\mathcal{M}, y \models \mathit{Boxes}_{\vec{x}} \wedge \neg\chi_i$. 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 $\mathcal{M}, y \models \mathit{Boxes}_{\vec{x}} \wedge \neg\chi_i$. 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 $\mathcal{M}', (Z_0, \ldots, Z_0) \models \mathit{Neg\_in\_z_0}$. More precisely we prove that for all $[J]\chi \in \mathit{Neg\_in\_z_0}$, $\mathcal{M}', (Z_0, \ldots, Z_0) \models \langle J\rangle\neg\chi$. We know that $\mathcal{M}, z_0 \models \langle J\rangle\neg\chi$. So there exists $y \in W$ such that yR_Jz_0 and $\mathcal{M}, y \models \neg\chi$. The case $J = \mathit{AGT}$ 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 $\mathit{Boxes}_{\vec{x}}$ only contains Boolean formulas χ' where formulas $[J']\chi'$ are subformulas of Z_0, where $J' \subseteq J$. Hence $\mathcal{M}, y \models \mathit{Boxes}_{\vec{x}}$. 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 $\mathcal{M}, y \models \neg\chi \wedge \mathit{Boxes}_{\vec{x}}$. By definition of V', $\mathcal{M}', \vec{x} \models \neg\chi$.

The conclusion of the proof is left to the reader. □

A.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. □

A.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 $\vdash_{\mathsf{NCL}} \varphi$. □

A.6. Proof of Theorem 3

The satisfiability problem of dfKSTIT is NP-complete if $\mathit{card}(\mathit{AGT}) = 1$ and PSPACE-complete if $\mathit{card}(\mathit{AGT}) \geqslant 2$.

function sat(Σ, i)
  ($\exists$) **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.

  ($\forall$) **choose** $S' \in \beta$
  ($\forall$) **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

Fig. 9. An algorithm for the KSTIT-satisfiability problem of a given set of *df*KSTIT-formulas Σ.

Proof. $\boxed{card(AGT) = 1}$

Let us consider the case $card(AGT) = 1$. In this case there are only three operators: [Ø], [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 cannot appear after a [Ø] operator, we can prove by a selected points argument that if a *df*KSTIT-formula is KSTIT-satisfiable, then it is in a polynomial sized model (in [35], it is done for $S5$).

$\boxed{card(AGT) \geqslant 2}$

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

Now let us prove that the KSTIT-satisfiability problem of a given *df*KSTIT-formula is in PSPACE. As APTIME = PSPACE [15], it is sufficient to prove that this problem is in APTIME. Fig. 9 shows an alternating procedure sat(Σ, i) where Σ is a set of *df*KSTIT-formulas and $i \in AGT$. For all $i \in AGT$, 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 Fig. 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. □

A.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] cannot be applied here because we are dealing with syntactic fragments.

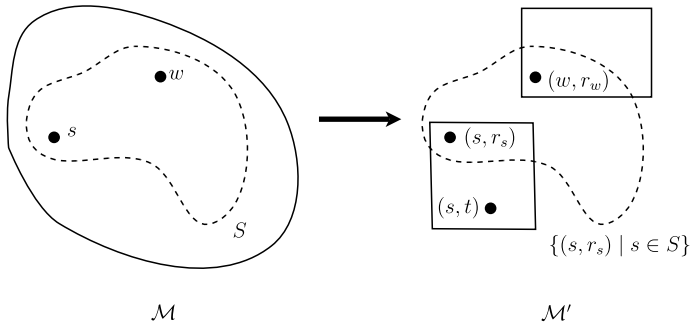


Fig. 10. From KNCL-model $\mathcal{M}$ to the KSTIT-model $\mathcal{M}'$.

⇑ We can prove that a KSTIT-model is a KNCL-model. The proof is similar to the proof of Proposition 1.

⇓ 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 Theorem 2 as shown in Fig. 10.

Consider the set:

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

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

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

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

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

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 $\mathcal{M}', (s, r_s) \models \varphi$.

(*df*STIT) 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 $\mathcal{M}', (s, r_s) \models \mathtt{K}_i\varphi$. □

A.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 $(\mathtt{K}_i\chi_1 \wedge \cdots \wedge \mathtt{K}_i\chi_n) \leftrightarrow \mathtt{K}_i(\chi_1 \wedge \cdots \wedge \chi_n)$
26. $\vdash_{\mathsf{KNCL}} \mathcal{KB}^* \rightarrow \mathtt{REJOICE}_u pass_u$ from 25 and $\mathcal{KB}^* = \mathtt{K}_t\mathtt{DES}_u pass_u \wedge \mathtt{K}_t\mathtt{K}_u pass_u \wedge \mathtt{K}_t\mathtt{K}_u\langle\emptyset\rangle[u]\neg studied_u \wedge \mathtt{K}_t\mathtt{K}_u[\emptyset]([u]\neg studied_u \rightarrow \neg pass_u)$ □

A.9. Proof of Proposition 2

The following schemata are KSTIT^+-valid:

(Red_p) $\quad \big[\|\theta\|\big]p \leftrightarrow p$

$(Red_\neg)$ $\quad \big[\|\theta\|\big]\neg\varphi \leftrightarrow \neg\big[\|\theta\|\big]\varphi$

$(Red_\wedge)$ $\quad \big[\|\theta\|\big](\varphi_1 \wedge \varphi_2) \leftrightarrow \big(\big[\|\theta\|\big]\varphi_1 \wedge \big[\|\theta\|\big]\varphi_2\big)$

$(Red_{[J]})$ $\quad \big[\|\theta\|\big][J]\varphi \leftrightarrow [J]\big[\|\theta\|\big]\varphi$

$(Red_{\mathtt{K}_i})$ $\quad \big[\|\theta\|\big]\mathtt{K}_i\varphi \leftrightarrow \big(\big(\theta \rightarrow \mathtt{K}_i\big(\theta \rightarrow \big[\|\theta\|\big]\varphi\big)\big) \wedge \big(\neg\theta \rightarrow \mathtt{K}_i\big(\neg\theta \rightarrow \big[\|\theta\|\big]\varphi\big)\big)\big)$

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 $M, w \models \mathrm{K}_i(\theta \to [\|\theta\|]\varphi)$) and (if $M, w \models \neg\theta$ then $M, w \models \mathrm{K}_i(\neg\theta \to [\|\theta\|]\varphi)$),
IFF $M, w \models (\theta \to \mathrm{K}_i(\theta \to [\|\theta\|]\varphi)) \wedge (\neg\theta \to \mathrm{K}_i(\neg\theta \to [\|\theta\|]\varphi))$.

$M, w \models [\|\theta\|][J]\varphi$,
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 $M, w \models [J][\|\theta\|]\varphi$. □

A.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 $\mathrm{K}_i\varphi_1 \in \mathcal{L}_{df\mathsf{KSTIT}^+}$ then $red(\mathrm{K}_i\varphi_1) \in \mathcal{L}_{df\mathsf{KSTIT}}$.
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 $\mathrm{K}_i\varphi_1 \in \mathcal{L}_{df\mathsf{KSTIT}^+}$. 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 $\mathrm{K}_i red(\varphi_1) \in \mathcal{L}_{df\mathsf{KSTIT}}$. The latter implies $red(\mathrm{K}_i\varphi_1) \in \mathcal{L}_{df\mathsf{KSTIT}}$, as $\mathrm{K}_i red(\varphi_1) = red(\mathrm{K}_i\varphi_1)$.

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 $\varphi = \mathrm{K}_i\varphi_1$.

Suppose $red(\theta) \in \mathcal{L}_{df\mathsf{KSTIT}}$ and $\mathrm{K}_i\varphi_1 \in \mathcal{L}_{df\mathsf{KSTIT}}$. 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

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

By definition of *red*, we have that

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

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

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 $red([\|\theta\|]\varphi_1) \in \mathcal{L}_{df\mathsf{KSTIT}}$. □

A.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 dfKSTIT (Corollary 3), $red(\varphi)$ is also provable there. $df\mathsf{KSTIT}^+$ being a conservative extension of dfKSTIT, $red(\varphi)$ 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 $df\mathsf{KSTIT}^+$. □

A.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 A.8 in this appendix), 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 $[\|\theta\|](\varphi_1 \wedge \cdots \wedge \varphi_n)$ is equivalent to $[\|\theta\|]\varphi_1 \wedge \cdots \wedge [\|\theta\|]\varphi_n$ (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 $\texttt{REJOICE}_u pass_u$ abbreviates $\texttt{K}_u pass_u \wedge \texttt{DES}_u pass_u \wedge \texttt{K}_u\langle AGT \setminus \{u\}\rangle \neg pass_u$, from the previous validity F, we have:

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

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

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

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

1. $\models_{\text{KSTIT}^+} \mathcal{KB}^{***} \rightarrow \texttt{K}_t\mathcal{KB}^{***}$ by Axiom 4 and Axiom 5 for $\texttt{K}_t$
2. $\models_{\text{KSTIT}^+} \texttt{K}_t(\mathcal{KB}^{***} \rightarrow [\|\texttt{K}_t pass_u\|]\texttt{REJOICE}_u pass_u)$ by the previous validity G and necessitation rule for $\texttt{K}_t$
3. $\models_{\text{KSTIT}^+} \mathcal{KB}^{***} \rightarrow \texttt{K}_t[\|\texttt{K}_t pass_u\|]\texttt{REJOICE}_u pass_u$ by 1, 2 and Axiom K for $\texttt{K}_t$
4. $\models_{\text{KSTIT}^+} \mathcal{KB}^{***} \rightarrow \texttt{K}_t[say(t, pass_u)]\texttt{REJOICE}_u pass_u$ by 3, the validity $[\|\theta\|]\varphi \rightarrow [!\theta]\varphi$, Axiom K and necessitation rule for $\texttt{K}_t$. □

References

[1] C. Adam, A. Herzig, D. Longin, A logical formalization of the OCC theory of emotions, Synthese 168 (2) (2009) 201–248.
[2] T. Ågotnes, W. van der Hoek, M. Wooldridge, Quantified coalition logic, in: Proceedings of the Twentieth International Joint Conference on Artificial Intelligence (IJCAI'07), AAAI Press, 2007, pp. 1181–1186.
[3] T. Ågotnes, H. van Ditmarsch, Coalitions and announcements, in: Proceedings of the Seventh International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2008), ACM Press, 2008, pp. 673–680.
[4] R. Alur, T. Henzinger, Alternating-time temporal logic, Journal of the ACM 49 (2002) 672–713.
[5] P. Balbiani, O. Gasquet, A. Herzig, F. Schwarzentruber, N. Troquard, Coalition games over Kripke semantics, in: C. Dégremont, L. Keiff, H. Rückert (Eds.), Festschrift in Honour of Shahid Rahman, College Publications, 2008, pp. 1–12.
[6] P. Balbiani, O. Gasquet, E. Lorini, F. Schwarzentruber, An alternating procedure for the satisfiability problem of $S5_n$, Tech. Rep. IRIT/RT–2009-1-FR, IRIT, 2009.
[7] P. Balbiani, A. Herzig, N. Troquard, Alternative axiomatics and complexity of deliberative STIT theories, Journal of Philosophical Logic 37 (4) (2008) 387–406.
[8] J. Bates, The role of emotion in believable agents, Communications of the ACM 37 (7) (1994) 122–125.
[9] N. Belnap, M. Perloff, M. Xu, Facing the Future: Agents and Choices in Our Indeterminist World, Oxford, 2001.
[10] P. Blackburn, M. de Rijke, Y. Venema, Modal Logic, Cambridge University Press, 2001.
[11] J. Broersen, 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, 2010, pp. 215–219.
[12] J. Broersen, A. Herzig, N. Troquard, Embedding Alternating-time temporal logic in strategic STIT logic of agency, Journal of Logic and Computation 16 (5) (2006) 559–578.
[13] J. Broersen, A. Herzig, N. Troquard, Normal Coalition Logic and its conformant extension, in: D. Samet (Ed.), Theoretical Aspects of Rationality and Knowledge (TARK), Brussels, 25/06/2007–27/06/2007, Presses universitaires de Louvain, 2007, pp. 91–101.
[14] C. Castelfranchi, 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, 2005, pp. 258–276.
[15] A.K. Chandra, D.C. Kozen, L.J. Stockmeyer, Alternation, J. ACM 28 (1) (1981) 114–133.
[16] H. Chockler, J.Y. Halpern, Responsibility and blame: A structural-model approach, Journal of Artificial Intelligence Research 22 (2004) 93–115.
[17] P.R. Cohen, H.J. Levesque, Intention is choice with commitment, Artificial Intelligence 42 (2–3) (1990) 213–261.
[18] F. de Rosis, C. Pelachaud, I. Poggi, V. Carofiglio, B. De Carolis, 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 (2003) 81–118.
[19] W.W. Dijk, M. Zeelenberg, Investigating the appraisal patterns of regret and disappointment, Motivation and Emotion 26 (4) (2002) 321–331.
[20] M.S. El-Nasr, J. Yen, T.R. Ioerger, FLAME: Fuzzy logic adaptive model of emotions, Autonomous Agents and Multi-Agent Systems 3 (3) (2000) 219–257.
[21] C. Elliot, The affective reasoner: A process model for emotions in a multi-agent system, Ph.D. thesis, Northwestern University, Institute for Learning Sciences, 1992.
[22] N. Frijda, The Emotions, Cambridge University Press, 1986.
[23] N.H. Frijda, P. Kuipers, E. Ter Schure, Relations among emotion, appraisal, and emotional action readiness, Journal of Personality and Social Psychology 57 (2) (1989) 212–228.
[24] D.M. Gabbay, A. Kurucz, F. Wolter, M. Zakharyaschev, Many-Dimensional Modal Logics: Theory and Applications, Studies in Logic and the Foundations of Mathematics, vol. 148, Elsevier, North-Holland, 2003.
[25] O. Gasquet, A. Herzig, Translating non-normal modal logics into normal modal logics, in: A. Jones, M. Sergot (Eds.), Proceedings International Workshop on Deontic Logic in Computer Science (DEON'94), TANO, Oslo, 1993.
[26] R.M. Gordon, The Structure of Emotions, Cambridge University Press, Cambridge, 1987.
[27] J. Gratch, S. Marsella, A domain-independent framework for modelling emotions, Journal of Cognitive Systems Research 5 (4) (2004) 269–306.
[28] J.Y. Halpern, Y. Moses, A guide to completeness and complexity for modal logics of knowledge and belief, Artificial Intelligence 54 (2) (1992) 319–379.
[29] A. Herzig, F. Schwarzentruber, Properties of logics of individual and group agency, in: Proceedings of Advances in Modal Logic 2008, College Publication, 2008, pp. 133–149.
[30] J.F. Horty, Agency and Deontic Logic, Oxford University Press, 2001.
[31] J.F. Horty, N. Belnap, The deliberative STIT: A study of action, omission, and obligation, Journal of Philosophical Logic 24 (6) (1995) 583–644.
[32] D. Kahneman, Varieties of counterfactual thinking, in: N.J. Roese, J.M. Olson (Eds.), What Might Have Been: The Social Psychology of Counterfactual Thinking, Erlbaum, 1995.
[33] D. Kahneman, D.T. Miller, Norm theory: comparing reality to its alternatives, Psychological Review 93 (2) (1986) 136–153.
[34] D. Kahneman, A. Tversky, The psychology of preferences, Scientific American 246 (1982) 160–173.

[35] R.E. Ladner, The computational complexity of provability in systems of modal propositional logic, SIAM Journal on Computing 6 (3) (1977) 467–480.
[36] R.S. Lazarus, Emotion and Adaptation, Oxford University Press, New York, 1991.
[37] G. Loomes, R. Sugden, Regret theory: an alternative theory of rational choice under uncertainty, Economic Journal 92 (4) (1982) 805–824.
[38] G. Loomes, R. Sugden, Testing for regret and disappointment in choice under uncertainty, Economic Journal 97 (1987) 118–129.
[39] E. Lorini, C. Castelfranchi, The cognitive structure of surprise: looking for basic principles, Topoi: An International Review of Philosophy 26 (1) (2007) 133–149.
[40] J.-J.C. Meyer, Reasoning about emotional agents, International Journal of Intelligent Systems 21 (6) (2006) 601–619.
[41] J.J.C. Meyer, W. van der Hoek, B. van Linder, A logical approach to the dynamics of commitments, Artificial Intelligence 113 (1–2) (1999) 1–40.
[42] A. Ortony, G.L. Clore, A. Collins, The Cognitive Structure of Emotions, Cambridge University Press, 1988.
[43] M. Pauly, A modal logic for coalitional power in games, Journal of Logic and Computation 12 (1) (2002) 149–166.
[44] R.W. Picard, Affective Computing, MIT Press, 1997.
[45] J.A. Plaza, Logics of public communications, in: M. Emrich, M. Pfeifer, M. Hadzikadic, Z. Ras (Eds.), Proceedings of the 4th International Symposium on Methodologies for Intelligent Systems, 1989, pp. 201–216.
[46] W.S. Reilly, J. Bates, Building emotional agents, Tech. Rep., CMUCS-92-143, School of Computer Science, Carnegie Mellon University, 1992.
[47] R. Reisenzein, Emotional experience in the computational belief-desire theory of emotion, Emotion Review 1 (3) (2009) 214–222.
[48] N.J. Roese, Counterfactual thinking, Psychological Bulletin 121 (1) (1997) 133–148.
[49] N.J. Roese, L.J. Sanna, A.D. Galinsky, The mechanics of imagination: automaticity and control in counterfactual thinking, in: R.R. Hassin, J.S. Uleman, J.A. Bargh (Eds.), The New Unconscious, Oxford University Press, 2005.
[50] I.J. Roseman, A model of appraisal in the emotion system, in: K.R. Scherer, A. Schorr, T. Johnstone (Eds.), Appraisal Processes in Emotion: Theory, Methods, Research, Oxford University Press, Oxford, 2001.
[51] I.J. Roseman, A.A. Antoniou, P.E. Jose, Appraisal determinants of emotions: constructing a more accurate and comprehensive theory, Cognition and Emotion 10 (1996) 241–277.
[52] K. Scherer, Appraisal considered as a process of multilevel sequential checking, in: K.R. Scherer, A. Schorr, T. Johnstone (Eds.), Appraisal Processes in Emotion: Theory, Methods, Research, Oxford University Press, Oxford, 2001.
[53] K.R. Scherer, A. Schorr, T. Johnstone (Eds.), Appraisal Processes in Emotion: Theory, Methods, Research, Oxford University Press, Oxford, 2001.
[54] F. Schwarzentruber, Décidabilité et complexité de la logique normale des coalitions, Master's thesis, Univ. Paul Sabatier Toulouse III, 2007.
[55] J. Searle, Intentionality: An Essay in the Philosophy of Mind, Cambridge University Press, New York, 1983.
[56] C. Smith, R. Lazarus, Emotion and adaptation, in: J. Pervin (Ed.), Handbook of Personality: Theory & Research, Guilford Press, New York, 1990.
[57] R.C. Solomon, The Passions, University of Notre Dame Press, Notre Dame, 1976.
[58] B.R. Steunebrink, M. Dastani, J.-J.C. Meyer, A logic of emotions for intelligent agents, in: Proceedings of AAAI'07, AAAI Press, 2007, pp. 142–147.
[59] R. Sugden, Regret, recrimination and rationality, Theory and Decision 19 (1) (1985) 77–99.
[60] G. Taylor, Pride, Shame and Guilt: The Emotions of Self-Assessment, Oxford University Press, New York, 1985.
[61] N. Troquard, Independent agents in branching time, Ph.D. thesis, Univ. Paul Sabatier Toulouse III & Univ. degli studi di Trento, 2007.
[62] P. Turrini, J.-J.C. Meyer, C. Castelfranchi, Coping with shame and sense of guilt: a dynamic logic account, Journal of Autonomous Agents and Multi-Agent Systems 20 (3) (2009) 401–420.
[63] J. van Benthem, F. Liu, Dynamic logic of preference upgrade, Journal of Applied Non-Classical Logics 17 (2) (2007) 157–182.
[64] J. van Benthem, S. Minică, Towards a dynamic logic of questions, in: Proceedings of Second International Workshop on Logic, Rationality and Interaction (LORI-II), in: LNCS, vol. 5834, Springer-Verlag, 2009, pp. 27–41.
[65] W. van der Hoek, W. Jamroga, M. Wooldridge, A logic for strategic reasoning, in: Proceedings of the Fourth International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2005), ACM Press, New York, 2005, pp. 157–164.
[66] W. van der Hoek, M. Wooldridge, Cooperation, knowledge, and time: Alternating-time temporal epistemic logic and its applications, Studia Logica 75 (2003) 125–157.
[67] H. van Ditmarsch, W. van der Hoek, B. Kooi, Dynamic Epistemic Logic, Synthese Library, vol. 337, Springer, 2007.
[68] M. Zeelenberg, J. Beattie, J. van der Pligt, N.K. de Vries, Consequences of regret aversion: effects of expected feedback on risky decision making, Organizational Behavior and Human Decision Processes 65 (2) (1996) 148–158.
[69] M. Zeelenberg, W. van Dijk, A.S.R. Manstead, J. van der Pligt, On bad decisions and disconfirmed expectancies: The psychology of regret and disappointment, Cognition and Emotion 14 (4) (2000) 521–541.
[70] M. Zeelenberg, W. van Dijk, J. van der Pligt, A.S.R. Manstead, P. van Empelen, D. Reinderman, 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) (1998) 117–141.
[71] M. Zeelenberg, W.W. van Dijk, A.S.R. Manstead, Reconsidering the relation between regret and responsibility, Organizational Behavior and Human Decision Processes 74 (3) (1998) 254–272.

Synthese
DOI 10.1007/s11229-015-0960-5

S.I. : LOGIC AND THE FOUNDATIONS OF DECISION AND GAME THEORY (LOFT)

A minimal logic for interactive epistemology

Emiliano Lorini[1]

Received: 21 December 2014 / Accepted: 26 October 2015

Abstract We propose a minimal logic for interactive epistemology based on a qualitative representation of epistemic individual and group attitudes including knowledge, belief, strong belief, common knowledge and common belief. We show that our logic is sufficiently expressive to provide an epistemic foundation for various game-theoretic solution concepts including "1-round of deletion of weakly dominated strategies, followed by iterated deletion of strongly dominated strategies" (DWDS^1-IDSDS) and "2-rounds of deletion of weakly dominated strategies, followed by iterated deletion of strongly dominated strategies" (DWDS^2-IDSDS). Axiomatization and complexity results for the logic are given in the paper.

Keywords Epistemic logic · Epistemic game theory · Axiomatization · Complexity

1 Introduction

As emphasized by Aumann (1999), interactive epistemology deals with formal models of knowledge and belief when there is more than one rational agent or "player" in the context of interaction having not only knowledge and beliefs about substantive matters, but also knowledge and beliefs about the others' knowledge and beliefs. The aim of the epistemic program in game theory is to use such models in order to provide an epistemic foundation for various game-theoretic solution concepts.

Much of the work in the field of epistemic game theory is based either on a *quantitative* representation of uncertainty and epistemic attitudes or on a *semi-qualitative* one. Examples of the former are the analysis of the epistemic foundations for forward

✉ Emiliano Lorini
emiliano.lorini@irit.fr; lorini@irit.fr

1 IRIT–CNRS, Toulouse University, Toulouse, France

Published online: 15 December 2015

induction and for iterated admissibility based on Bayesian probabilities (Stalnaker 1998; Halpern and Pass 2009), conditional probabilities (Battigalli and Siniscalchi 2002) or lexicographic probabilities (Brandenburger et al. 2008). An example of the latter is the analysis of the epistemic foundation for iterated weak dominance by Lorini (2013) based on Spohn's theory of uncertainty (Spohn 1988). There are also few examples of *qualitative* models of epistemic attitudes and uncertainty which have been used in the field of epistemic game theory. For instance, Baltag et al. (2009) have proposed an analysis of the epistemic foundation for backward induction based on a purely qualitative notion of plausibility. The distinction between purely quantitative, semi-qualitative and qualitative approaches to uncertainty has been widely discussed in the AI literature (see, e.g., Pearl 1993; Weydert 1994). While in purely quantitative approaches belief states are characterized by classical probabilistic measures or by alternative numerical accounts, such as lexicographic probabilities or conditional probabilities (Battigalli and Siniscalchi 2002), in semi-qualitative approaches, such as Spohn's theory, belief states are described by rough qualitative measures assigning orders of magnitude. Finally, qualitative approaches do not use any numerical representation of uncertainty but simply a plausibility ordering on possible worlds structures inducing an epistemic-entrenchment-like ordering on propositions.

The aim of the present work is to propose a *minimal logic* for interactive epistemology based on a qualitative representation of epistemic individual and group attitudes including knowledge, belief, strong belief, common knowledge and common belief. We call LEG our logic, which stands for Logic of Epistemic attitudes in Games. Our logic is "minimal" in the sense we make as few assumptions as possible about the representation of epistemic states. Specifically, we assume that players in a game have only two kinds of epistemic attitudes: (i) each player envisages a set of worlds or states (also called epistemic alternatives) that defines the player's *information set*, and (ii) among these epistemic alternatives the player can distinguish possible alternatives, that define the player's *belief set*, from impossible ones. Our qualitative semantics for uncertainty and epistemic attitudes is simpler than existing qualitative semantics used in epistemic game theory (Baltag et al. 2009; Board 1998), as we do not assume a plausibility ordering over possible worlds. The advantage of using such epistemic-minimal models over usual epistemic-plausibility models is that some of the underlying assumptions built in epistemic-plausibility models are not needed to provide epistemic characterizations of solution concepts. In particular, epistemic-minimal models only require that a player is capable of assessing whether an envisaged situation is *possible* or not, without requiring that she is capable of comparing the plausibility of two *possible* situations she envisages and of assessing whether one situation is more plausible than the other.[1] In other words, epistemic-minimal models can be seen as special cases of epistemic-plausibility models in which the epistemic representation is binary (i.e., envisaged worlds are either possible or impossible).

We show that the logic LEG is sufficiently expressive to provide foundation for various game-theoretic solution concepts. The most interesting result will be a charac-

[1] Here we take the term "envisaged" to be synonymous of the term "imagined". Clearly, there are situations that one can imagine that she considers impossible. For example, a person can imagine a situation in which she is the president of French republic and, at the same time, considers this situation impossible.

terization in LEG of the epistemic foundation for the "2-rounds of deletion of weakly dominated strategies, followed by iterated deletion of strongly dominated strategies" (DWDS^2-IDSDS).

As emphasized by Stalnaker (1998), DWDS^2-IDSDS is an interesting solution concept as it is tightly related with the concept of forward induction. The latter has been used to explain some empirical evidences and to justify the behavior of rational players in the context of extensive games such as the well-known "battle of the sexes with outside option" (Cooper et al. 1993). This game has been widely discussed in the game-theoretic literature (see, e.g., Kohlberg and Mertens 1986; Osborne and Rubinstein 1994, Ex. 110.1) and, more specifically, in the area of epistemic game theory (see, e.g., Brandeburger 2007; Battigalli and Siniscalchi 2002). The game goes as follows. Player 1 and player 2 have to play a simultaneous coordination game. Each player has to choose between two available actions A and B. The worst scenario for both players is to miscoordinate (i.e., playing A while the other plays B or vice versa). Furthermore, the players have diverging preferences regarding the best outcome for themselves: player 1 prefers coordination on (A, A) while player 2 prefers coordination on (B, B). Player 1 has a first-move advantage since, prior to playing the coordination game with player 2, she is offered the possibility of a fixed outside option. If she chooses to enter (IN), both players will play the coordination game. If she takes the outside option (OUT), the game will end with an outcome that is for both players better than the outcome resulting from coordinating on the less preferred option, but worse than the outcome resulting from coordinating on the most preferred option. The extensive form of the game is depicted in Fig. 1. The game has two subgame perfect Nash equilibria in pure strategies that can be computed by backward induction, namely $((IN, A), A)$ and $((OUT, B), B)$. Forward induction reasoning provides a refinement of subgame

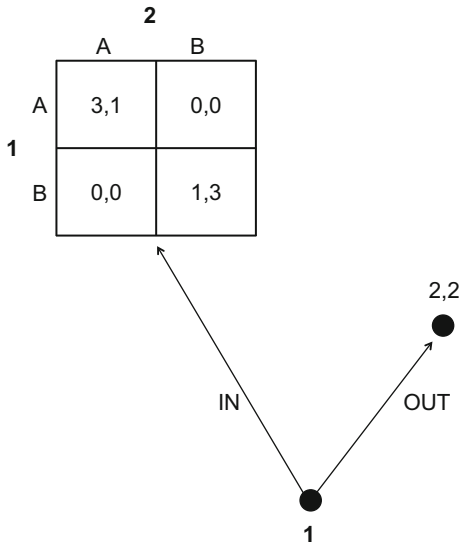


Fig. 1 Battle of the sexes with outside option: extensive form

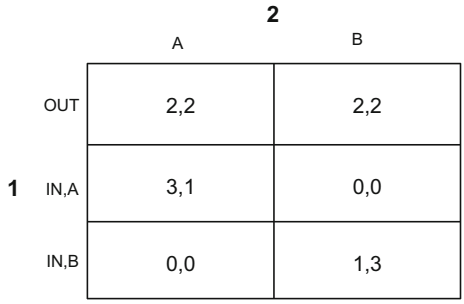


Fig. 2 Battle of the sexes with outside option: strategic form

perfect Nash equilibrium as it allows us to rule out the strategy profile $((OUT, B), B)$ and to keep the strategy profile $((IN, A), A)$. The idea is that, by entering the game, player 1 "signals" to player 2 that she expects to get a payoff higher than 2 and, consequently, that she will play action A (otherwise she would have played OUT). Thus, player 1 can take advantage of this and guarantee to herself the highest possible payoff of 3. Indeed, player 1's signal ensures that player 2 will choose action A in order to maximize his payoff and that the two players will coordinate on the (A, A) option.

It turns out that the forward induction solution $((IN, A), A)$ can be computed by "two rounds of deletion of weakly dominated strategies, followed by one round of deletion of strongly dominated strategies" in the strategic form of the game depicted in Fig. 2. This is a special case of DWDS^2-IDSDS in which the procedure terminates after three rounds of iteration.[2] Indeed, at the first round, strategy (IN, B) of player 1 can be eliminated since it is strongly dominated and thus weakly dominated by strategy OUT. In the resulting subgame strategy B of player 2 can be eliminated since it is weakly dominated by strategy A. Finally, strategy OUT of player 1 can be eliminated since it is strongly dominated by strategy (IN, A). $((IN, A), A)$ is the only strategy profile which survives.

The main motivation for introducing a minimal logic for interactive epistemology is to show that interesting results about the epistemic foundation for solution concepts in game theory can be proved, even if we make very few assumptions about the representation of the players' epistemic states. The rest of the paper is organized as follows. In Sect. 2 we introduce a semantics for LEG based on the concept of epistemic-doxastic game model. Section 3 is devoted to present the formal language of LEG. Section 4 provides a formalization of a qualitative version of the concept of strong belief in LEG. This concept is indeed crucial to characterize the epistemic conditions of DWDS^2-IDSDS. In Sect. 5, LEG is used to provide an epistemic foundation for various game-theoretic solution concepts. Finally, Sect. 6 is devoted to the axiomatization and complexity results for LEG. A selection of proofs of the main results given in the paper are collected in a technical annex at the end of the paper.

[2] Notice that the game can be easily generalized to account for "two rounds of deletion of weakly dominated strategies, followed by n rounds of deletion of strongly dominated strategies".

2 Normal form games and epistemic game models

Let $Atm = \{p, q, \ldots\}$ be a countable set of atomic propositions and let $\Gamma = (Agt, S_1, \ldots, S_n, U_1, \ldots, U_n)$ be a normal form game in the usual sense where:

- $Agt = \{1, \ldots, n\}$ is a finite set of players or agents;
- For every $i \in Agt$, S_i is player i's finite set of strategies whose elements are denoted by $s_i, s_i', \ldots$;
- For every $i \in Agt$, $U_i : \prod_{i \in Agt} S_i \longrightarrow \mathbb{R}$ is player i's utility function assigning a real number (the utility value for i) to every strategy profile.

We let $2^{Agt*} = 2^{Agt} \setminus \{\emptyset\}$ denote the set of all non-empty sets of players (*alias* coalitions). For notational convenience we write $-i$ instead of $Agt \setminus \{i\}$. For every $G \in 2^{Agt*}$, we define the set of strategies for the coalition G to be $S_G = \prod_{i \in G} S_i$. Elements of S_G are denoted by $s_G, s_G', \ldots$ For notational convenience, we write S instead of S_{Agt} and we denote elements of S by $s, s', \ldots$. Moreover, for every coalition $G \in 2^{Agt*}$, for every strategy $s_G \in S_G$ and for every player $i \in G$, we write $s_G[i]$ to denote the position in the vector s_G corresponding to player i.

The following definition introduces the concept of epistemic game model for the game Γ and for the set of atomic propositions Atm.

Definition 1 (*Epistemic game model*) An epistemic-doxastic game model or, simply, epistemic game model (EGM) is a tuple $M = (W, \mathcal{E}_1, \ldots, \mathcal{E}_n, \mathcal{B}_1, \ldots, \mathcal{B}_n, \mathsf{ch}_1, \ldots, \mathsf{ch}_n, \mathcal{V})$ where:

- W is a nonempty set of possible worlds or states;
- For every $i \in Agt$, $\mathcal{E}_i$ is an equivalence relation on W;
- For every $i \in Agt$, $\mathcal{B}_i$ is a serial relation between on W such that:
  - (**C1**) $\mathcal{B}_i \subseteq \mathcal{E}_i$;
  - (**C2**) For all $w, v \in W$: if $w \mathcal{E}_i v$ then $\mathcal{B}_i(w) = \mathcal{B}_i(v)$;
- For every $i \in Agt$, ch_i is a total function $\mathsf{ch}_i : W \longrightarrow S_i$ such that:
  - (**C3**) For all $w, v \in W$ and for all $s_i \in S_i$: if $\mathsf{ch}_i(w) = s_i$ and $w \mathcal{E}_i v$ then $\mathsf{ch}_i(v) = s_i$;
  - (**C4**) For all $w \in W$, for all $i \in Agt$ and for all $s_{-i} \in S_{-i}$: there is $u \in W$ such that $w \mathcal{E}_i u$ and $\mathsf{ch}_{-i}(u) = s_{-i}$.
- $\mathcal{V} : W \longrightarrow 2^{Atm}$ is a valuation function for the set of atomic propositions Atm.

We define $\mathcal{E} = \bigcup_{i \in Agt} \mathcal{E}_i$ and $\mathcal{B} = \bigcup_{i \in Agt} \mathcal{B}_i$. Moreover, we let $\mathcal{E}^+$ and $\mathcal{B}^+$ be the transitive closure of the relation $\mathcal{E}$ and the transitive closure of the relation $\mathcal{B}$, respectively. Such relations are needed to define common knowledge and common belief.

For any binary relation $\mathcal{R}$ on the set of worlds W and for every world $w \in W$, we define $\mathcal{R}(w) = \{v \in W : w \mathcal{R} v\}$.

We call $\mathcal{E}_i(w) = \{v \in W : w \mathcal{E}_i v\}$ player i's *information set* at world w, that is, the set of worlds that player i envisages at world w. Moreover, we call $\mathcal{B}_i(w) = \{v \in W : w \mathcal{B}_i v\}$ player i's *belief set* at world w, that is, the set of worlds that player i thinks to be possible at world w. The function ch_i specifies the strategy chosen by player i at a certain world. We generalize this class of functions to coalitions by introducing a function $\mathsf{ch}_G : W \longrightarrow S_G$ for every $G \in 2^{Agt*}$ and defining it as follows:

for all $w \in W$ and for all $s_G \in S_G$, $\mathsf{ch}_G(w) = s_G$ if and only if $\mathsf{ch}_i(w) = s_G[i]$ for all $i \in G$.

The Condition (**C1**) just says that a player's belief set is a subset of the player's information set. According to the Condition (**C2**), if two worlds are in the same information set of player i, then player i has the same belief set at these two worlds. The Condition (**C3**) is an ex-interim condition according to which an agent i chooses the strategy s_i if and only if she knows this. Finally, according to the Condition (**C4**), for every strategy s_{-i} of the other players a player i envisages a world in which this strategy is played. This corresponds to a well-known assumption in the context of interactive epistemology: the assumption that knowledge of the players is cautious, in the sense that every player must envisage all possible choices of the other players (see, e.g., Mas-Colell et al. 1995, Chap. 8 and also Börgers 1994, Brandenburger 1992, Börgers and Samuelson 1992, Samuelson 1992, Asheim and Dufwenberg 2003, and Asheim 2001, for further discussion about this requirement). Notice that the Condition (**C4**) excludes that if there exists a correlation between a player's choice and the choices of the others, then the player rules out from her information set all choices of the others that are not correlated with her current choice. According to the present model, all choices of the other players that are not correlated with the player's current choice should be included in the player's information set but excluded from her belief set (as they are impossible). For example, suppose that player 1 believes that there is a strict correlation between her choice and the choice of player 2. In particular, she *believes* that if she chooses A then player 2 will not choose B. Now, player 1 chooses A. Hence, because of the ex-interim Condition (**C3**), player 1 knows this. Consequently, the worlds in which player 2 chooses B should be excluded from player 1's belief set. However, the cautiousness Condition (**C4**) guarantees that they are included in player 1's information set.

3 Logic of epistemic attitudes in games: formal language

The language of the logic of epistemic attitudes in games (LEG) is parametrized by the game Γ and the set of atomic propositions Atm. It is denoted by $\mathcal{L}_{\mathsf{LEG}}(\Gamma, Atm)$ and defined by the following grammar:

$$\varphi ::= p \mid choose_i(s_i) \mid poss_i \mid \neg\varphi \mid \varphi \wedge \psi \mid \mathsf{K}_i\varphi \mid \mathsf{B}_i\varphi \mid \mathsf{CK}\varphi \mid \mathsf{CB}\varphi$$

where p ranges over Atm, i ranges over Agt and s_i ranges over S_i.

The formulas of the language $\mathcal{L}_{\mathsf{LEG}}(\Gamma, Atm)$ have the following meaning:

- $choose_i(s_i)$: player i chooses strategy s_i;
- $poss_i$: the current world is considered possible by player i;
- $\mathsf{K}_i\varphi$: player i knows that φ is true;
- $\mathsf{B}_i\varphi$: player i believes that φ is true;
- $\mathsf{CK}\varphi$: the players have common knowledge that φ;
- $\mathsf{CB}\varphi$: the players have common belief that φ.

For notational convenience, we abbreviate $\widehat{\mathsf{K}}_i\varphi \stackrel{\mathtt{def}}{=} \neg\mathsf{K}_i\neg\varphi$ and $\widehat{\mathsf{B}}_i\varphi \stackrel{\mathtt{def}}{=} \neg\mathsf{B}_i\neg\varphi$. Furthermore, we define:

$$\mathsf{EK}\varphi \stackrel{\mathtt{def}}{=} \bigwedge_{i\in Agt} \mathsf{K}_i\varphi$$

$$\mathsf{EB}\varphi \stackrel{\mathtt{def}}{=} \bigwedge_{i\in Agt} \mathsf{B}_i\varphi$$

$$choose(s) \stackrel{\mathtt{def}}{=} \bigwedge_{i\in Agt} choose_i(s[i])$$

where $\mathsf{EK}\varphi$, $\mathsf{EB}\varphi$ and $choose(s)$ have to be read respectively "everybody knows that φ is true", "everybody believes that φ is true" and "the players choose the strategy profile s".

The following definition provides the truth conditions of formulas in $\mathcal{L}_{\mathsf{LEG}}(\Gamma, Atm)$ relative to the class of epistemic game models. Notice in particular the truth condition of formula $poss_i$ according to which $poss_i$ is true at certain world w if and only if w is an element of player i's belief set at w.

Definition 2 (*Truth conditions*) Let M be an epistemic game model and let w be a world in M. Then:

$$\begin{array}{ll}
M, w \models p & \text{iff } p \in \mathcal{V}(w) \\
M, w \models poss_i & \text{iff } w \in \mathcal{B}_i(w) \\
M, w \models choose_i(s_i) & \text{iff } \mathsf{ch}_i(w) = s_i \\
M, w \models \neg\varphi & \text{iff } M, w \not\models \varphi \\
M, w \models \varphi \wedge \psi & \text{iff } M, w \models \varphi \text{ and } M, w \models \psi \\
M, w \models \mathsf{K}_i\varphi & \text{iff } M, v \models \varphi \text{ for all } v \in \mathcal{E}_i(w) \\
M, w \models \mathsf{B}_i\varphi & \text{iff } M, v \models \varphi \text{ for all } v \in \mathcal{B}_i(w) \\
M, w \models \mathsf{CK}\varphi & \text{iff } M, v \models \varphi \text{ for all } v \in \mathcal{E}^+(w) \\
M, w \models \mathsf{CB}\varphi & \text{iff } M, v \models \varphi \text{ for all } v \in \mathcal{B}^+(w)
\end{array}$$

We write $\models \varphi$ to mean that the LEG-formula φ is *valid* relative to the class of epistemic game models. That is, $\models \varphi$ iff, for every EGM M and for every w in M, we have $M, w \models \varphi$. We say that φ is *satisfiable* relative to the class of epistemic game models iff, $\neg\varphi$ is not valid relative to the class of epistemic game models.

Moreover, for every epistemic game model $= (W, \mathcal{E}_1, \ldots, \mathcal{E}_n, \mathcal{B}_1, \ldots, \mathcal{B}_n, \mathsf{ch}_1, \ldots, \mathsf{ch}_n, \mathcal{V})$ and LEG formula φ, $||\varphi||_M = \{w \in W : M, w \models \varphi\}$ denotes the truth set of φ relative to M. For notational convenience, when the context is clear we omit the subscript and write $||\varphi||$ instead of $||\varphi||_M$.

4 The concept of strong belief

The following abbreviation provides a purely qualitative characterization of the concept of strong belief:

$$\mathsf{SB}_i\varphi \stackrel{\text{def}}{=} \mathsf{K}_i\varphi \vee (\mathsf{K}_i(poss_i \rightarrow \varphi) \wedge \mathsf{K}_i(\neg poss_i \rightarrow \neg\varphi))$$

where $\mathsf{SB}_i\varphi$ has to be read "player i strongly believes φ". Quantitative and semi-qualitative characterizations of this concept in a probabilistic framework and in a semantics for epistemic logic with plausibility orderings over states can be found in Battigalli and Siniscalchi (2002) and Baltag and Smets (2009), respectively. The general intuition is that player i has a strong belief that φ if and only if the worlds that i envisages in which φ is true are strictly more possible than the states that i envisages in which φ is false.

As the following proposition highlights, the previous abbreviation correctly characterizes the concept of strong belief:

Proposition 1 *Let* $M = (W, \mathcal{E}_1, \ldots, \mathcal{E}_n, \mathcal{B}_1, \ldots, \mathcal{B}_n, \mathsf{ch}_1, \ldots, \mathsf{ch}_n, \mathcal{V})$ *be an epistemic game model and let* $||\varphi||_{w,i} = \{v \in W : v \in ||\varphi|| \text{ and } w\mathcal{E}_i v\}$ *be player i's set of epistemic alternatives at world w in which* φ *is true. Then,* $M, w \models \mathsf{SB}_i\varphi$ *iff:*

- $||\varphi||_{w,i} \neq \emptyset$, *and*
- *For all* $v \in ||\varphi||_{w,i}$ *and for all* $u \in ||\neg\varphi||_{w,i}$, $\pi_i(u) < \pi_i(v)$;

where $\pi_i : W \longrightarrow \{0, 1\}$ *such that for all* $v \in W$*:*

- $\pi_i(v) = 0$ *iff* $v \notin \mathcal{B}_i(v)$, *and*
- $\pi_i(v) = 1$ *iff* $v \in \mathcal{B}_i(v)$.

Following possibility theory (Dubois and Prade 1998), we consider two ways of lifting the possibility degree of a world to the possibility of a formula viewed as a set of worlds. The first possibility measure introduced in the context of possibility theory is the so-called weak possibility measure (or potential possibility). The weak possibility of a formula φ corresponds to the maximal degree of plausibility of an envisaged world in which φ is true. More formally:

Definition 3 (*Weak possibility measure (or potential possibility)*) Let $M = (W, \mathcal{E}_1, \ldots, \mathcal{E}_n, \mathcal{B}_1, \ldots, \mathcal{B}_n, \mathsf{ch}_1, \ldots, \mathsf{ch}_n, \mathcal{V})$ be an epistemic game model and let $||\varphi||_{w,i} = \{v \in W : v \in ||\varphi|| \text{ and } w\mathcal{E}_i v\}$. The weak possibility of formula φ for agent i at world w, denoted by $\Pi_{w,i}(\varphi)$, is defined as follows:

$$\Pi_{w,i}(\varphi) = \begin{cases} \max\limits_{v \in ||\varphi||_{w,i}} \pi_i(v) & \text{if } \ ||\varphi||_{w,i} \neq \emptyset \\ 0 & \text{if } \ ||\varphi||_{w,i} = \emptyset \end{cases}$$

where the function π_i is defined as in Proposition 1.

The second possibility measure introduced in the context of possibility theory is the so-called strong possibility measure (or guaranteed possibility). The strong possibility of a formula φ corresponds to the minimal degree of plausibility of an envisaged world in which φ is true. More formally:

Definition 4 (*Strong possibility measure (or guaranteed possibility)*) Let $M = (W, \mathcal{E}_1, \ldots, \mathcal{E}_n, \mathcal{B}_1, \ldots, \mathcal{B}_n, \mathsf{ch}_1, \ldots, \mathsf{ch}_n, \mathcal{V})$ be an epistemic game model and let

$||\varphi||_{w,i} = \{v \in W : v \in ||\varphi|| \text{ and } w\mathcal{E}_i v\}$. The strong possibility of formula φ for agent i at world w, denoted by $\Delta_{w,i}(\varphi)$, is defined as follows:

$$\Delta_{w,i}(\varphi) = \begin{cases} \min\limits_{v \in ||\varphi||_{w,i}} \pi_i(v) & \text{if} \quad ||\varphi||_{w,i} \neq \emptyset \\ 1 & \text{if} \quad ||\varphi||_{w,i} = \emptyset \end{cases}$$

where the function π_i is defined as in Proposition 1.

Two basic decomposability properties of the weak possibility measure and the strong possibility measures are given in the following proposition.

Proposition 2 *Let M be an epistemic game model, let w be a world in M and let $i \in Agt$. Then,*

$$\Pi_{w,i}(\varphi \vee \psi) = \max(\Pi_{w,i}(\varphi), \Pi_{w,i}(\psi))$$
$$\Delta_{w,i}(\varphi \vee \psi) = \min(\Delta_{w,i}(\varphi), \Delta_{w,i}(\psi))$$

For instance, in order to prove the second property, we have to distinguish two cases. Suppose $||\varphi \vee \psi||_{w,i} \neq \emptyset$. Then:

$$\begin{aligned} \Delta_{w,i}(\varphi \vee \psi) &= \min_{v \in ||\varphi \vee \psi||_{w,i}} \pi_i(v) \\ &= \min_{v \in ||\varphi||_{w,i} \cup ||\psi||_{w,i}} \pi_i(v) \\ &= \min(\min_{v \in ||\varphi||_{w,i}} \pi_i(v), \min_{v \in ||||\psi||_{w,i}} \pi_i(v)) \\ &= \min(\Delta_{w,i}(\varphi), \Delta_{w,i}(\psi)) \end{aligned}$$

Suppose $||\varphi \vee \psi||_{w,i} = \emptyset$. Then $||\varphi||_{w,i} = \emptyset$ and $||\psi||_{w,i} = \emptyset$. Thus, $\Delta_{w,i}(\varphi \vee \psi) = \min(\Delta_{w,i}(\varphi), \Delta_{w,i}(\psi)) = 1$.

The following proposition highlights the connection between the concept of strong belief and preceding concepts of weak possibility and strong possibility. Specifically, a certain agent i strongly believes φ if and only if, i knows that φ, or the strong possibility of φ for i is equal to 1 and the weak possibility of $\neg\varphi$ for i is equal to 0.

Proposition 3 *Let M be an epistemic game model, let w be a world in M and let $i \in Agt$. Then, $M, w \models \mathsf{SB}_i\varphi$ iff: (i) $||\neg\varphi||_{w,i} = \emptyset$ or, (ii) $\Delta_{w,i}(\varphi) = 1$ and $\Pi_{w,i}(\neg\varphi) = 0$.*

5 Epistemic game theory in LEG

In this section we use the logic LEG to characterize the epistemic conditions of some well-know solution concepts in game theory: the procedure of iterated deletion of strongly dominated strategies (IDSDS procedure), the procedure of 1-round of deletion of weakly dominated strategies followed by iterated deletion of strongly

dominated strategies (DWDS^1-IDSDS procedure, and the procedure of 2-rounds of deletion of weakly dominated strategies followed by iterated deletion of strongly dominated strategies (DWDS^2-IDSDS procedure). The concepts of strong dominance and weak dominance we consider are relative to pure strategies.

The latter solution concepts are introduced in Sect. 5.1. In Sect. 5.2 the logic LEG is used to formally characterize three different forms of rationality which have been discussed in the field of epistemic game theory: weak rationality, strong rationality and perfect rationality. In Sect. 5.3 three theorems about the epistemic and rationality conditions of IDSDS, DWDS^1-IDSDS and DWDS^2-IDSDS will be given.

5.1 Solution concepts

The following concept of subgame is needed to formally characterize the solution concepts.

Definition 5 (*Subgame*) Let $\Gamma = (Agt, S_1, \ldots, S_n, U_1, \ldots, U_n)$ and $\Gamma' = (Agt', S'_1, \ldots, S'_n, U'_1, \ldots, U'_n)$ be normal form games. Then, Γ' is a subgame of Γ if and only if:

- $Agt = Agt'$;
- For every $i \in Agt$, $S'_i \subseteq S_i$;
- For every $i \in Agt$, $U'_i = U_i|_{\prod_{i \in Agt} S'_i}$ where U'_i is the restriction of U_i to the set of strategy profiles $\prod_{i \in Agt} S'_i$.

A strategy s_i of a given player i is a strongly dominated strategy if and only if, there exists another strategy s'_i of i such that, for all strategies s_{-i} of the other players, playing s'_i while the others play s_{-i} is for i better than playing s_i while the others play s_{-i}. An example of strongly dominated strategy is cooperation in the Prisoner Dilemma (PD) game: whether the opponent chooses to cooperate or defect, defection yields a higher payoff than cooperation. More formally:

Definition 6 (*Strongly dominated strategies*) Let $\Gamma = (Agt, S_1, \ldots, S_n, U_1, \ldots, U_n)$ be a normal form game. The set

$$\mathsf{SD}_i^{\Gamma} = \{s_i \in S_i : \exists s'_i \in S_i \text{ s.t. } \forall s_{-i} \in S_{-i}, U_i(s_i, s_{-i}) < U_i(s'_i, s_{-i})\}$$

is the set of strategies of player i that are strongly dominated in Γ.

A strategy s_i of a given player i is a weakly dominated strategy if and only if, there exists another strategy s'_i of i such that, for all strategies s_{-i} of the others, playing s'_i while the others play s_{-i} is for i at least as good as playing s_i while the others play s_{-i} and there is at least one strategy s'_{-i} of the others such that playing s'_i while the others play s'_{-i} is for i better than playing s_i while the others play s'_{-i}. More formally:

Definition 7 (*Weakly dominated strategies*) Let $\Gamma = (Agt, S_1, \ldots, S_n, U_1, \ldots, U_n)$ be a normal form game. The set

$$\begin{aligned}\mathsf{WD}_i^{\Gamma} = \{s_i \in S_i : \exists s'_i \in S_i \text{ s.t. } \forall s_{-i} \in S_{-i}, U_i(s_i, s_{-i}) \leq U_i(s'_i, s_{-i}) \quad \text{and} \\ \exists s'_{-i} \in S_{-i} \text{ s.t. } U_i(s_i, s_{-i}) < U_i(s'_i, s_{-i})\}\end{aligned}$$

is the set of strategies of player i that are weakly dominated in Γ.

The so-called iterated deletion of strongly dominated strategies (IDSDS) (or iterated strong dominance) is a procedure that starts with the original game Γ and, at each step, for every player i removes from the game all i's strongly dominated strategies, thereby generating a subgame of the original game, and that repeats this process again and again. The IDSDS procedure can be inductively defined as follows.

Definition 8 (*IDSDS Procedure*) Let $\Gamma = (Agt, S_1, \ldots, S_n, U_1, \ldots, U_n)$ be a normal form game. Iterated deletion of strongly dominated strategies (IDSDS) is the procedure defined recursively as follows. For all $i \in Agt$:

- Let $S_{i,0}^{IDSDS} = S_i$ and $\Gamma^0 = \Gamma$,
- For $m \geq 1$, let $S_{i,m}^{IDSDS} = S_{i,m-1}^{IDSDS} \setminus \mathsf{SD}_i^{\Gamma^{m-1}}$, where Γ^{m-1} is the subgame of Γ with strategy sets $S_{i,m-1}^{IDSDS}$.

For every $m \geq 0$ and $G \in 2^{Agt*}$, let $S_{G,m}^{IDSDS} = \prod_{i \in G} S_{i,m}^{IDSDS}$ and let $S_m^{IDSDS} = S_{Agt,m}^{IDSDS}$.

Finally, let $S_i^{IDSDS} = \bigcap_{m \in \mathbb{N}} S_{i,m}^{IDSDS}$. For every $G \in 2^{Agt*}$, let $S_G^{IDSDS} = \prod_{i \in G} S_i^{IDSDS}$ and $S^{IDSDS} = S_{Agt}^{IDSDS}$.

The procedure of n-rounds of deletion of weakly dominated strategies followed by iterated deletion of strongly dominated strategies (DWDS^n-IDSDS) is a procedure that starts with the original game Γ and, at each step, for every player i removes from the game all i's weakly dominated strategies thereby generating a subgame of the original game, and that repeats this process for n rounds. Then, after n rounds it applies iterated deletion of strongly dominated strategies starting with the game Γ^n. The DWDS^n-IDSDS procedure can be inductively defined as follows.

Definition 9 (DWDS^n-IDSDS *Procedure*) Let $\Gamma = (Agt, S_1, \ldots, S_n, U_1, \ldots, U_n)$ be a normal form game. n-iteration of deletion of weakly dominated strategies (DWDS^n) followed by iterated deletion of strongly dominated strategies (IDSDS) is the procedure defined recursively as follows. For all $i \in Agt$:

- Let $S_{i,0}^{DWDS^n-IDSDS} = S_i$ and $\Gamma^0 = \Gamma$,
- For $1 \leq m \leq n$, let $S_{i,m}^{DWDS^n-IDSDS} = S_{i,m-1}^{DWDS^n-IDSDS} \setminus \mathsf{WD}_i^{\Gamma^{m-1}}$,
- For $m > n$, let $S_{i,m}^{DWDS^n-IDSDS} = S_{i,m-1}^{DWDS^n-IDSDS} \setminus \mathsf{SD}_i^{\Gamma^{m-1}}$, where Γ^{m-1} is the subgame of Γ with strategy sets $S_{i,m-1}^{DWDS^n-IDSDS}$.

For every m such that $0 \leq m \leq n$ and $G \in 2^{Agt*}$, let $S_{G,m}^{DWDS^n-IDSDS} = \prod_{i \in G} S_{i,m}^{DWDS^n-IDSDS}$ and $S_m^{DWDS^n-IDSDS} = S_{Agt,m}^{DWDS^n-IDSDS}$.

Finally, let $S_i^{DWDS^n-IDSDS} = \bigcap_{m \in \mathbb{N}} S_{i,m}^{DWDS^n-IDSDS}$. For every $G \in 2^{Agt*}$, let $S_G^{DWDS^n-IDSDS} = \prod_{i \in G} S_i^{DWDS^n-IDSDS}$ and $S^{DWDS^n-IDSDS} = S_{Agt}^{DWDS^n-IDSDS}$.

Note that, if $n = 0$, DWDS^n-IDSDS is nothing but IDSDS. Moreover, if $n = \infty$, DWDS^n-IDSDS corresponds to the procedure of iterated deletion of weakly dominated strategies.

The previous procedures IDSDS and DWDS^n-IDSDS are well-known concepts in game theory. In this paper we focus on the special cases of DWDS^n-IDSDS with $n = 1$ and $n = 2$. For instance, when $n = 1$, the procedure DWDS^n-IDSDS corresponds to the variant of the so-called *Dekel-Fudenberg procedure* (Dekel and Fudenberg 1990) for pure strategies.[3] A Bayesian justification of the concept of weak dominance given in Definition 7 can be found in Börgers (1993). Börgers shows that, if the only part of players' preferences that is taken as exogenously given is their (ordinal) preferences over pure strategy profiles, then a pure strategy s_i of a certain player i is not weakly dominated by another pure strategy given a certain subset $S'_{-i} \subseteq S_{-i}$ of the set of strategies of the other players if and only if, there exists a subjective probabilistic belief with full support on S'_{-i} of the player and some von Neumann and Morgenstern-utility function that agrees with the player's ordinal preferences such that the strategy s_i maximizes expected utility given S'_{-i}. (This can be seen as a pure strategy-variant of the classical result about rationalizability by Pearce 1984.)

5.2 Types of rationality

The following three sections are devoted to define three rationality criteria that will be used in Sect. 5.3 to give an epistemic foundation of the concepts of strong dominance, weak dominance, IDSDS, DWDS^1-IDSDS and DWDS^2-IDSDS.

5.2.1 Weak rationality

The first concept of rationality we consider is weak rationality. Player i is said to be *weakly rational* in choosing s_i, denoted by $\mathsf{WRat}_i(s_i)$, if and only if action s_i is not *strongly* dominated within her set of doxastic alternatives. Formally:

$$\mathsf{WRat}_i(s_i) \stackrel{\mathtt{def}}{=} \bigwedge_{s'_i \neq s_i} \left(\bigvee_{s_{-i} \in S_{-i} : U_i(s'_i, s_{-i}) \leq U_i(s_i, s_{-i})} \widehat{\mathsf{B}}_i \mathit{choose}_{-i}(s_{-i}) \right)$$

We moreover define the fact that i is weakly rational, denoted by WRat_i:

$$\mathsf{WRat}_i \stackrel{\mathtt{def}}{=} \bigvee_{s_i \in S_i} (\mathit{choose}_i(s_i) \wedge \bigwedge_{s'_i \neq s_i} \left(\bigvee_{s_{-i} \in S_{-i} : U_i(s'_i, s_{-i}) \leq U_i(s_i, s_{-i})} \widehat{\mathsf{B}}_i \mathit{choose}_{-i}(s_{-i})) \right)$$

5.2.2 Strong rationality

Let us now move to the concept strong rationality. Player i is said to be *strongly rational* in choosing s_i, denoted by $\mathsf{SRat}_i(s_i)$, if and only if action s_i is not *weakly*

[3] Indeed the concept of weak dominance used by Dekel and Fudenberg is relative to mixed strategies in the sense that, a given pure strategy s_i of player i is weakly dominated if and only if, there exists a mixed strategy $\sigma_i \in \Delta(S_i)$ of player i such that $U_i(s_i, s'_{-i}) \leq U_i(\sigma_i, s'_{-i})$ for all $s'_{-i} \in S_{-i}$ and $U_i(s_i, s'_{-i}) < U_i(\sigma_i, s'_{-i})$ for some $s'_{-i} \in S_{-i}$, where $\Delta(S_i)$ is the set of all probability measures over S_i.

dominated within her set of doxastic alternatives. Formally:

$$\mathsf{SRat}_i(s_i) \stackrel{\mathtt{def}}{=} \bigwedge_{s_i' \neq s_i} \left(\left(\bigvee_{s_{-i} \in S_{-i} : U_i(s_i', s_{-i}) < U_i(s_i, s_{-i})} \widehat{\mathsf{B}}_i \mathit{choose}_{-i}(s_{-i}) \right) \vee \left(\bigwedge_{s_{-i} \in S_{-i} : U_i(s_i, s_{-i}) < U_i(s_i', s_{-i})} \mathsf{B}_i \neg \mathit{choose}_{-i}(s_{-i}) \right) \right)$$

We moreover define the fact that i is strongly rational, denoted by SRat_i:

$$\mathsf{SRat}_i \stackrel{\mathtt{def}}{=} \bigvee_{s_i \in S_i} (\mathit{choose}_i(s_i) \wedge \mathsf{SRat}_i(s_i))$$

5.2.3 Perfect rationality

Finally, we consider the concept perfect rationality. The concept defined here is a qualitative version of Stalnaker's lexicographic concept of perfect rationality (Stalnaker 1996; Stalnaker 1998). Player i is said to be *perfectly rational* in choosing s_i, denoted by $\mathsf{PRat}_i(s_i)$, if and only if:

- s_i is not weakly dominated within i's belief set (i.e., i is strongly rational in choosing s_i); and
- s_i is not weakly dominated within i's information set, after having discarded all weakly dominated strategies within i's belief set (i.e., after having discarded all strategies of i that are not strongly rational choices).

Formally:

$$\mathsf{PRat}_i(s_i) \stackrel{\mathtt{def}}{=} \mathsf{SRat}_i(s_i) \wedge \bigwedge_{s_i' \neq s_i} (\mathsf{SRat}_i(s_i') \rightarrow \left(\left(\bigvee_{s_{-i} \in S_{-i} : U_i(s_i', s_{-i}) < U_i(s_i, s_{-i})} \widehat{\mathsf{K}}_i \mathit{choose}_{-i}(s_{-i}) \right) \vee \left(\bigwedge_{s_{-i} \in S_{-i} : U_i(s_i, s_{-i}) < U_i(s_i', s_{-i})} \mathsf{K}_i \neg \mathit{choose}_{-i}(s_{-i}) \right) \right)$$

We moreover define the fact that i is perfectly rational, denoted by PRat_i:

$$\mathsf{PRat}_i \stackrel{\mathtt{def}}{=} \bigvee_{s_i \in S_i} (\mathit{choose}_i(s_i) \wedge \mathsf{PRat}_i(s_i))$$

5.3 Epistemic foundation

Our first result is an epistemic characterization of the concepts of strong dominance and weak dominance as defined in Definitions 6 and 7.

Theorem 1 *Let $i \in Agt$. Then:*

1. $s_i \notin \mathsf{SD}_i^{\Gamma}$ *if and only if there exists an epistemic game model* $M = (W, \mathcal{E}_1, \ldots, \mathcal{E}_n, \mathcal{B}_1, \ldots, \mathcal{B}_n, \mathsf{ch}_1, \ldots, \mathsf{ch}_n, \mathcal{V})$ *and a world* $w \in W$ *such that* $M, w \models \mathsf{WRat}_i(s_i)$*;*
2. $s_i \notin \mathsf{WD}_i^{\Gamma}$ *if and only if there exists an epistemic game model* $M = (W, \mathcal{E}_1, \ldots, \mathcal{E}_n, \mathcal{B}_1, \ldots, \mathcal{B}_n, \mathsf{ch}_1, \ldots, \mathsf{ch}_n, \mathcal{V})$ *and a world* $w \in W$ *such that* $M, w \models \mathsf{PRat}_i(s_i)$.

It is worth noting that the second item of the previous Theorem 1 no longer holds if we replace the concept of perfect rationality, expressed by the formula $\mathsf{PRat}_i(s_i)$, by the concept of strong rationality, expressed by the formula $\mathsf{SRat}_i(s_i)$. To show this, it is sufficient to consider the following epistemic game model $M = (W, \mathcal{E}_1, \mathcal{E}_2, \mathcal{B}_1, \mathcal{B}_2, \mathsf{ch}_1, \mathsf{ch}_2, \mathcal{V})$ for the two-player normal form game with strategy set $S = \{(A, A), (A, B), (B, A), (B, B)\}$ depicted in Fig. 3:

- $W = \{w_{(A,A)}, w_{(A,B)}, w_{(B,A)}, w_{(B,B)}, \}$;
- $\mathcal{E}_1(w_{(A,A)}) = \{w_{(A,A)}, w_{(A,B)}\}$, $\mathcal{E}_1(w_{(A,B)}) = \{w_{(A,A)}, w_{(A,B)}\}$, $\mathcal{E}_1(w_{(B,A)}) = \{w_{(B,A)}, w_{(B,B)}\}$, $\mathcal{E}_1(w_{(B,B)}) = \{w_{(B,A)}, w_{(B,B)}\}$;
- $\mathcal{E}_2(w_{(A,A)}) = \{w_{(A,A)}, w_{(B,A)}\}$, $\mathcal{E}_2(w_{(A,B)}) = \{w_{(A,B)}, w_{(B,B)}\}$, $\mathcal{E}_2(w_{(B,A)}) = \{w_{(A,A)}, w_{(B,A)}\}$, $\mathcal{E}_2(w_{(B,B)}) = \{w_{(A,B)}, w_{(B,B)}\}$;
- $\mathcal{B}_1(w_{(A,A)}) = \{w_{(A,A)}\}$, $\mathcal{B}_1(w_{(A,B)}) = \{w_{(A,B)}\}$, $\mathcal{B}_1(w_{(B,A)}) = \{w_{(B,A)}\}$, $\mathcal{B}_1(w_{(B,B)}) = \{w_{(B,B)}\}$;
- $\mathcal{B}_2(w_{(A,A)}) = \{w_{(A,A)}\}$, $\mathcal{B}_2(w_{(A,B)}) = \{w_{(A,B)}\}$, $\mathcal{B}_2(w_{(B,A)}) = \{w_{(B,A)}\}$, $\mathcal{B}_2(w_{(B,B)}) = \{w_{(B,B)}\}$;
- $\mathsf{ch}_1(w_{(A,A)}) = \mathsf{ch}_1(w_{(A,B)}) = A$ and $\mathsf{ch}_1(w_{(B,A)}) = \mathsf{ch}_1(w_{(B,B)}) = B$;
- $\mathsf{ch}_2(w_{(A,A)}) = \mathsf{ch}_2(w_{(B,A)}) = A$ and $\mathsf{ch}_2(w_{(A,B)}) = \mathsf{ch}_2(w_{(B,B)}) = B$;
- $\mathcal{V}(w_{(A,A)}) = \mathcal{V}(w_{(A,B)}) = \mathcal{V}(w_{(B,A)}) = \mathcal{V}(w_{(B,B)}) = Atm$.

It is easy to check that player 1's strategy A is weakly dominated by strategy B Thus, $A \in \mathsf{WD}_1^{\Gamma}$. But $M, w_{(A,A)} \models \mathsf{SRat}_1(A)$. Intuitively speaking, the reason why

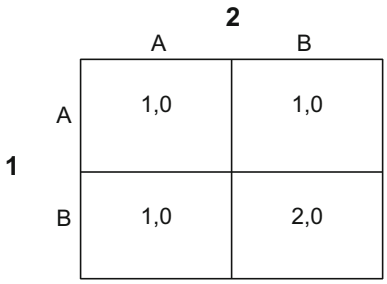


Fig. 3 Counterexample

the concept of strong rationality is not sufficient to characterize weak dominance is that the latter requires that, when deciding what to do, a player should take into account all strategies of the other players. While this is an important feature of the concept of perfect rationality, it is not contemplated by the concept of strong rationality.

The following Theorem 2 is a well-known result in epistemic game theory (see, e.g., Stalnaker 1994; Bonanno 2008; Board 2002): if the players have common belief that every player is weakly rational, then the strategy profile which is played must survive Iterated Deletion of Strongly Dominated Strategies (IDSDS). A similar result, where domination means domination by mixed strategies, has also been shown by Tan and Werlang (1988) and Brandenburger and Dekel (1987).

Theorem 2 *Let* $\mathsf{AllWRat}_G \stackrel{\mathtt{def}}{=} \bigwedge_{i \in G} \mathsf{WRat}_i$ *for any* $G \in 2^{Agt*}$. *Then:*

$$\models \mathsf{CB}\ \mathsf{AllWRat}_{Agt} \rightarrow \bigvee_{s \in S^{IDSDS}} choose(s)$$

According to the following Theorem 3, if the players have common belief that every agent is perfectly rational, then the strategy profile which is played must survive one iteration of DWDS followed by IDSDS. This theorem is also proved by Lorini (2013).

Theorem 3 *Let* $\mathsf{AllPRat}_G \stackrel{\mathtt{def}}{=} \bigwedge_{i \in G} \mathsf{PRat}_i$ *for any* $G \in 2^{Agt*}$. *Then:*

$$\models \mathsf{CB}\ \mathsf{AllPRat}_{Agt} \rightarrow \bigvee_{s \in S^{DWDS^1-IDSDS}} choose(s)$$

An alternative characterization of the epistemic conditions of DWDS^1-IDSDS in a probabilistic setting has been given by Stalnaker (1996). Brandenburger (1992) and Börgers (1994) provide characterizations of the so-called Dekel-Fudenberg procedure, the variant of DWDS^1-IDSDS which uses mixed strategies instead of pure strategies.

The last theorem of this section is about the epistemic foundation for DWDS^2-IDSDS. Under the assumption that the players have common belief that for every player i and for every non-weakly dominated strategy of the other players there must be a world envisaged by i in which this strategy is played and the other players are perfectly rational (the so-called 'completeness assumption' Brandenburger et al. 2008; Brandeburger 2007), the common belief that every player is perfectly rational and has a robust belief about the perfect rationality of the other players is a sufficient condition for DWDS^2-IDSDS. As far as we know, this theorem has never been proved before in a purely qualitative setting.

Theorem 4 *Let*

$$\mathsf{ComplAss} \stackrel{\mathtt{def}}{=} \bigwedge_{i \in Agt} \bigwedge_{s_{-i} \in S_{-i} : s_{-i} \in S_{-i,1}^{DWDS^2-IDSDS}} \widehat{\mathsf{K}}_i(choose_{-i}(s_{-i}) \wedge \mathsf{AllPRat}_{-i}).$$

Then:

$$\models \mathsf{CB}(\mathsf{ComplAss} \wedge \mathsf{AllPRat}_{Agt} \wedge \bigwedge_{i \in Agt} \mathsf{SB}_i\ \mathsf{AllPRat}_{-i}) \rightarrow \bigvee_{s \in S^{DWDS^2-IDSDS}} choose(s)$$

Notice that the completeness assumption ComplAss just guarantees that the epistemic game model is large enough to include, for every non-weakly dominated strategy of a player, a world in which this strategy is 'rationally' chosen by the player.

We conclude this section with the following Theorem 5 stating the converse of the preceding Theorems 2, 3 and 4. Specifically, the theorem states that:

- For every strategy profile that survives IDSDS we can find an epistemic game model in which this strategy profile is played and the players have common belief that every player is weakly rational;
- For every strategy profile that survives DWDS^1-IDSDS we can find an epistemic game model in which this strategy profile is played and the players have common belief that every player is strongly rational;
- For every strategy profile that survives DWDS^2-IDSDS we can find an epistemic game model in which this strategy profile is played and the players have common belief that the completeness assumption ComplAss holds and that every player is perfectly rational and has a robust belief about the perfect rationality of the other players.

Thus, together with Theorems 2, 3 and 4, the following Theorem 5 provides a characterization of the epistemic conditions of IDSDS, DWDS^1-IDSDS and DWDS^2-IDSDS.

Theorem 5 *Let $s \in S$. Then:*

1. *If $s \in S^{IDSDS}$ then there exists an epistemic game model $M = (W, \mathcal{E}_1, \ldots, \mathcal{E}_n, \mathcal{B}_1, \ldots, \mathcal{B}_n, \mathsf{ch}_1, \ldots, \mathsf{ch}_n, \mathcal{V})$ and a world $w \in W$ such that $M, w \models choose(s) \wedge \mathsf{CB}\ \mathsf{AllWRat}_{Agt}$;*
2. *If $s \in S^{DWDS^1-IDSDS}$ then there exists an epistemic game model $M = (W, \mathcal{E}_1, \ldots, \mathcal{E}_n, \mathcal{B}_1, \ldots, \mathcal{B}_n, \mathsf{ch}_1, \ldots, \mathsf{ch}_n, \mathcal{V})$ and a world $w \in W$ such that $M, w \models choose(s) \wedge \mathsf{CB}\ \mathsf{AllPRat}_{Agt}$;*
3. *If $s \in S^{DWDS^2-IDSDS}$ then there exists an epistemic game model $M = (W, \mathcal{E}_1, \ldots, \mathcal{E}_n, \mathcal{B}_1, \ldots, \mathcal{B}_n, \mathsf{ch}_1, \ldots, \mathsf{ch}_n, \mathcal{V})$ and a world $w \in W$ such that $M, w \models choose(s) \wedge \mathsf{CB}(\mathsf{ComplAss} \wedge \mathsf{AllPRat}_{Agt} \wedge \bigwedge_{i \in Agt} \mathsf{SB}_i\ \mathsf{AllPRat}_{-i})$.*

The next section is devoted to study the mathematical and computational properties of the logic LEG.

6 Axiomatization and complexity results

In this section we provide a sound and complete axiomatization for the logic LEG relative to the class of epistemic game models as well as a complexity result for the satisfiability problem.

Theorem 6 *The set of validities of the logic* LEG *relative to the class of epistemic game models is completely axiomatized by the principles given in Fig. 4.*

As Fig. 4 highlights, the axiomatization of the logic LEG is organized in four categories of axioms *plus* the tautologies of propositional logic, the usual rule of inference *modus ponens* and the rule of necessitation for the common knowledge operator. Notice that the rules of necessitation for common belief, knowledge and belief do not need to be added, as they are deducible from the rule of necessitation for common knowledge, Axioms 2(f), 4(a) and 4(c). The axiomatization includes the usual axioms for knowledge and common knowledge, namely the S5 principles for the knowledge operators, the fixed-point axiom and the induction axiom for common knowledge. Similarly, it includes the usual axioms for belief and common belief, namely the KD principles for the belief operators, the fixed-point axiom and the induction axiom for common belief. Notice that Axioms 4 and 5 for belief, namely $\mathsf{B}_i\varphi \rightarrow \mathsf{B}_i\mathsf{B}_i\varphi$ and $\neg\mathsf{B}_i\varphi \rightarrow \mathsf{B}_i\neg\mathsf{B}_i\varphi$, do not need to be added as they are deducible from Axioms 2(a), 2(b), 2(c), 2(d), 4(a) and 4(b). As an example, we give the Hilbert-style syntactic proof of the theorem $\neg\mathsf{B}_i\varphi \rightarrow \mathsf{B}_i\neg\mathsf{B}_i\varphi$:

1. $\vdash \neg\mathsf{B}_i\varphi \rightarrow \widehat{\mathsf{K}}_i\neg\mathsf{B}_i\varphi$
   By Axiom 2(b)
2. $\vdash \widehat{\mathsf{K}}_i\varphi \rightarrow \mathsf{K}_i\widehat{\mathsf{K}}_i\widehat{\mathsf{K}}_i\varphi$
   By Axiom 2(d)
3. $\vdash \mathsf{K}_i\widehat{\mathsf{K}}_i\widehat{\mathsf{K}}_i\varphi \rightarrow \mathsf{K}_i\widehat{\mathsf{K}}_i\varphi$
   By Axioms 2(a), 2(c) and necessitation for K_i
4. $\vdash \widehat{\mathsf{K}}_i\varphi \rightarrow \mathsf{K}_i\widehat{\mathsf{K}}_i\varphi$
   From 2 and 3
5. $\vdash \widehat{\mathsf{K}}_i\neg\mathsf{B}_i\varphi \rightarrow \mathsf{K}_i\widehat{\mathsf{K}}_i\neg\mathsf{B}_i\varphi$
   By item 4
6. $\vdash \mathsf{K}_i\widehat{\mathsf{K}}_i\neg\mathsf{B}_i\varphi \rightarrow \mathsf{K}_i\neg\mathsf{B}_i\varphi$
   By Axioms 2(a), 4(b) and necessitation for K_i
7. $\vdash \mathsf{K}_i\neg\mathsf{B}_i\varphi \rightarrow \mathsf{B}_i\neg\mathsf{B}_i\varphi$
   By Axiom 4(a)
8. $\vdash \neg\mathsf{B}_i\varphi \rightarrow \mathsf{B}_i\neg\mathsf{B}_i\varphi$
   From 1, 5, 6 and 7

Axiom 3(c) is a *local* version of the Axiom T for the belief operator B_i in the sense of Ditmarsch et al. (2012). It is *local* since the property $\mathsf{B}_i\varphi \rightarrow \varphi$ holds only if the atomic formula $poss_i$ is true. That is, the formula $\mathsf{B}_i\varphi \rightarrow \varphi$ may hold in the actual world if $poss_i$ is true, but not in all worlds as there exists some worlds in which $poss_i$ is false.

The third group of axioms including Axioms 4(a), 4(b) and 4(c) captures the basic relationships between knowledge and belief, and between common knowledge and common belief. In particular, we have the principle "knowing implies believing", the principle "believing implies knowing that one believes" and the principle "commonly knowing implies commonly believing". Finally, the fourth group of axioms is about the properties of knowledge and choice. In particular, Axioms 5(a) and 5(b) together

- **Axioms for LEG**:
  - (1) All tautologies of classical propositional logic
  - (2) Axioms for knowledge and common knowledge
    - (a) $\mathsf{K}_i(\varphi \to \psi) \to (\mathsf{K}_i\varphi \to \mathsf{K}_i\psi)$
    - (b) $\mathsf{K}_i\varphi \to \varphi$
    - (c) $\mathsf{K}_i\varphi \to \mathsf{K}_i\mathsf{K}_i\varphi$
    - (d) $\varphi \to \mathsf{K}_i\widehat{\mathsf{K}}_i\varphi$
    - (e) $\mathsf{CK}(\varphi \to \psi) \to (\mathsf{CK}\varphi \to \mathsf{CK}\psi)$
    - (f) $\mathsf{CK}\varphi \to \mathsf{K}_i\varphi$
    - (g) $\mathsf{CK}\varphi \to \mathsf{K}_i\mathsf{CK}\varphi$
    - (h) $\mathsf{CK}(\varphi \to \mathsf{EK}\varphi) \to (\varphi \to \mathsf{CK}\varphi)$
  - (3) Axioms for belief and common belief
    - (a) $\mathsf{B}_i(\varphi \to \psi) \to (\mathsf{B}_i\varphi \to \mathsf{B}_i\psi)$
    - (b) $\neg(\mathsf{B}_i\varphi \wedge \mathsf{B}_i\neg\varphi)$
    - (c) $poss_i \to (\mathsf{B}_i\varphi \to \varphi)$
    - (d) $\mathsf{CB}(\varphi \to \psi) \to (\mathsf{CB}\varphi \to \mathsf{CB}\psi)$
    - (e) $\mathsf{CB}\varphi \to \mathsf{EB}\varphi$
    - (f) $\mathsf{CB}\varphi \to \mathsf{EBCB}\varphi$
    - (g) $\mathsf{CB}(\varphi \to \mathsf{EB}\varphi) \to (\mathsf{EB}\varphi \to \mathsf{CB}\varphi)$
  - (4) Interrelation between knowledge and belief, and between common knowledge and common belief
    - (a) $\mathsf{K}_i\varphi \to \mathsf{B}_i\varphi$
    - (b) $\mathsf{B}_i\varphi \to \mathsf{K}_i\mathsf{B}_i\varphi$
    - (c) $\mathsf{CK}\varphi \to \mathsf{CB}\varphi$
  - (5) Axioms for knowledge and choice
    - (a) $\bigvee_{s_i \in S_i} choose_i(s_i)$
    - (b) $choose_i(s_i) \to \neg choose_i(s_i')$ if $s_i \neq s_i'$
    - (c) $choose_i(s_i) \to \mathsf{K}_i\, choose_i(s_i)$
    - (d) $\widehat{\mathsf{K}}_i\, choose_{-i}(s_{-i})$
- **Rules of inference for LEG**:
  - (6) From φ and $\varphi \to \psi$ infer ψ
  - (7) From φ infer $\mathsf{CK}\varphi$

Fig. 4 Axiomatization of LEG

say that a player chooses exactly one strategy. According to Axiom 5(c), a player knows the strategy that she chooses. According to Axiom 5(d), for every strategy of the other players, a player envisages a situation in which this strategy is played.

Notice that the concept of strong belief has defined in Sect. 4 as an abbreviation. Thus, in the axiomatization of the logic LEG there is no specific principle for the operator SB_i. An equivalent solution would consist in introducing the strong belief operator as a primitive operator in the object language of the logic LEG, instead of defining it as an abbreviation, and in adding the following reduction axiom for strong belief to the axiomatization of the logic LEG:

$$\mathsf{SB}_i\varphi \leftrightarrow \mathsf{K}_i\varphi \vee (\mathsf{K}_i(poss_i \to \varphi) \wedge \mathsf{K}_i(\neg poss_i \to \neg\varphi))$$

The following theorem provides a complexity result for the satisfiability problem of LEG.

Theorem 7 *The satisfiability problem of LEG relative to the class of epistemic game models is decidable in exponential time.*

The interesting aspect of this complexity result is that, although the logic LEG allows to represent properties of the worlds in the object language via the special atomic formulas $poss_i$, it has the same complexity as the logic of common knowledge.

7 Conclusion

We have presented a logic for interactive epistemology called LEG (Logic of Epistemic attitudes and Games). The logic LEG is based on a qualitative representation of epistemic individual and group attitudes including knowledge, belief, strong belief, common knowledge and common belief. An application of the logic LEG to the characterization of the epistemic conditions of solution concepts has been given in the paper. This includes "1-round of deletion of weakly dominated strategies, followed by iterated deletion of strongly dominated strategies" (DWDS^1-IDSDS) and "2-rounds of deletion of weakly dominated strategies, followed by iterated deletion of strongly dominated strategies" (DWDS^2-IDSDS). In the last section of the paper, we have provided a sound and complete axiomatization for this logic as well as a complexity result for its satisfiability problem.

We envisage two directions of future research. First of all, we plan to generalize Theorem 4 to the the characterization of the epistemic conditions of "n-rounds of deletion of weakly dominated strategies, followed by iterated deletion of strongly dominated strategies" (DWDS^n-IDSDS), for any arbitrary integer n. Secondly, we plan to extend our analysis to solution concepts for games in extensive form. This will require to extend the logic LEG by operators of temporal logic in order to model how epistemic attitudes of the players evolve during the game.

Appendix 1: Proof of Theorem 1

Let $i \in Agt$. Then:

1. $s_i \notin \mathsf{SD}_i^{\Gamma}$ if and only if there exists an epistemic game model $M = (W, \mathcal{E}_1, \ldots, \mathcal{E}_n, \mathcal{B}_1, \ldots, \mathcal{B}_n, \mathsf{ch}_1, \ldots, \mathsf{ch}_n, \mathcal{V})$ and a world $w \in W$ such that $M, w \models \mathsf{WRat}_i(s_i)$;
2. $s_i \notin \mathsf{WD}_i^{\Gamma}$ if and only if there exists an epistemic game model $M = (W, \mathcal{E}_1, \ldots, \mathcal{E}_n, \mathcal{B}_1, \ldots, \mathcal{B}_n, \mathsf{ch}_1, \ldots, \mathsf{ch}_n, \mathcal{V})$ and a world $w \in W$ such that $M, w \models \mathsf{PRat}_i(s_i)$.

Proof We only prove the second item, as the first item can be proved in a very similar way.

As to the left-to-right direction, let us assume that $s_i \notin \mathsf{WD}_i^{\Gamma}$. We are going to build an epistemic game model $M = (W, \mathcal{E}_1, \ldots, \mathcal{E}_n, \mathcal{B}_1, \ldots, \mathcal{B}_n, \mathsf{ch}_1, \ldots, \mathsf{ch}_n, \mathcal{V})$ and a world $w \in W$ such that $M, w \models \mathsf{PRat}_i(s_i)$. The model M is defined as follows:

- $W = \{w_s : s \in S\}$,
- For all $i \in Agt$, $\mathcal{E}_i = \mathcal{B}_i = \{(w_s, w_{s'}) : w_s, w_{s'} \in W \text{ and } s[i] = s'[i]\}$,
- For all $w_s \in W$ and for all $i \in Agt$, $\mathsf{ch}_i(w_s) = s[i]$,
- For all $w_s \in W$, $\mathcal{V}(w_s) = Atm$.

It is straightforward to verify that $M, w_s \models \mathsf{PRat}_i(s_i)$ for all $w_s \in W$ such that $\mathsf{ch}_i(w_s) = s_i$.

As to the right-to-left direction, we prove it by reductio ad absurdum. Let us suppose that $s_i \in \mathsf{WD}_i^{\Gamma}$. Moreover, let $M = (W, \mathcal{E}_1, \ldots, \mathcal{E}_n, \mathcal{B}_1, \ldots, \mathcal{B}_n, \mathsf{ch}_1, \ldots, \mathsf{ch}_n, \mathcal{V})$ be an epistemic game model and let $w \in W$ such that $M, w \models \mathsf{PRat}_i(s_i)$. $s_i \in \mathsf{WD}_i^{\Gamma}$ means that there exists $s_i' \in S_i$ such that $\forall s_{-i} \in S_{-i}, U_i(s_i, s_{-i}) \leq U_i(s_i', s_{-i})$ and $\exists s'_{-i} \in S_{-i}$ such that $U_i(s_i, s_{-i}) < U_i(s_i', s_{-i})$. By the Condition (**C4**) on epistemic game models and the definition of perfect rationality, the latter implies that $M, w \models \neg\mathsf{PRat}_i(s_i)$. This is in contradiction with the initial assumption. □

Appendix 2: Proof of Theorem 4

Let

$$\mathsf{ComplAss} \stackrel{\mathtt{def}}{=} \bigwedge_{i \in Agt} \bigwedge_{s_{-i} \in S_{-i} : s_{-i} \in S_{-i,1}^{DWDS^2 - IDSDS}} \widehat{\mathsf{K}}_i(choose_{-i}(s_{-i}) \wedge \mathsf{AllPRat}_{-i}).$$

Then:

$$\models \mathsf{CB}(\mathsf{ComplAss} \wedge \mathsf{AllPRat}_{Agt} \wedge \bigwedge_{i \in Agt} \mathsf{SB}_i\ \mathsf{AllPRat}_{-i}) \rightarrow \bigvee_{s \in S^{DWDS^2 - IDSDS}} choose(s)$$

Proof Instead of proving

$$\models \mathsf{CB}(\mathsf{ComplAss} \wedge \mathsf{AllPRat}_{Agt} \wedge \bigwedge_{i \in Agt} \mathsf{SB}_i\ \mathsf{AllPRat}_{-i}) \rightarrow \bigvee_{s \in S^{DWDS^2 - IDSDS}} choose(s)$$

we simply prove that for all $s \notin S^{DWDS^2 - IDSDS}$:

$$\models \mathsf{CB}(\mathsf{ComplAss} \wedge \mathsf{AllPRat}_{Agt} \wedge \bigwedge_{i \in Agt} \mathsf{SB}_i\ \mathsf{AllPRat}_{-i}) \rightarrow \neg choose(s)$$

Indeed, we have the following valid equivalence:

$$\models \bigvee_{s \in S^{DWDS^2 - IDSDS}} choose(s) \leftrightarrow \bigwedge_{s \notin S^{DWDS^2 - IDSDS}} \neg choose(s)$$

The proof is by induction. The proof of the inductive case goes exactly as the proof of the inductive case in the proof of Theorem 3.

Let us prove the base case first.

Base case. For all $s \notin S_2^{DWDS^2 - IDSDS}$ we prove that:

(A) $\models \mathsf{CB}(\mathsf{ComplAss} \wedge \mathsf{AllPRat}_{Agt} \wedge \bigwedge_{i \in Agt} \mathsf{SB}_i\ \mathsf{AllPRat}_{-i}) \rightarrow \neg choose(s)$

To prove (A), it is sufficient to prove the following validity (B), as we have the following validity:

$$\models \mathsf{CB}(\mathsf{ComplAss} \wedge \mathsf{AllPRat}_{Agt} \wedge \bigwedge_{i \in Agt} \mathsf{SB}_i\ \mathsf{AllPRat}_{-i}) \rightarrow (\mathsf{ComplAss} \wedge \mathsf{AllPRat}_{Agt} \wedge \bigwedge_{i \in Agt} \mathsf{SB}_i\ \mathsf{AllPRat}_{-i})$$

For all $s \notin S_2^{DWDS^2-IDSDS}$ we have that:

(B) $\models (\mathsf{ComplAss} \wedge \mathsf{AllPRat}_{Agt} \wedge \bigwedge_{i \in Agt} \mathsf{SB}_i\ \mathsf{AllPRat}_{-i}) \rightarrow \neg choose(s)$

And to prove (B), it is sufficient to prove that if $s[i] \notin S_{i,2}^{DWDS^2-IDSDS}$ then:

(C) $\models \Big(\bigwedge_{s_{-i} \in S_{-i} : s_{-i} \in S_{-i,1}^{DWDS^2-IDSDS}} \widehat{\mathsf{K}}_i(choose_{-i}(s_{-i}) \wedge \mathsf{AllPRat}_{-i}) \wedge \mathsf{PRat}_i \wedge \mathsf{SB}_i\ \mathsf{AllPRat}_{-i} \Big) \rightarrow \neg choose_i(s[i])$

Let us prove (C) by reductio ad absurdum. We assume that $s[i] \notin S_{i,2}^{DWDS^2-IDSDS}$ and $M, w \models \bigwedge_{s_{-i} \in S_{-i} : s_{-i} \in S_{-i,1}^{DWDS^2-IDSDS}} \widehat{\mathsf{K}}_i(choose_{-i}(s_{-i}) \wedge \mathsf{AllPRat}_{-i}) \wedge \mathsf{PRat}_i \wedge \mathsf{SB}_i\ \mathsf{AllPRat}_{-i} \wedge choose_i(s[i])$ for some arbitrary epistemic game model $M = (W, \mathcal{E}_1, \ldots, \mathcal{E}_n, \mathcal{B}_1, \ldots, \mathcal{B}_n, \mathsf{ch}_1, \ldots, \mathsf{ch}_n, \mathcal{V})$ and world w in M. We are going to show that these two facts are inconsistent.

The rest of the proof makes use of the following Lemma 1.

Lemma 1 *Let $M = (W, \mathcal{E}_1, \ldots, \mathcal{E}_n, \mathcal{B}_1, \ldots, \mathcal{B}_n, \mathsf{ch}_1, \ldots, \mathsf{ch}_n, \mathcal{V})$ be an epistemic game model, let $w \in W$ and let $s_{-i}, s'_{-i} \in S_{-i}$. Then, if*

1. $M, w \models \bigwedge_{s_{-i} \in S_{-i} : s_{-i} \in S_{-i,1}^{DWDS^2-IDSDS}} \widehat{\mathsf{K}}_i(choose_{-i}(s_{-i}) \wedge \mathsf{AllPRat}_{-i}) \wedge \mathsf{SB}_i\ \mathsf{AllPRat}_{-i}$
2. $s_{-i} \in S_{-i,1}^{DWDS^2-IDSDS}$ *and* $s'_{-i} \notin S_{-i,1}^{DWDS^2-IDSDS}$

then $\Pi_{w,i}(choose_{-i}(s_{-i})) > \Pi_{w,i}(choose_{-i}(s'_{-i}))$.

Proof Assume that $M, w \models \bigwedge_{s_{-i} \in S_{-i} : s_{-i} \in S_{-i,1}^{DWDS^2-IDSDS}} \widehat{\mathsf{K}}_i(choose_{-i}(s_{-i}) \wedge \mathsf{AllPRat}_{-i}) \wedge \mathsf{SB}_i\ \mathsf{AllPRat}_{-i}$ and that $s_{-i} \in S_{-i,1}^{DWDS^2-IDSDS}$ and $s'_{-i} \notin S_{-i,1}^{DWDS^2-IDSDS}$. By Proposition 1, $M, w \models \mathsf{SB}_i\ \mathsf{AllPRat}_{-i}$ implies that for all $v \in ||\mathsf{AllPRat}_{-i}||_{w,i}$ and for all $u \in ||\neg \mathsf{AllPRat}_{-i}||_{w,i}$, $\pi_i(u) < \pi_i(v)$. Thus, by the second item of Theorem 1 and the fact that $M, w \models \widehat{\mathsf{K}}_i(choose_{-i}(s_{-i}) \wedge \mathsf{AllPRat}_{-i})$ and that $s_{-i} \in S_{-i,1}^{DWDS^2-IDSDS}$ and $s'_{-i} \notin S_{-i,1}^{DWDS^2-IDSDS}$, it follows that there exists $v \in ||choose_{-i}(s_{-i})||_{w,i}$ such that, for all $u \in ||choose_{-i}(s'_{-i})||_{w,i}$, $\pi_i(u) < \pi_i(v)$. The latter implies that $\Pi_{w,i}(choose_{-i}(s_{-i})) > \Pi_{w,i}(choose_{-i}(s'_{-i}))$. □

From $M, w \models \bigwedge_{s_{-i} \in S_{-i} : s_{-i} \in S_{-i,1}^{DWDS^2-IDSDS}} \widehat{\mathsf{K}}_i(choose_{-i}(s_{-i}) \wedge \mathsf{AllPRat}_{-i}) \wedge \mathsf{SB}_i\ \mathsf{AllPRat}_{-i}$, by Lemma 1, it follows that:

(D) if $s'_{-i} \in S^{DWDS^2-IDSDS}_{-i,1}$ and $s''_{-i} \notin S^{DWDS^2-IDSDS}_{-i,1}$ then $\Pi_{w,i}(choose_{-i}(s'_{-i})) > \Pi_{w,i}(choose_{-i}(s''_{-i}))$.

$s[i] \notin S^{DWDS^2-IDSDS}_{i,2}$ implies that:

(E1) $s[i] \notin S^{DWDS^2-IDSDS}_{i,1}$ or

(E2) $s[i] \in S^{DWDS^2-IDSDS}_{i,1}$ and $s[i] \notin S^{DWDS^2-IDSDS}_{i,2}$

We split the proof in the two subcases: (E1) and (E2).

Proof for the case (E1).

$s[i] \notin S^{DWDS^2-IDSDS}_{i,1}$ implies that:

(F1) There is $s'_i \in S^{DWDS^2-IDSDS}_i$ such that: (1) $s'_i \neq s[i]$ and (2) $\langle s[i], s'_{-i}\rangle <_i \langle s'_i, s'_{-i}\rangle$ for some $s'_{-i} \in S^{DWDS^2-IDSDS}_{-i}$ and (3) $\langle s[i], s''_{-i}\rangle \leq_i \langle s'_i, s''_{-i}\rangle$ for all $s''_{-i} \in S^{DWDS^2-IDSDS}_{-i}$.

From (F1), by the Condition (**C4**) on epistemic game models, it follows that:

(G1) There are $s'_i \in S^{DWDS^2-IDSDS}_i$ and $s'_{-i} \in S^{DWDS^2-IDSDS}_{-i}$ and $v \in W$ such that: (1) $s'_i \neq s[i]$ and (2) $w\mathcal{E}_i v$ and (3) $M, v \models choose_{-i}(s'_{-i})$ and (4) $\langle s[i], s'_{-i}\rangle <_i \langle s'_i, s'_{-i}\rangle$ and (5) for all $u \in W$ such that $w\mathcal{E}_i u$ and for all $s''_{-i} \in S^{DWDS^2-IDSDS}_{-i}$: if $M, u \models choose_{-i}(s''_{-i})$ then $\langle s[i], s''_{-i}\rangle \leq_i \langle s'_i, s''_{-i}\rangle$.

But (G1) is in contradiction with $M, w \models \mathsf{PRat}_i$ and $M, w \models choose_i(s[i])$.

Proof for the case (E2).

(E2) implies that:

(F2) There is $s'_i \in S^{DWDS^2-IDSDS}_{i,1}$ such that: (1) $s'_i \neq s[i]$ and (2) $\langle s[i], s'_{-i}\rangle <_i \langle s'_i, s'_{-i}\rangle$ for some $s'_{-i} \in S^{DWDS^2-IDSDS}_{-i,1}$ and (3) $\langle s[i], s''_{-i}\rangle \leq_i \langle s'_i, s''_{-i}\rangle$ for all $s''_{-i} \in S^{DWDS^2-IDSDS}_{-i,1}$.

By the Condition (**C4**) on epistemic game models, (F2) together with $M, w \models \mathsf{PRat}_i$ and $M, w \models choose_i(s[i])$ imply that:

(G2) There are $s'_{-i} \in S^{DWDS^2-IDSDS}_{-i,1}$ and $s''_{-i} \notin S^{DWDS^2-IDSDS}_{-i,1}$ such that $\Pi_{w,i}(choose_{-i}(s'_{-i})) \leq \Pi_{w,i}(choose_{-i}(s''_{-i}))$.

But (G2) is in contradiction with (D). This proves the base case.

Let us now prove the inductive case.

Inductive case. For $m > 1$, we assume that if $s \notin S^{DWDS^2-IDSDS}_m$ then:

(Inductive Hypothesis) $\models \mathsf{CB}(\mathsf{ComplAss} \wedge \mathsf{AllPRat}_{Agt} \wedge \bigwedge_{i \in Agt} \mathsf{SB}_i\, \mathsf{AllPRat}_{-i}) \rightarrow \neg choose(s)$

We are going to prove that if $s \notin S^{DWDS^2-IDSDS}_{m+1}$ then:

(A) $\models \mathsf{CB}(\mathsf{ComplAss} \wedge \mathsf{AllPRat}_{Agt} \wedge \bigwedge_{i \in Agt} \mathsf{SB}_i\, \mathsf{AllPRat}_{-i}) \rightarrow \neg choose(s)$

Let us take an arbitrary epistemic game model M and world w and assume that $M, w \models \mathsf{CB}(\mathsf{ComplAss} \wedge \mathsf{AllPRat}_{Agt} \wedge \bigwedge_{i \in Agt} \mathsf{SB}_i\ \mathsf{AllPRat}_{-i})$ and $M, w \models choose(s)$. We are going to show that $s \in S_{m+1}^{DWDS^2-IDSDS}$.

$M, w \models \mathsf{CB}(\mathsf{ComplAss} \wedge \mathsf{AllPRat}_{Agt} \wedge \bigwedge_{i \in Agt} \mathsf{SB}_i\ \mathsf{AllPRat}_{-i})$ implies $\models \mathsf{AllPRat}_{Agt}$ which in turn implies:

(A) $\models \bigwedge_{i \in Agt} \mathsf{SRat}_i$

Moreover, we have the following validity by the property $\models \mathsf{CB}_{Agt}\varphi \rightarrow \mathsf{B}_i\mathsf{CB}_{Agt}\varphi$ for every $i \in Agt$:

(B) $\models \mathsf{CB}(\mathsf{ComplAss} \wedge \mathsf{AllPRat}_{Agt} \wedge \bigwedge_{i \in Agt} \mathsf{SB}_i\ \mathsf{AllPRat}_{-i}) \rightarrow \bigwedge_{i \in Agt} \mathsf{B}_i\mathsf{CB}(\mathsf{ComplAss} \wedge \mathsf{AllPRat}_{Agt} \wedge \bigwedge_{i \in Agt} \mathsf{SB}_i\ \mathsf{AllPRat}_{-i})$

Therefore, from $M, w \models \mathsf{CB}(\mathsf{ComplAss} \wedge \mathsf{AllPRat}_{Agt} \wedge \bigwedge_{i \in Agt} \mathsf{SB}_i\ \mathsf{AllPRat}_{-i})$ we infer that:

(C) $M, w \models \bigwedge_{i \in Agt} \mathsf{B}_i\mathsf{CB}(\mathsf{ComplAss} \wedge \mathsf{AllPRat}_{Agt} \wedge \bigwedge_{i \in Agt} \mathsf{SB}_i\ \mathsf{AllPRat}_{-i})$

By the inductive hypothesis, Axiom K and the rule of necessitation for the belief operator B_i, from (C) it follows that if $s' \notin S_m^{DWDS^1-IDSDS}$ then:

(D) $M, w \models \bigwedge_{i \in Agt} \mathsf{B}_i \neg choose(s')$

From (B), (D) and $M, w \models choose(s)$ it follows that for every $i \in Agt$ and for all $s'_i \in S_i$ either there is $s' \in S_m^{DWDS^1-IDSDS}$ such that $\langle s'_i, s'_{-i}\rangle <_i \langle s[i], s'_{-i}\rangle$ or for all $s' \in S_m^{DWDS^1-IDSDS}$ we have $\langle s'_i, s'_{-i}\rangle \leq_i \langle s[i], s'_{-i}\rangle$. The latter implies that for every $i \in Agt$ we have $s[i] \in S_{i,m+1}^{DWDS^1-IDSDS}$ which is equivalent to $s \in S_{m+1}^{DWDS^1-IDSDS}$. □

Appendix 3: Proof of Theorem 5

Let $s \in S$. Then:

1. If $s \in S^{IDSDS}$ then there exists an epistemic game model $M = (W, \mathcal{E}_1, \ldots, \mathcal{E}_n, \mathcal{B}_1, \ldots, \mathcal{B}_n, \mathsf{ch}_1, \ldots, \mathsf{ch}_n, \mathcal{V})$ and a world $w \in W$ such that $M, w \models choose(s) \wedge \mathsf{CB}\ \mathsf{AllWRat}_{Agt}$;
2. If $s \in S^{DWDS^1-IDSDS}$ then there exists an epistemic game model $M = (W, \mathcal{E}_1, \ldots, \mathcal{E}_n, \mathcal{B}_1, \ldots, \mathcal{B}_n, \mathsf{ch}_1, \ldots, \mathsf{ch}_n, \mathcal{V})$ and a world $w \in W$ such that $M, w \models choose(s) \wedge \mathsf{CB}\ \mathsf{AllPRat}_{Agt}$;
3. If $s \in S^{DWDS^2-IDSDS}$ then there exists an epistemic game model $M = (W, \mathcal{E}_1, \ldots, \mathcal{E}_n, \mathcal{B}_1, \ldots, \mathcal{B}_n, \mathsf{ch}_1, \ldots, \mathsf{ch}_n, \mathcal{V})$ and a world $w \in W$ such that $M, w \models choose(s) \wedge \mathsf{CB}(\mathsf{ComplAss} \wedge \mathsf{AllPRat}_{Agt} \wedge \bigwedge_{i \in Agt} \mathsf{SB}_i\ \mathsf{AllPRat}_{-i})$.

Proof We only prove the third item, as the first item and the second item can be proved in a very similar way.

We build the following epistemic game model $M = (W, \mathcal{E}_1, \ldots, \mathcal{E}_n, \mathcal{B}_1, \ldots, \mathcal{B}_n, \mathsf{ch}_1, \ldots, \mathsf{ch}_n, \mathcal{V})$:

– $W = \{w_s{:}s \in S\}$,
– For all $i \in Agt$, $\mathcal{E}_i = \{(w_s, w_{s'}){:}w_s, w_{s'} \in W \text{ and } s[i] = s'[i]\}$,
– For all $i \in Agt$, $\mathcal{B}_i = \{(w_s, w_{s'}){:}w_s \mathcal{E}_i w_{s'} \text{ and } s'_{-i} \in S_{-i,1}^{DWDS^2-IDSDS}\}$,
– For all $w_s \in W$ and for all $i \in Agt$, $\mathsf{ch}_i(w_s) = s[i]$,
– For all $w_s \in W$, $\mathcal{V}(w_s) = Atm$.

It is routine to check that, for all $w_s \in W$ and for all $i \in Agt$:

(A) $s[i] \in S_{i,1}^{DWDS^2-IDSDS}$ iff $M, w_s \models \mathsf{PRat}_i$.

The right-to-left direction follows from Theorem 1 above. The left-to-right direction follows from the construction of the previous epistemic game model M and the definition of PRat_i.

By the previous item (A), from the definition of SB_i and the construction of the previous epistemic game model M, for all $w_s \in W$ and for all $i \in Agt$, we have:

(B) $M, w_s \models \mathsf{SB}_i\ \mathsf{AllPRat}_{-i}$.

By the previous item (A), from the definition of $\mathsf{ComplAss}$ and the construction of the previous epistemic game model M, for all $w_s \in W$, we have:

(C) $M, w_s \models \mathsf{ComplAss}$.

By the previous item (A), from the construction of the previous epistemic game model M, we have that:

(D) If $s \in S_1^{DWDS^2-IDSDS}$ then $M, w_{s'} \models \mathsf{AllPRat}_{Agt}$ for all $w_{s'} \in \mathcal{B}^+(w_s)$.

Now, let $s \in S^{DWDS^2-IDSDS}$. Hence, $s \in S_1^{DWDS^2-IDSDS}$. Thus, from the previous items (B), (C) and (D) it follows that $M, w_s \models choose(s) \wedge \mathsf{CB}(\mathsf{ComplAss} \wedge \mathsf{AllPRat}_{Agt} \wedge \bigwedge_{i \in Agt} \mathsf{SB}_i\ \mathsf{AllPRat}_{-i})$. □

Appendix 4: Proof of Theorem 6

The set of validities of the logic LEG relative to the class of epistemic game models is completely axiomatized by the principles given in Fig. 4.

Proof Proving that the axioms given in Fig. 4 are sound with respect to the class of epistemic game models (EGMs) and that the inference rules preserve validity is just a routine task and we do not give it here.

As to completeness, the proof consists of three steps.

Step 1 Let us first observe that the LEG semantics in terms of epistemic game models is equivalent to a LEG semantics in terms of multi-relational Kripke models in which formulas $choose_i(s_i)$ and $poss_i$ are seen as atomic propositions.

Definition 10 (*Multi-relational Kripke model*) A multi-relational Kripke model (MKM) for the strategic game $\Gamma = (Agt, \{S_i\}_{i \in Agt}, \{U_i\}_{i \in Agt})$ and for the set of atomic propositions Atm is a tuple $M = (W, \{\mathcal{E}_i\}_{i \in Agt}, \{\mathcal{B}_i\}_{i \in Agt}, \pi)$ where W, $\{\mathcal{E}_i\}_{i \in Agt}$ and $\{\mathcal{B}_i\}_{i \in Agt}$ are as in Definition 1 and π is the valuation function:

$$\pi : W \longrightarrow 2^{Atm \cup \Phi \cup \Psi}$$

with $\Phi_i = \{choose_i(s_i) : s_i \in S_i\}$, $\Phi = \bigcup_{i \in Agt} \Phi_i$ and $\Psi = \{poss_i : i \in Agt\}$, which satisfies the following conditions:

(**C3***) For all $w, v \in W$, for all $i \in Agt$ and for all $s_i \in S_i$: if $choose_i(s_i) \in \pi(w)$ and $w\mathcal{E}_i v$ then $choose_i(s_i) \in \pi(v)$;
(**C4***) For all $w \in W$, for all $i \in Agt$ and for all $s_{-i} \in S_{-i}$: there is $u \in W$ such that $w\mathcal{E}_i u$ and $choose_j(s_j) \in \pi(u)$ for all $j \in Agt \setminus \{i\}$;
(**C5**) For all $w \in W$ and for all $i \in Agt$: $\pi_i(w)$ is a singleton with $\pi_i(w) = \pi(w) \cap \Phi_i$;
(**C6**) For all $w \in W$ and for all $i \in Agt$: if $poss_i \in \pi(w)$ then $w\mathcal{B}_i w$.

The truth conditions of formulas in $\mathcal{L}_{\mathsf{LEG}}(\Gamma, Atm)$ relative to MKMs are like the truth conditions relative to EGMs except for formulas $choose_i(s_i)$ and $poss_i$, which are interpreted by means of the valuation function π as follows:

$$\begin{array}{ll} M, w \models poss_i & \text{iff } poss_i \in \pi(w) \\ M, w \models choose_i(s_i) & \text{iff } choose_i(s_i) \in \pi(w) \end{array}$$

We write $\models_{MKM} \varphi$ to mean that the LEG-formula φ is *valid* relative to the class of MKMs.

We have the following equivalence result:

Lemma 2 *Let $\varphi \in \mathcal{L}_{LEG}(\Gamma, Atm)$. Then, $\models \varphi$ iff $\models_{MKM} \varphi$.*

Step 2 The second step consists in introducing the class of *weak* multi-relational Kripke models (WMKMs) that are like MKMs except that they do not necessarily satisfy Conditions (**C3***), (**C4***) and (**C5**).

We write $\models_{WMKM} \varphi$ to mean that the LEG-formula φ is *valid* relative to the class of WMKMs.

Moreover, for any finite set Δ of LEG-formulas, we write $\Delta \models_{WMKM} \varphi$ to mean that φ is a *logical consequence* of the set of formulas Δ relative to the class of WMKMs. That is, $\Delta \models_{WMKM} \varphi$ iff, for every *weak* multi-relational Kripke model M, if $M, w \models \bigwedge_{\psi \in \Delta} \psi$ for all $w \in W$, then $M, w \models \varphi$ for all $w \in W$.

The following Proposition 4 highlights that the validity problem relative to the class of MKMs is reducible to the logical consequence problem relative to the class of WMKMs.

Proposition 4 *Let*

$$\begin{aligned} \Delta_0 = & \{\bigvee_{s_i \in S_i} choose_i(s_i) : i \in Agt\} \cup \\ & \{choose_i(s_i) \rightarrow \neg choose_i(s_i') : i \in Agt \text{ and } s_i, s_i' \in S_i \text{ with } s_i \neq s_i'\} \cup \\ & \{choose_i(s_i) \rightarrow \mathsf{K}_i choose_i(s_i) : i \in Agt \text{ and } s_i \in S_i\} \cup \\ & \{\widehat{\mathsf{K}}_i choose_{-i}(s_{-i}) : i \in Agt \text{ and } s_{-i} \in S_{Agt \setminus \{i\}}\} \end{aligned}$$

Then, for every LEG-formula φ, $\models_{MKM} \varphi$ iff $\Delta_0 \models_{WMKM} \varphi$.

Proof We just need to observe that the (global) axioms in Δ_0 force a *weak* multi-relational Kripke models to satisfy Conditions (**C3***), (**C4***) and (**C5**). That is, M is a WMKM in which the formula $\bigwedge_{\psi \in \Delta_0} \psi$ is true (i.e., $M, w \models \bigwedge_{\psi \in \Delta_0} \psi$ for all w in M) iff M is a MKM. Therefore, the class of WMKMs in which the formula $\bigwedge_{\psi \in \Delta_0} \psi$ is true coincides with the class of MKMs. □

The following Proposition 5 highlights that, thanks to the common knowledge modality CK, the logical consequence problem relative to the class of WMKMs can be reduced to the validity problem relative to the class of WMKMs.

Proposition 5 *For every* LEG*-formula* φ *and for every finite set* Δ *of* LEG*-formulas,* $\Delta \models_{WMKM} \varphi$ *iff* $\models_{WMKM} \mathsf{CK} \bigwedge_{\psi \in \Delta} \psi \rightarrow \varphi$.

Step 3 The third step consists in providing an axiomatization result for LEG relative to the class of WMKMs.

Lemma 3 *The set of validities of the logic* LEG *relative to the class of WMKMs is completely axiomatized by the groups of axioms (1), (2), (3) and (4) and by the rules of inference (6) and (7) in Fig. 4.*

Proof The proof just consists in adapting the completeness proof by Kraus and Lehmann (1988). Kraus and Lehmann provide an axiomatization result for the logic of belief, knowledge, common belief and common knowledge whose language is defined by the following grammar:

$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \psi \mid \mathsf{K}_i\varphi \mid \mathsf{B}_i\varphi \mid \mathsf{CK}\varphi \mid \mathsf{CB}\varphi$$

and which is interpreted over structures that are like WMKMs except that they do not necessarily satisfy Condition (**C6**). Their axiomatization consists of the groups of axioms (1), (2) and (4), the rules of inference (6) and (7) and Axioms (3a), (3b), (3d), (3e), (3f) and (3g) in Fig. 4. It is a routine task to slightly modify Kraus and Lehmann's proof by adding Condition (**C6**) in the semantics and proving that the resulting structures are completely axiomatized by their axioms and inference rules *plus* Axiom (3c).[4] □

The last element we need for proving Theorem 6 is the following Proposition 6. Let $\vdash_{\mathsf{LEG}} \varphi$ and $\Vdash_{\mathsf{LEG}} \varphi$ mean, respectively, that the LEG-formula φ is provable via the groups of axioms (1), (2), (3), (4) and (5) and the rules of inference (6) and (7) in Fig. 4 and that the LEG-formula φ is provable via the groups of axioms (1), (2), (3) and (4) and the rules of inference (6) and (7) in Fig. 4.

[4] Kraus and Lehmann's proof can be easily adapted by adding the following condition to their definition of *standard set* (Kraus and Lehmann 1988, Def. 2.8):

– If $\mathsf{B}_i\varphi \in D$, $poss_i \in D$ and D is a standard set then $\varphi \in D$.

This modification of the definition of standard set guarantees that the binary relation $\sim_i$ over standard sets defined in (Kraus and Lehmann 1988, Def. 2.10) correctly satisfies the Condition (**C6**) (i.e., if D is a standard set and $poss_i \in D$ then $D \sim_i D$).

Proposition 6 *For every LEG-formula φ, if $\Vdash\!\vdash_{LEG} \mathsf{CK} \bigwedge_{\psi\in\Delta_0} \psi \rightarrow \varphi$ then $\vdash_{LEG} \varphi$, where Δ_0 is defined as in Proposition 4.*

Proof Suppose $\Vdash\!\vdash_{\mathsf{LEG}} \mathsf{CK} \bigwedge_{\psi\in\Delta_0} \psi \rightarrow \varphi$. Hence, $\vdash_{\mathsf{LEG}} \mathsf{CK} \bigwedge_{\psi\in\Delta_0} \psi \rightarrow \varphi$.

By the inference rule (7) (viz. necessitation for CK) and the group of axioms (5), we have $\vdash_{\mathsf{LEG}} \bigwedge_{\psi\in\Delta_0} \mathsf{CK}\psi$. By Axiom 2(e), we can derive $\vdash_{\mathsf{LEG}} \mathsf{CK} \bigwedge_{\psi\in\Delta_0} \psi$.

Consequently, by the inference rule (6) (viz. modus ponens), we have that $\vdash_{\mathsf{LEG}} \varphi$. □

Propositions 4, 5 and 6 together with Lemmas 2 and 3 are sufficient to prove Theorem 6.

Suppose that $\models_{\mathsf{LEG}} \varphi$. Hence, by Lemma 2, $\models_{MKM} \varphi$. Hence, by Propositions 4 and 5, $\models_{WMKM} \mathsf{CK} \bigwedge_{\psi\in\Delta_0} \psi \rightarrow \varphi$. By Lemma 3, it follows that $\Vdash\!\vdash_{\mathsf{LEG}} \mathsf{CK} \bigwedge_{\psi\in\Delta_0} \psi \rightarrow \varphi$. Hence, by Proposition 6, $\vdash_{\mathsf{LEG}} \varphi$. □

Appendix 5: Proof of Theorem 7

The satisfiability problem of LEG relative to the class of epistemic game models is decidable in exponential time.

Proof The following Lemma 4 follows from the fact that LEG is an extension of the logic of common knowledge whose satisfiability problem is ExpTime-complete Fagin et al. (1995).

Lemma 4 *The satisfiability problem of LEG relative to the class of epistemic game models is ExpTime-hard.*

In order to prove that the satisfiability problem of LEG is in ExpTime, we are going to embed LEG into the decidable logic S5-PDL, i.e., the variant of propositional dynamic logic PDL Harel et al. (2000) in which atomic programs are interpreted by means of equivalence relations.

Let $Atm^+ = Atm \cup \{choose_i(s_i){:}i \in Agt \text{ and } s_i \in S_i\} \cup \{poss_i{:}i \in Agt\}$. The language $\mathcal{L}_{\mathrm{S5-PDL}}(Atm^+, Agt)$ of S5-PDL is defined as follows:

$$\pi ::= \sim_i \mid \pi_1;\pi_2 \mid \pi_1 \cup \pi_2 \mid \pi^* \mid ?\varphi$$
$$\varphi ::= p \mid \neg\varphi \mid \varphi \wedge \varphi \mid [\pi]\varphi$$

where p ranges over Atm^+ and i ranges over Agt. The dual of the operator $[\pi]$ is defined in the standard way as follows: $\langle\pi\rangle\varphi \stackrel{\text{def}}{=} \neg[\pi]\neg\varphi$.

S5-PDL models are tuples $M = \langle W, \mathcal{R}_{\sim_1}, \ldots, \mathcal{R}_{\sim_n}, \mathcal{I}\rangle$ where:

- W is a set of worlds;
- Every $\mathcal{R}_{\sim_i}$ is an equivalence relation on W;
- $\mathcal{I} : Atm^+ \longrightarrow 2^W$ is a valuation function.

Binary relations for complex programs are defined in the standard way as follows:

$$\begin{aligned}\mathcal{R}_{\pi_1;\pi_2} &= \mathcal{R}_{\pi_1};\mathcal{R}_{\pi_2}\\ \mathcal{R}_{\pi_1\cup\pi_2} &= \mathcal{R}_{\pi_1}\cup\mathcal{R}_{\pi_2}\\ \mathcal{R}_{\pi^*} &= (\mathcal{R}_{\pi})^*\\ \mathcal{R}_{?\varphi} &= \{(w,w){:}w\in W \text{ and } M,w\models\varphi\}\end{aligned}$$

Truth conditions of S5-PDL formulae are the standard ones for Boolean operators plus the following one for the operator $[\pi]$:

$$M,w\models[\pi]\varphi \text{ iff } M,v\models\varphi \text{ for all } v\in\mathcal{R}_{\pi}(w)$$

For any formula φ of the language $\mathcal{L}_{\mathrm{S5-PDL}}(Atm^{+},Agt)$, we write $\models_{\mathrm{S5-PDL}}\varphi$ if φ is S5-PDL *valid*, that is, if φ is true in all S5-PDL models (i.e., for all S5-PDL models M and for all worlds w in M, we have $M,w\models\varphi$). We say that φ is S5-PDL *satisfiable* if $\neg\varphi$ is not S5-PDL valid. Moreover, we shall say that φ is a global logical consequence in S5-PDL of a finite set of global axioms $\Delta=\{\chi_1,\ldots,\chi_n\}$, denoted by $\Delta\models_{\mathrm{S5-PDL}}\varphi$, if and only if for every S5-PDL model M, if Δ is true in M (i.e., for every world w in M, we have $M,w\models\chi_1\wedge\ldots\wedge\chi_n$) then φ is true in M too (i.e., for every world w in M, we have $M,w\models\varphi$).

Proposition 7 *The satisfiability problem of S5-PDL is in ExpTime.*

Proof The logic S5-PDL is polynomially embeddable into PDL extended with converse, by simulating S5 program $\sim_i$ with the composite program $(x_i\cup -x_i)^*$ where x_i is an arbitrary atomic program interpreted by means of the binary relation $\mathcal{R}_{x_i}$, and $-x_i$ is the converse of x_i . (Note indeed that relation $\mathcal{R}_{(x_i\cup -x_i)^*}$ is an equivalence relation.) The satisfiability problem of PDL with converse has been proved to be ExpTime-complete Vardi (1985). It follows that the satisfiability problem of S5-PDL is in ExpTime. □

We define the following translation from the LEG language $\mathcal{L}_{\mathsf{LEG}}(\Gamma,Atm)$ to $\mathcal{L}_{S5-\mathrm{PDL}}(Atm^{+},Agt)$ for $p\in Atm$ and $i\in Agt$:

$$\begin{aligned}tr(p) &= p\\ tr(choose_i(s_i)) &= choose_i(s_i)\\ tr(poss_i) &= poss_i\\ tr(\neg\varphi) &= \neg tr(\varphi)\\ tr(\varphi\wedge\psi) &= tr(\varphi)\wedge tr(\psi)\\ tr(\mathsf{K}_i\varphi) &= [\sim_i]tr(\varphi)\\ tr(\mathsf{B}_i\varphi) &= [\sim_i](poss_i\rightarrow tr(\varphi))\\ tr(\mathsf{CK}\varphi) &= [(\sim_1\cup\ldots\cup\sim_n)^*]tr(\varphi)\\ tr(\mathsf{CB}\varphi) &= [(\sim_1;?poss_1\cup\ldots\cup\sim_n;?poss_n)^*]tr(\varphi)\end{aligned}$$

As the following proposition highlights, the validity problem in LEG can be reduced to the problem of (global) logical consequence in S5-PDL.

Proposition 8 *A LEG formula φ is LEG valid, i.e., $\models_{LEG} \varphi$, if and only if $tr(\varphi)$ is a logical consequence of the set of global axioms Δ_0 in S5-PDL, i.e., $\Delta_0 \models_{S5-PDL} tr(\varphi)$, where:*

$$\begin{aligned}\Delta_0 = &\{\bigvee_{s_i \in S_i} choose_i(s_i) : i \in Agt\} \cup \\ &\{choose_i(s_i) \rightarrow \neg choose_i(s_i') : i \in Agt \text{ and } s_i, s_i' \in S_i \text{ with } s_i \neq s_i'\} \cup \\ &\{choose_i(s_i) \rightarrow \mathsf{K}_i choose_i(s_i) : i \in Agt \text{ and } s_i \in S_i\} \cup \\ &\{\widehat{\mathsf{K}}_i choose_{-i}(s_{-i}) : i \in Agt \text{ and } s_{-i} \in S_{Agt \setminus \{i\}}\}\end{aligned}$$

Proof (Sketch) ($\Rightarrow$) Take an arbitrary EGM $M = (W, \mathcal{E}_1, \ldots, \mathcal{E}_n, \mathcal{B}_1, \ldots, \mathcal{B}_n, \mathsf{ch}_1, \ldots, \mathsf{ch}_n, \mathcal{V})$ which satisfies formula φ. We build a corresponding S5-PDL model $M' = \langle W, \mathcal{R}_{\sim_1}, \ldots, \mathcal{R}_{\sim_n}, \mathcal{I}\rangle$ such that:

- For all $i \in AGT$, $\mathcal{R}_{\sim_i} = \mathcal{E}_i$;
- For all $p \in Atm$, $\mathcal{I}(p) = \mathcal{V}(p)$;
- For all $i \in Agt$, for all $s_i \in S_i$ and for all $w \in W$, $w \in \mathcal{I}(poss_i)$ iff $w \in \mathcal{B}_i(w)$;
- For all $i \in Agt$ and for all $w \in W$, $w \in \mathcal{I}(choose_i(s_i))$ iff $\mathsf{ch}_i(w) = s_i$.

By induction on the structure of φ, it is easy to check that $tr(\varphi)$ is satisfied by M' and that for every $\chi \in \Delta_0$, $M', w \models \chi$ for all $w \in W$.

($\Leftarrow$) Take an arbitrary S5-PDL model $M = \langle W, \mathcal{R}_{\sim_1}, \ldots, \mathcal{R}_{\sim_n}, \mathcal{I}\rangle$ which satisfies $tr(\varphi)$ such that that for all $\chi \in \Delta_0$ and for all $w \in W$, $M, w \models \chi$. We build a corresponding EGM $M' = (W, \mathcal{E}_1, \ldots, \mathcal{E}_n, \mathcal{B}_1, \ldots, \mathcal{B}_n, \mathsf{ch}_1, \ldots, \mathsf{ch}_n, \mathcal{V})$ such that:

- For all $i \in AGT$, $\mathcal{E}_i = \mathcal{R}_{\sim_i}$;
- For all $i \in AGT$, $\mathcal{B}_i = \{(w, v) : w \mathcal{E}_i v \text{ and } M, v \models poss_i\}$;
- For all $i \in Agt$, for all $s_i \in S_i$ and for all $w \in W$, $\mathsf{ch}_i(w) = s_i$ iff $w \in \mathcal{I}(choose_i(s_i))$;
- For all $p \in Atm$, $\mathcal{V}(p) = \mathcal{I}(p)$.

Again, by induction on the structure of φ, it is easy to check that φ is satisfied by M'. □

From Propositions 7 and 8 we obtain the following upper boud for the complexity of the satisfiability problem of LEG.

Lemma 5 *The satisfiability problem of LEG relative to the class of epistemic game models is in ExpTime.*

Proof It is a routine task to verify that the problem of global logical consequence in S5-PDL with a finite number of global axioms is reducible to the problem of validity in S5-PDL. In particular, if $\Delta = \{\chi_1, \ldots, \chi_m\}$ we have $\Delta \models_{\text{S5-PDL}} \varphi$ if and only if $\models_{\text{S5-PDL}} [\mathbf{any}^*](\chi_1 \wedge \ldots \chi_m) \rightarrow \varphi$ where **any** is the special program defined as $\mathbf{any} \stackrel{\text{def}}{=} (\bigcup_{i \in Agt} \sim_i \cup \equiv)$. Hence, by Proposition 8 and by the fact that

Δ_0 is finite, it follows that a LEG formula φ is LEG valid if and only if $\models_{\text{S5-PDL}} [\mathbf{any}^*](\bigwedge_{\chi \in \Delta_0} \chi) \rightarrow tr(\varphi)$. Consequently, given the fact that tr is a polynomial reduction of LEG to S5-PDL and given also the fact that the satisfiability problem of S5-PDL is in ExpTime (Proposition 7), it follows that the satisfiability problem of LEG is also in ExpTime. □

Theorem 7 follows from Lemmas 4 and 5. □

References

Asheim, G., & Dufwenberg, M. (2003). Admissibility and common belief. *Games and Economic Behavior*, *42*, 208–234.

Asheim, G. B. (2001). Proper rationalizability in lexicographic beliefs. *International Journal of Game Theory*, *30*, 453–478.

Aumann, R. (1999). Interactive epistemology I: Knowledge. *International Journal of Game Theory*, *28*(3), 263–300.

Baltag, A., & Smets, S. (2009). Talking your way into agreement: Belief merge by persuasive communication. In *Proceedings of the Second Multi-Agent Logics, Languages, and Organisations Federated Workshops (MALLOW)*, vol. 494 of *CEUR Workshop Proceedings*.

Baltag, A., Smets, S., & Zvesper, J. A. (2009). Keep 'hoping' for rationality: a solution to the backward induction paradox. *Synthese*, *169*(2), 301–333.

Battigalli, P., & Siniscalchi, M. (2002). Strong belief and forward induction reasoning. *Journal of Economic Theory*, *106*(2), 356–391.

Board, O. (1998). Belief revision and rationalizability. In: *Proceedings of TARK'98* (pp. 201–213). Morgan Kaufmann.

Board, O. (2002). Knowledge, beliefs, and game-theoretic solution concepts. *Oxford Review of Economic Policy*, *18*, 418–432.

Bonanno, G. (2008). A syntactic approach to rationality in games with ordinal payoffs. In: *Proceedings of LOFT 2008*, Texts in Logic and Games Series (pp. 59–86). Amsterdam University Press.

Börgers, T. (1994). Weak dominance and approximate common knowledge. *Journal of Economic Theory*, *64*, 265–276.

Börgers, T. (1993). Pure strategy dominance. *Econometrica*, *61*(2), 423–430.

Börgers, T., & Samuelson, L. (1992). "Cautious" utility maximization and iterated weak dominance. *International Journal of Game Theory*, *21*, 13–25.

Brandeburger, A. (2007). The power of paradox: Some recent developments in interactive epistemology. *International Journal of Game Theory*, *35*(4), 465–492.

Brandenburger, A. (1992). Lexicographic probabilities and iterated admissibility. In P. Dasgupta, D. Gale, O. Hart, & E. Maskin (Eds.), *Economic analysis of markets and games* (pp. 282–290). Cambridge: MIT Press.

Brandenburger, A., & Dekel, E. (1987). Rationalizability and correlated equilibria. *Econometrica*, *55*, 1391–1402.

Brandenburger, A., Friedenberg, A., & Keisler, J. (2008). Admissibility in games. *Econometrica*, *76*, 307–352.

Cooper, R., De Jong, D., Forsythe, R., & Ross, T. (1993). Forward induction in the battle-of-the-sexes games. *The American Economic Review*, *83*(5), 1303–1316.

Dekel, E., & Fudenberg, D. (1990). Rational behavior with payoff uncertainty. *Journal of Economic Theory*, *52*, 243–267.

Dubois, D., & Prade, H. (1998). Possibility theory: Qualitative and quantitative aspects. In D. Gabbay & P. Smets (Eds.), *Handbook of defeasible reasoning and uncertainty management systems* (Vol. 1, pp. 169–226). Dordrecht: Kluwer.

Fagin, R., Halpern, J., Moses, Y., & Vardi, M. (1995). *Reasoning about knowledge*. Cambridge: MIT Press.

Halpern, J. Y., & Pass, R. (2009). A logical characterization of iterated admissibility. In: Heifetz, A. (ed.). *Proceedings of TARK 2009* (pp. 146–155).

Harel, D., Kozen, D., & Tiurin, J. (2000). *Dynamic logic*. Cambridge: MIT Press.

Kohlberg, E., & Mertens, J. F. (1986). On the strategic stability of equilibria. *Econometrica*, *54*, 1003–1037.

Kraus, S., & Lehmann, D. J. (1988). Knowledge, belief and time. *Theoretical Computer Science*, *58*, 155–174.

Lorini, E. (2013). On the epistemic foundation for iterated weak dominance: An analysis in a logic of individual and collective attitudes. *Journal of Philosophical Logic*, *42*(6), 863–904.

Mas-Colell, M., Winston, A., & Green, J. (1995). *Microeconomic theory*. Oxford: Oxford University Press.

Osborne, M., & Rubinstein, A. (1994). *A course in game theory*. Cambridge: MIT Press.

Pacuit, E., & Roy, O. (forthcoming). Epistemic foundations of game theory. In: *Stanford encyclopedia of philosophy*.

Pearce, D. (1984). Rationalizable strategic behavior and the problem of perfection. *Econometrica*, *52*(4), 1029–1050.

Pearl, J. (1993). From conditional oughts to qualitative decision theory. In D. Heckerman & E. H. Mamdani (Eds.), *Proceedings of UAI'93* (pp. 12–22). Morgan Kaufmann.

Samuelson, L. (1992). Dominated strategies and common knowledge. *Games and Economic Behavior*, *4*, 284–313.

Spohn, W. (1988). Ordinal conditional functions: A dynamic theory of epistemic states. In W. L. Harper & B. Skyrms (Eds.), *Causation in decision, belief change and statistics* (pp. 105–134). Dordrecht: Kluwer.

Stalnaker, R. (1994). On the evaluation of solution concepts. *Theory and Decision*, *37*, 49–73.

Stalnaker, R. (1996). Knowledge, belief and counterfactual reasoning in games. *Economics and Philosophy*, *12*, 133–163.

Stalnaker, R. (1998). Belief revision in games: Forward and backward induction. *Mathematical Social Sciences*, *36*, 31–56.

Tan, T., & Werlang, S. R. C. (1988). The bayesian foundations of solution concepts of games. *Journal of Economic Theory*, *45*, 370–391.

van Ditmarsch, H., van der Hoek, W., & Kooi, B. (2012). Local properties in modal logic. *Artificial Intelligence*, *187–188*, 133–155.

Vardi, M. Y. (1985). The taming of converse: Reasoning about two-way computations. In: *Proc. of the Conference on Logic of Programs*, vol. 193 of *Lecture Notes in Computer Science*. Springer.

Weydert, E. (1994). General belief measures. In R. L. de Mántaras & D. Poole (Eds.) *Proceedings of UAI'94* (pp. 575–582). Morgan Kaufmann.

The logic of acceptance: grounding institutions on agents' attitudes

Emiliano Lorini, Dominique Longin, Benoit Gaudou, Andreas Herzig
Institut de Recherche en Informatique de Toulouse (IRIT)
118 Route de Narbonne, F-31062, Toulouse, France
{lorini,longin,gaudou,herzig}@irit.fr

January 23, 2009

Abstract

In the recent years, several formal approaches to the specification of normative multi-agent systems and artificial institutions have been proposed. The aim of this paper is to advance the state of the art in this area by proposing an approach in which a normative multi-agent system is conceived to be autonomous, in the sense that it is able to create, maintain, and eventually change its own institutions by itself, without the intervention of an external designer in this process. In our approach the existence and the dynamics of an institution (norms, rules, institutional facts, *etc.*) are determined by the (individual and collective) *acceptances* of its members, and its dynamics depends on the dynamics of these acceptances.

In order to meet this objective, we propose the logic $\mathcal{AL}$ (*Acceptance Logic*) in which the acceptance of a proposition by the agents *qua* members of an institution is introduced. Such propositions are true w.r.t. an institutional context and correspond to facts that are instituted in an attitude-dependent way.

The second part of the paper is devoted to the logical characterization of some important notions in the theory of institutions. We provide a formalization of the concept of *constitutive rule*, expressed by a statement of the form "X counts as Y in the context of institution x". Then, we formalize the concepts of obligation and permission (so called *regulative rules*). In our approach constitutive rules and regulative rules of a certain institution are attitude-dependent facts which are grounded on the acceptances of the members of the institution.

Keywords

Modal logic, institutions, acceptance, normative systems, multi-agent systems

Contents

1 Introduction

The problem of devising artificial institutions and modeling their dynamics is a fundamental problem in the multi-agent system (MAS) domain [Dignum and Dignum, 2001]. Following [North, 1990, p. 3], artificial institutions can be conceived as "the rules of the game in a society or the humanly devised constraints that structure agents' interaction". Starting from this concept of institution, many researchers working in the field of normative MAS have been interested in developing models which describe the different kinds of rules and norms that agents have to deal with. In some models of artificial institutions norms are conceived as means to achieve coordination among agents and agents are supposed to comply with them and to obey the authorities of the system [Esteva et al., 2001]. More sophisticated models of institutions leave to the agents' autonomy the decision whether to comply or not with the specified rules and norms of the institution [Ågotnes et al., 2007, Lopez y Lopez et al., 2004]. However, all previous models abstract away from the legislative source of the norms of an institution, and from how institutions are created, maintained and changed by their members. More precisely, while it is widely shared in the MAS field that, in order to face complex and dynamical problems, individual agents must be autonomous, less emphasis is devoted to the fact that MASs themselves for exactly the same reasons should be conceived and designed to be autonomous. In fact, etymologically, autonomous means self-binding ('auto' and 'nomos'), and an autonomous MAS should be the vision of an artificial society that is able to create, maintain, and eventually change its own institutions by itself, without the intervention of the external designer in this process.

The aim of this work is to advance the state of the art on artificial institutions and normative multi-agent systems by proposing a logical model in which the existence and the dynamics of an institution (norms, rules, institutional facts, *etc.*) are determined by the individual and collective attitudes of the agents which identify themselves as members of the institution. In particular, we propose a model in which an institution is grounded on the (individual and collective) *acceptances* of its members, and its dynamics depends on the dynamics of these acceptances. On this aspect we agree with [Mantzavinos et al., 2004], when the authors say that (p. 77):

> "only because institutions are anchored in peoples minds do they ever become behaviorally relevant. The *elucidation of the internal aspect is the crucial step* in adequately explaining the emergence, evolution, and effects of institutions." [Emphasis added].

This relationship between acceptance and institutions has been emphasized in the philosophical doctrine of Legal Positivism [Hart, 1992]. According to Hart, the foundations of a normative system or institution consist of adherence to, or acceptance of, an ultimate rule of recognition by which the validity of any rule of the institution may be evaluated. [1]

[1] In Hart's theory, the rule of recognition is the rule which specifies the ultimate criteria of validity in a legal system.

Other authors working in the field of multi-agent systems have advocated the need for a bottom up approach to the explanation of the origin and the evolution of institutions. According to these authors, institutions and their dynamics should be anchored in the agents' attitudes [Conte et al., 1998, Boella and van der Torre, 2007]. For instance, in agreement with Hart's theory, [Conte et al., 1998] have stressed that the existence of a norm in an institution (but also in a group, organization, *etc.*) depends on the recognition and acceptance of the norm by the members of the institution. In their perspective, agents in a multi-agent system contribute to the enforcement and the propagation of the norm in the social context.

The fundamental concept in our paper is that of acceptance *qua* member of an institution. This notion will be informally presented in Section 2. In Section 3 we will introduce a modal logic (called $\mathcal{AL}$ for *Acceptance Logic*) which enables to reason about acceptances of agents and groups of agents. We call the former *individual acceptances*, and the latter *collective acceptances*. In Section 4 we will study the logical properties of the notion of acceptance and its interactions with classical notions such that of individual (private) belief and that of mutual belief. On the basis of the concept of acceptance *qua* member of an institution, we will specify how a group of agents can create and maintain normative and institutional facts which hold only in an attitude-dependent way. That is, it is up to the agents, and not to the external designer, to support such facts (Section 5). Then, we will distinguish regulative components and non-regulative components of an institution [Searle, 1995] (Section 6). On the one hand, we will formalize the concept of *constitutive rule*, that is, the kind of rules accepted by the members of an institution which express classifications between different concepts and establish the relations between "brute" physical facts and institutional facts within the context of the institution (Section 6.1). Since [Searle, 1995, Searle, 1969] and [Jones and Sergot, 1996], these rules have been expressed in terms of assertions of the form "X counts as Y in the context of institution x" (*e.g.* in the institutional context of US, a piece of paper with a certain shape, color, *etc.* counts as a five-dollar bill). On the other hand, *regulative rules* will be formalized through a notion of obligation and a notion of permission by studying a reduction of deontic logic to the logic of acceptance (Section 6.2). Section 7 will be devoted to show how the logic of acceptance $\mathcal{AL}$ can be appropriately refined in order to capture some essential properties of legal institutions in which a special kind of agents called *legislators* are introduced. We will discuss some general principles which seem adequate for a formal characterization of legal institutions. Finally, in Section 8, we will compare our proposal with related logical works on institutions and normative systems. Special emphasis will be devoted to the comparison between our approach and the modal logic of normative systems and "counts-as" proposed by Grossi et al. [Grossi et al., 2006]. Proofs of the main theorems presented in the paper are collected in the annex.

2 The concept of acceptance

Some conceptual clarifications of the concept of acceptance *qua* member of an institution are needed because of the crucial role it plays in explaining the maintenance of social institutions.

Several authors have emphasized the difference between acceptance and belief as particular kinds of *individual* attitudes. Whereas private beliefs have been studied for decades [Hintikka, 1962] as representative of doxastic mental states, acceptances have only been examined since [Stalnaker, 1984] and since [Cohen, 1992]. Some authors (*e.g.* [Clarke, 1994]) claim that acceptance implies belief (at least to some minimal degree as argued in [Tollefsen, 2003]). On the contrary, in [Stalnaker, 1984] acceptance is considered to be stronger than belief. Although belief and acceptance seem very close, several authors [Bratman, 1992, Cohen, 1992, Tuomela, 2000] have argued for the importance of keeping the two notions independent. We here agree with this point of view (see Section 4.3).

For the aims of this paper we are particularly interested in a particular feature of acceptance, namely the fact that acceptance is context-dependent (on this point see also [Engel, 1998]). In our approach, this feature is directly encoded in the formal definition of acceptance (see Section 3.1). In fact, one can decide (say for prudential reasons) to reason and act by "accepting" the truth of a proposition in a specific context, and reject the very same proposition in a different context. We will explore the role of acceptance in institutional contexts. Institutional contexts are conceived here as rule-governed social practices on the background of which the agents reason. For example, take the case of a game like Clue. The institutional context is the rule-governed social practice which the agents conform to in order to be competent players. On the background of such contexts, we are interested in the agents' attitudes that can be formally captured. In the context of Clue, for instance, an agent accepts that something has happened *qua* player of Clue. The state of acceptance *qua* member of an institution is the kind of acceptance one is committed to when one is "functioning as a member of the institution" [Tuomela, 2002]. In these situations it may happen that the agent's acceptances are in conflict with his/her beliefs. For instance, a lawyer who is trying to defend a client in a murder case accepts *qua* lawyer that his/her client is innocent, even she/he believes the contrary.

There exist others differences between belief and acceptance that are not encoded in our formalization of acceptance. According to [Hakli, 2006], the key difference between belief and acceptance is that the former is aimed at truth, whilst the latter depends on an agent's decision. More precisely, while a belief that p is an attitude constitutively aimed at the truth of p, an acceptance is the output of "a decision to treat p as true in one's utterances and actions" without being *necessarily* (see [Tuomela, 2000] for instance) connected to the actual truth of the proposition.

In the present paper the notion of acceptance *qua* member of an institution is also applied to the collective level named *collective acceptance*. The idea of collective attitudes is developed by Searle [Searle, 1995] among others: without supposing the existence of any collective consciousness, he argues that attitudes can be ascribed to a group of agents and that "the forms of collective intentionality cannot (...) be reduced to something else" [Searle, 1995][2].

Collective attitudes such as collective acceptance have been studied in social philosophy in opposition to the traditional notions of *mutual belief* and *mutual knowl-*

[2] A deeper discussion on this point remains out of the scope of this paper. Some interesting arguments for collective intentionality can be found in [Tollefsen, 2002].

edge that are very popular in artificial intelligence and theoretical computer science [Fagin et al., 1995, Lewis, 1969]. It has been stressed that, while mutual belief is strongly linked to individual beliefs and can be reduced to them, collective attitudes such as collective acceptance cannot be reduced to a composition of individual attitudes. This aspect is particularly emphasized by Gilbert [Gilbert, 1987] who follows Durkheim's non-reductionist view of collective attitudes [Durkheim, 1982]. According to Gilbert, any proper group attitude cannot be defined only as a label on a particular configuration of individual attitudes, as mutual belief is. In [Gilbert, 1989] it is suggested that a collective acceptance of a set of agents C is based on the fact that the agents in C identify themselves as members of a certain group, institution, team, organization, *etc.* and recognize each other as members of the same group, institution, team, organization, *etc.* (this is the view that we adopt in our formalization of acceptance, see Section 3). But mutual belief (and mutual knowledge) does not entail this aspect of mutual recognition and identification with respect to the same social context.

In accordance with [Tuomela, 2002], in this paper we consider collective acceptance with respect to institutional contexts as an attitude that is held by a set of agents *qua* members of the same institution. A collective acceptance held by a set of agents C *qua* members of a certain institution x is the kind of acceptance the agents in C are committed to when they are "functioning together as members of the institution x", that is, when the agents in C identify and recognize each other as members of the institution x. For example, in the context of the institution Greenpeace agents (collectively) accept that their mission is to protect the Earth *qua* members of Greenpeace. The state of acceptance *qua* members of Greenpeace is the kind of acceptance these agents are committed to when they are functioning together as members of Greenpeace, that is, when they identify and recognize each other as members of Greenpeace.

3 Acceptance logic

The logic $\mathcal{AL}$ (*Acceptance Logic*) enables expressing that some agents identify themselves as members of a certain institution and what (groups of) agents accept while functioning together as members of an institution. The principles of $\mathcal{AL}$ clarify the relationships between individual acceptances (acceptances of individual agents) and collective acceptances (acceptances of groups of agents).

3.1 Syntax

The syntactic primitives of $\mathcal{AL}$ are the following: a finite non-empty set of agents AGT; a countable set of atomic formulas ATM; and a finite set of labels $INST$ denoting institutions. We note $2^{AGT\star} = 2^{AGT} \setminus \{\emptyset\}$ the set of all non-empty subsets of AGT. The language $\mathcal{L}_{\mathcal{AL}}$ of the logic $\mathcal{AL}$ is given by the following BNF:

$$\varphi \ ::= \ \bot \mid p \mid \neg\varphi \mid \varphi \vee \varphi \mid \mathcal{A}_{C:x}\varphi$$

where p ranges over ATM, C ranges over $2^{AGT\star}$ and x ranges over $INST$. We define $\wedge$, $\rightarrow$, $\leftrightarrow$ and $\top$ from $\vee$, $\neg$ and $\bot$ in the usual manner.

The formula $\mathcal{A}_{C:x}\varphi$ reads "the agents in C accept that φ while functioning together as members of the institution x". For notational convenience, we write $i{:}x$ instead of $\{i\}{:}x$.

For example, $\mathcal{A}_{C:Greenpeace}protectEarth$ expresses that the agents in C accept that the mission of Greenpeace is to protect the Earth while functioning together as activists in the context of Greenpeace; and $\mathcal{A}_{i:Catholic}PopeInfallibility$ expresses that agent i accepts that the Pope is infallible while functioning as a member of the Catholic Church.

The formula $\mathcal{A}_{C:x}\bot$ has to be read "agents in C are not functioning together as members of the institution x", because we assume that functioning as a member of an institution is, at least in this minimal sense, a rational activity. Conversely, $\neg\mathcal{A}_{C:x}\bot$ has to be read "agents in C are functioning together as members of the institution x". Thus, $\neg\mathcal{A}_{C:x}\bot \wedge \mathcal{A}_{C:x}\varphi$ stands for "agents in C are functioning together as members of the institution x and they accept that φ while functioning together as members of x" or simply "agents in C accept that φ *qua* members of the institution x". Therefore $\neg\mathcal{A}_{C:x}\varphi$ has to be read "agents in C do not accept that φ be true *qua* members of x".

3.2 $\mathcal{AL}$ frames

We use a standard possible worlds semantics. Let the set of all couples of non-empty subsets of agents and institutional contexts be

$$\Delta = 2^{AGT^\star} \times INST.$$

A *frame* of the logic of acceptance $\mathcal{AL}$ ($\mathcal{AL}$ *frame*) is a couple

$$\mathcal{F} = \langle W, \mathscr{A} \rangle$$

where:

- W is a non-empty set of possible worlds;
- $\mathscr{A} : \Delta \rightarrow W \times W$ maps every $C{:}x \in \Delta$ to a relation $\mathscr{A}_{C:x}$ between possible worlds in W.

We note $\mathscr{A}_{C:x}(w) = \{w' : \langle w, w' \rangle \in \mathscr{A}_{C:x}\}$ the set of worlds that the agents in C accept at w while functioning together as members of the institution x.

We impose the following constraints on $\mathcal{AL}$ frames, for any world $w \in W$, institutional context $x \in INST$, and sets of agents $C, B \in 2^{AGT^\star}$ such that $B \subseteq C$:

(**S.1**) if $w' \in \mathscr{A}_{B:y}(w)$ then $\mathscr{A}_{C:x}(w') \subseteq \mathscr{A}_{C:x}(w)$

(**S.2**) if $w' \in \mathscr{A}_{B:y}(w)$ then $\mathscr{A}_{C:x}(w) \subseteq \mathscr{A}_{C:x}(w')$

(**S.3**) if $\mathscr{A}_{C:x}(w) \neq \emptyset$ then $\mathscr{A}_{B:x}(w) \subseteq \mathscr{A}_{C:x}(w)$

(**S.4**) if $w' \in \mathscr{A}_{C:x}(w)$ then $w' \in \bigcup_{i \in C} \mathscr{A}_{i:x}(w')$

(**S.5**) if $\mathscr{A}_{C:x}(w) \neq \emptyset$ then $\mathscr{A}_{B:x}(w) \neq \emptyset$

The constraint **S.1** is a generalized version of transitivity: given two sets of agents C, B such that $B \subseteq C$, if w' is a world that the agents in B accept at w while functioning together as members of the institution y and w'' is a world that the agents in C

accept at w' while functioning together as members of the institution x then, w'' is a world that the agents in C accept at w while functioning together as members of the institution x.

The constraint **S.2** is a generalized version of euclideanity: given two sets of agents C, B such that $B \subseteq C$, if w' is a world that the agents in B accept at w while functioning together as members of the institution y and w'' is a world that the agents in C accept at w while functioning together as members of the institution x then, w'' is a world that the agents in C accept at w' while functioning together as members of the institution x.

The constraint **S.3** is a property of conditional inclusion: given two sets of agents C, B such that $B \subseteq C$, if there exists a world w'' that the agents in C accept at w while functioning together as members of the institution x and w' is a world that the agents in B accept at w while functioning together as members of the institution x then, w' is also a world that the agents in C accept at w while functioning together as members of the institution x.

The constraint **S.4** is a sort of weak reflexivity: if w' is a world that the agents in C accept at w while functioning together as members of the institution x then, there exists some agent $i \in C$ such that w' is a world that agent i accepts at w', while functioning as a member of the institution x.

According to the last constraint **S.5**, given two sets of agents C, B such that $B \subseteq C$, if there exists a world w' that the agents in C accept at w while functioning together as members of the institution x then, there exists a world w'' that the agents in B accept at w while functioning together as members of the institution x.

3.3 $\mathcal{AL}$ models and validity

A *model* of the logic of acceptance $\mathcal{AL}$ ($\mathcal{AL}$ *model*) is a couple

$$\mathcal{M} = \langle \mathcal{F}, \mathscr{V} \rangle$$

where:

- $\mathcal{F}$ is a $\mathcal{AL}$ frame;
- $\mathscr{V} : ATM \rightarrow 2^W$ is valuation function associating a set of possible worlds $\mathscr{V}(p) \subseteq W$ to each atomic formula p of ATM.

Given $\mathcal{M} = \langle W, \mathscr{A}, \mathscr{V} \rangle$ and $w \in W$, the couple $\langle \mathcal{M}, w \rangle$ is a *pointed* $\mathcal{AL}$ *model*. Given a formula φ, we write $\mathcal{M}, w \models \varphi$ and say that φ is *true* at world w in $\mathcal{M}$. The notation $\mathcal{M}, w \not\models \varphi$ means that φ is *false* at world w in $\mathcal{M}$. The truth conditions for the formulas of the logic $\mathcal{AL}$ are:

- $\mathcal{M}, w \not\models \bot$;
- $\mathcal{M}, w \models p$ iff $w \in \mathscr{V}(p)$;
- $\mathcal{M}, w \models \neg\varphi$ iff $\mathcal{M}, w \not\models \varphi$;
- $\mathcal{M}, w \models \varphi \vee \psi$ iff $\mathcal{M}, w \models \varphi$ or $\mathcal{M}, w \models \psi$;
- $\mathcal{M}, w \models \mathcal{A}_{C:x}\varphi$ iff $\mathcal{M}, w' \models \varphi$ for all $w' \in \mathscr{A}_{C:x}(w)$.

A formula φ is *true in a $\mathcal{AL}$ model* $\mathcal{M}$ if and only if $\mathcal{M}, w \models \varphi$ for every world w in $\mathcal{M}$. φ is $\mathcal{AL}$ *valid* (noted $\models_{\mathcal{AL}} \varphi$) if and only if φ is true in all $\mathcal{AL}$ models. φ is $\mathcal{AL}$ *satisfiable* if and only if $\neg\varphi$ is not $\mathcal{AL}$ valid.

3.4 Axiomatization

The axiomatization of $\mathcal{AL}$ is as follows:

(**ProTau**)	All principles of propositional calculus
(**K**)	$\mathcal{A}_{C:x}(\varphi \to \psi) \to (\mathcal{A}_{C:x}\varphi \to \mathcal{A}_{C:x}\psi)$
(**PAccess**)	$\mathcal{A}_{C:x}\varphi \to \mathcal{A}_{B:y}\mathcal{A}_{C:x}\varphi$ if $B \subseteq C$
(**NAccess**)	$\neg\mathcal{A}_{C:x}\varphi \to \mathcal{A}_{B:y}\neg\mathcal{A}_{C:x}\varphi$ if $B \subseteq C$
(**Inc**)	$(\neg\mathcal{A}_{C:x}\bot \wedge \mathcal{A}_{C:x}\varphi) \to \mathcal{A}_{B:x}\varphi$ if $B \subseteq C$
(**Unanim**)	$\mathcal{A}_{C:x}(\bigwedge_{i \in C} \mathcal{A}_{i:x}\varphi \to \varphi)$
(**Mon**)	$\neg\mathcal{A}_{C:x}\bot \to \neg\mathcal{A}_{B:x}\bot$ if $B \subseteq C$
(**MP**)	From $\vdash \varphi$ and $\vdash \varphi \to \psi$ infer $\vdash \psi$
(**Nec**)	From $\vdash \varphi$ infer $\vdash \mathcal{A}_{C:x}\psi$

This axiomatization includes all tautologies of propositional calculus (**ProTau**) and the rule of inference *modus ponens* (**MP**). Axiom **K** and rule of *necessitation* (**Nec**) define a minimal normal modal logic. (See [Chellas, 1980, chap. 4].)

Axioms **PAccess** and **NAccess** express that a group of agents has always access to what is accepted (resp. not accepted) by its supergroups.

Axiom **PAccess** concerns the (positive) access to what is accepted by a supergroup: when the agents in a set C function together as members of the institution x, then for all $B \subseteq C$ the agents in B have access to all facts that are accepted by the agents in C. That is, if the agents in C accept that φ while functioning together as members of the institution x then, while functioning together as members of x, the agents of every subset B of C accept that the agents in C accept that φ.

Axiom **NAccess** concerns the (negative) access to what is not accepted by a supergroup: if the agents in C do not accept that φ while functioning together as members of the institution x then, while functioning together as members of x, the agents of every subset B of C accept that the agents in C do not accept that φ.

Example 1. *Suppose that three agents i, j, k, while functioning together as members of the UK trade union, accept that their mission is to increase teachers' wages, but they do not accept* qua *members of the trade union that their mission is to increase railway workers' wages:*

$\mathcal{A}_{\{i,j,k\}:Union} increaseTeacherWage$ *and* $\neg\mathcal{A}_{\{i,j,k\}:Union} increaseRailwayWage$.

By Axiom ***PAccess*** *we infer that, while functioning as a UK citizen, i accepts that i, j, k accept that their mission is to increase teachers' wages, while functioning together as members of the trade union:*

$$\mathcal{A}_{i:UK}\mathcal{A}_{\{i,j,k\}:Union} increaseTeacherWage.$$

By Axiom ***NAccess*** *we infer that, while functioning as a UK citizen,* i *accepts that* i, j, k *do not accept,* qua *members of the trade union, that their mission is to increase railway workers' wages:*

$$\mathcal{A}_{i:UK} \neg \mathcal{A}_{\{i,j,k\}:Union} increaseRailwayWage.$$

Axiom **Inc** says that, if the agents in C accept that φ *qua* members of x then for every subset B of C the agents in B accept φ while functioning together as members of x. This means that the facts accepted by the agents in C *qua* members of a certain institution x are necessarily accepted by the agents in all of C's subsets with respect to the same institution. Therefore Axiom **Inc** describes the *top down* process leading from C's collective acceptance to the individual acceptances of the agents in C.

Example 2. *Imagine three agents* i*,* j*,* k *that,* qua *players of the game Clue, accept that someone called Mrs. Red, has been killed:*

$$\neg \mathcal{A}_{\{i,j,k\}:Clue} \bot \wedge \mathcal{A}_{\{i,j,k\}:Clue} killedMrsRed.$$

By Axiom ***Inc*** *we infer that also the two agents* i, j*, while functioning as Clue players, accept that someone called Mrs. Red has been killed:*

$$\mathcal{A}_{\{i,j\}:Clue} killedMrsRed.$$

Axiom **Unanim** expresses a unanimity principle according to which the agents in C, while functioning together as members of x, accept that if each of them individually accepts that φ while functioning as a member of x, then φ is the case. This axiom describes the *bottom up* process leading from the individual acceptances of the members of C to the collective acceptance of the group C.

Finally, Axiom **Mon** expresses an intuitive property of monotonicity about institution membership. It says that, if the agents in C are functioning together as members of the institution x then, for every subset B of C, the agents in B are also functioning together as members of the institution x. As emphasized in Section 2, "the agents in C function together as members of institution x" means for us that "the agents in C identify and recognize each other as members of the same institution x". Thus, Axiom **Mon** can be rephrased as follows: if the agents in a set C identify and recognize each other as members of the institution x then, for every subset B of C, the agents in B also identify and recognize each other as members of x.

The following correspondences (in the sense of correspondence theory, see for instance [van Benthem, 2001, Blackburn et al., 2001]) exist between the axioms of the logic $\mathcal{AL}$ and the semantic constraints over $\mathcal{AL}$ frames given in Section 3.2 (see also proof of Theorem 1 in the Annex): Axiom **PAccess** corresponds to the constraint **S.1**, **NAccess** corresponds to **S.2**, **Inc** corresponds to **S.3**, **Unanim** corresponds to **S.4** and **Mon** corresponds to **S.5**.

We call $\mathcal{AL}$ the logic axiomatized by the principles given above: **ProTau**, **K**, **PAccess**, **NAccess**, **Inc**, **Unanim**, **Mon**, **MP**, **Nec**. We write $\vdash_{\mathcal{AL}} \varphi$ if formula φ is a theorem of $\mathcal{AL}$ and $\nvdash_{\mathcal{AL}} \varphi$ if formula φ is not a theorem.

We can prove that $\mathcal{AL}$ is sound and complete with respect to the class of $\mathcal{AL}$ frames.

Theorem 1. $\vdash_{\mathcal{AL}} \varphi$ *if and only if* $\models_{\mathcal{AL}} \varphi$.

By the standard filtration method we can also prove that the logic $\mathcal{AL}$ is decidable.

Theorem 2. *The logic* $\mathcal{AL}$ *is decidable.*

In the following section the properties of the concepts of individual acceptance, collective acceptance and institution membership will be studied. We will also study the relationships between acceptance and belief in a more formal way than in Section 2.

4 General properties

4.1 Properties of acceptance and institution membership

The following theorem highlights some interesting properties of collective acceptance and institution membership.

Theorem 3. *For every $x, y \in INST$ and $B, C \in 2^{AGT^\star}$ such that $B \subseteq C$:*

(3a) $\vdash_{\mathcal{AL}} \mathcal{A}_{C:x} \neg \mathcal{A}_{C:x} \bot$

(3b) $\vdash_{\mathcal{AL}} \mathcal{A}_{C:x} \bigwedge_{i \in C} \neg \mathcal{A}_{i:x} \bot$

(3c) $\vdash_{\mathcal{AL}} \mathcal{A}_{B:y} \mathcal{A}_{C:x} \varphi \leftrightarrow \mathcal{A}_{C:x} \varphi$

(3d) $\vdash_{\mathcal{AL}} \mathcal{A}_{B:y} \neg \mathcal{A}_{C:x} \varphi \leftrightarrow (\mathcal{A}_{B:y} \bot \vee \neg \mathcal{A}_{C:x} \varphi)$

(3e) $\vdash_{\mathcal{AL}} \mathcal{A}_{C:x} (\mathcal{A}_{C:x} \varphi \rightarrow \varphi)$

(3f) $\vdash_{\mathcal{AL}} (\mathcal{A}_{C:x} \bigwedge_{i \in C} \mathcal{A}_{i:x} \varphi) \leftrightarrow \mathcal{A}_{C:x} \varphi$

Theorem 3a expresses a property of institution membership. It says that the agents in a group C, while functioning together as members of the institution x, accept that they are functioning together as members of the institution x. Theorem 3b is another way to express the property of institution membership: it expresses that the agents in a group, while functioning together as members of a certain institution, accept that everyone of them is functioning as a member of the institution.

Example 3. *Suppose that, during a concert, the agents in C are functioning together as members of the Philharmonic Orchestra. Then, according to Theorem 3a, this fact is accepted by the group C. That is, while functioning together as members of the Philharmonic Orchestra, the agents in C accept that they are functioning together as members of the Philharmonic Orchestra: $\mathcal{A}_{C:Orchestra} \neg \mathcal{A}_{C:Orchestra} \bot$. Moreover, they accept that everyone of them is functioning as a member of the Philharmonic Orchestra: $\mathcal{A}_{C:Orchestra} \bigwedge_{i \in C} \neg \mathcal{A}_{i:Orchestra} \bot$.*

Theorem 3c and Theorem 3d together express that a group of agents B can never be wrong in ascribing a collective acceptance to its supergroup C and in recognizing that its supergroup C does not accept something. Furthermore, a group of agents B has always correct access to what is accepted (resp. not accepted) by its supergroups. The right to left direction of Theorem 3c is Axiom **PAccess**. The left to right direction means that, given two sets of agents B and C such that $B \subseteq C$, if the agents in B, while functioning together as members of institution y, accept that the agents in C accept φ while functioning together as members of institution x then, the agents in C accept φ

while functioning together as members of institution x. The right to left direction of Theorem 3d is Axiom **NAccess**. The left to right direction means that, given two sets of agents B and C such that $B \subseteq C$, if the agents in B, while functioning together as members of institution y, accept that the agents in C do not accept φ *qua* members of institution x then, either the agents in B do not function as members of y or the agents in C do not accept φ *qua* members of institution x.

Theorem 3e and Theorem 3f are variants of the unanimity Axiom **Unanim**. Theorem 3e says that for every set of agents C, the agents in C, while functioning together as members of x, accept that if they accept that φ while functioning together as members of x, then φ is the case. Theorem 3f expresses that: if the agents in C, while functioning together as members of x, accept that each of them individually accepts that φ while functioning as a member of x, then the agents in C, while functioning together as members of x, accept that φ is the case.

The following theorem highlights the relationship between the acceptance of a group of agents and the acceptances of its subgroups.

Theorem 4. *For every $x \in INST$ and $C_1, C_2, C_3 \in 2^{AGT}$ such that $C_3 \subseteq C_2 \subseteq C_1$ and $C_3 \neq \emptyset$:*

$$\vdash_{\mathcal{AL}} \mathcal{A}_{C_1:x}(\mathcal{A}_{C_2:x}\varphi \rightarrow \mathcal{A}_{C_3:x}\varphi)$$

Theorem 4 expresses that every group of agents has to accept the principle of inclusion formalized by Axiom **Inc**.

4.2 Discussion around the unanimity principle

Let us consider more in detail the unanimity property of our logic of acceptance expressed by Axiom **Unanim** (and Theorems 3e,3f). This property says that collective acceptances emerge from consensus. This is for us a necessary requirement for a notion of collective acceptance which is valid for all institutions and groups. We did not include stronger principles which explain how a collective acceptance of a group of agents C might be constructed. Nevertheless, one might go further and consider other kinds of principles which are specific to certain institutions and groups.

For example, one might want to extend the analysis to formal (legal) institutions in which special agents with the power to affect the acceptances of the other members of the institution are introduced. In legal institutions, one can formalize the rule according to which all facts that are accepted by the legislators of an institution must be universally accepted by all members of this institution. Suppose that x denotes a legal institution (*e.g.* EU, Association of Symbolic Logic, *etc.*) which has a non-empty set of agents called legislators, noted $Leg(x) \in 2^{AGT^\star}$. (See Section 7 for a precise definition of the function $Leg()$ and a more elaborate analysis of the concepts of legislator and legal institution.) From this, one can formalize a principle stating that everything that the legislators of the legal institution x accept is universally accepted in the legal institution x:

(**Legislators**) $$\mathcal{A}_{C:x}(\bigwedge_{i \in Leg(x)} \mathcal{A}_{i:x}\varphi \rightarrow \varphi)$$

The Principle **Legislators** says that, for every group of agents C, while functioning together as members of the institution x, the agents in C accept that if the legislators of x accept that φ, then φ is the case.

Another interesting principle for the construction of collective acceptance is majority. (In this case, unanimity is not required to obtain a consensus.) This kind of principle applies both to informal and formal institutions. The principle of majority could be introduced as a logical axiom for two specific sets of agents C and B such that $B \subseteq C$ and $|C \setminus B| < |B|$ (*i.e.* B represents the majority of agents in C):

(**Majority**) $$\mathcal{A}_{C:x}(\bigwedge_{i \in B} \mathcal{A}_{i:x}\varphi \rightarrow \varphi)$$

The Principle **Majority** says that, for every group of agents C, while functioning together as members of the institution x, the agents in C accept that if the majority of them accept that φ, then φ is the case. The following example by Pettit [Pettit, 2001] shows how the majority principle would work.

Example 4. *Imagine a three-member court which has to make a judgment on whether a defendant is liable (noted l) for a breach of contract. The three judges i, j and k accept a majority rule to decide on the issue. That is, i, j and k, while functioning together as members of the court, accept that if the majority of them accepts that the defendant is liable (resp. not liable), then the defendant is liable (resp. not liable). Formally, for any B such that $B \subseteq \{i, j, k\}$ and $|B| = 2$ we have:*

$$\mathcal{A}_{\{i,j,k\}:court}(\bigwedge_{i \in B} \mathcal{A}_{i:court}l \rightarrow l) \wedge \mathcal{A}_{\{i,j,k\}:court}(\bigwedge_{i \in B} \mathcal{A}_{i:court}\neg l \rightarrow \neg l)$$

Therefore, if the three judges accept that two of them accept that the defendant is liable, i.e. $\mathcal{A}_{\{i,j,k\}:court}(\mathcal{A}_{i:court}l \wedge \mathcal{A}_{j:court}l)$*, by the Principle **Majority** and Axiom **K** it follows that the three judges have to accept that the judge is liable,* i.e. $\mathcal{A}_{\{i,j,k\}:court}l$*).*

It has to be noted that the previous principle of majority cannot be generalized to all sets of agents without incurring the following very counterintuitive consequence.

Proposition 1. *If we suppose that the Principle **Majority** is valid for any B, C such that $B \subseteq C$ and $|C \setminus B| < |B|$ then, the following consequence is derivable, for $i \neq j$:*

$$(\mathcal{A}_{AGT:x}\mathcal{A}_{\{i,j\}:x}\varphi \wedge \neg\mathcal{A}_{AGT:x}\bot) \rightarrow \mathcal{A}_{AGT:x}\varphi$$

This means that, when the majority principle is generalized to all sets of agents, we can infer that: if all agents, *qua* members of institution x, accept that two of them accept φ while functioning together as members of institution x then, the acceptances of the two agents propagate to all agents in such a way that all agents accept φ *qua* members of institution x.

4.3 Relationships between acceptance and belief

As said in Section 2, there is a large literature about the distinction between belief and acceptance. For us, belief and acceptance are clearly different concepts in several

senses. In this section we focus on the distinction between acceptance, individual belief and mutual belief. Our aim is to provide further clarifications of the concept of acceptance in terms of its relationships with other kinds of agents' attitudes rather than proposing an extension of the logic $\mathcal{AL}$ with individual belief and mutual belief and studying its mathematical properties. Here, we just show how modal operators for belief and mutual belief can be integrated into the logic $\mathcal{AL}$ on the basis of some intuitive interaction principles relating acceptance and belief.

For convenience, we note $Bel_i\varphi$ the formula that reads "the agent i believes that φ is true", and we suppose that belief operators of type Bel_i are defined as usual in a KD45 modal logic [Hintikka, 1962]. Belief operators Bel_i are interpreted in terms of accessibility relations $\mathscr{B}_i$ on the set of possible worlds W. These accessibility relations are supposed to be serial, transitive and euclidean. We write $\mathscr{B}_i(w)$ for the set $\{w' \ : \ \langle w, w'\rangle \in \mathscr{B}_i\}$. $\mathscr{B}_i(w)$ is the set of worlds that are possible according to agent i. The truth condition is:

$$\mathcal{M}, w \models Bel_i\varphi \quad \text{iff} \quad \mathcal{M}, w' \models \varphi \text{ for every } w' \in \mathscr{B}_i(w)$$

Moreover we introduce the notion of mutual belief which has been extensively studied both in the computer science literature [Fagin et al., 1995] and in the philosophical literature [Lewis, 1969]. Given a set of agents $C \subseteq AGT$, $\mathcal{MB}_C\varphi$ reads "there is a mutual belief in C that φ", that is, "everyone in C believes that φ, everyone in C believes that everyone in C believes that φ, everyone in C believes that everyone in C believes that everyone in C believes that φ, and so on". The mutual belief of a set of agents C is interpreted in terms of the transitive closure $\mathscr{B}_C^+$ of the union of the accessibility relations $\mathscr{B}_i$ for every agent $i \in C$, that is:

$$\mathcal{M}, w \models \mathcal{MB}_C\varphi \quad \text{iff} \quad \mathcal{M}, w' \models \varphi \text{ for every } w' \in \mathscr{B}_C^+(w)$$

Let the concept of "everybody in group C believes φ" be defined as follows:

$$E_C\varphi \stackrel{def}{=} \bigwedge_{i \in C} Bel_i\varphi$$

As shown in [Fagin et al., 1995], the following axioms and rules of inference provide a sound and complete axiomatization of the logic of individual belief and mutual belief:

(**KD45$_{\mathbf{Bel}}$**)	All KD45-principles for the operators Bel_i
(**FixPoint**)	$\vdash \mathcal{MB}_C\varphi \leftrightarrow E_C(\varphi \wedge \mathcal{MB}_C\varphi)$
(**InductionRule**)	From $\vdash \varphi \rightarrow E_C(\varphi \wedge \psi)$ infer $\vdash \varphi \rightarrow \mathcal{MB}_C\psi$

The first interesting thing to note is that, although collective acceptance and mutual belief have different natures (see the discussion in Section 2), they share the Fix Point property. The following Theorem 5 highlights this aspect.

Theorem 5.

$$\vdash_{\mathcal{AL}} \mathcal{A}_{C:x}\varphi \leftrightarrow \bigwedge_{i \in C} \mathcal{A}_{i:x}(\varphi \wedge \mathcal{A}_{C:x}\varphi)$$

Nevertheless we cannot argue that our concept of collective acceptance is stronger than the concept of mutual belief, in particular because the **InductionRule** does not hold in $\mathcal{AL}$. This is due to the non-reductionist feature of the collective acceptance: it cannot be reduced to a particular configuration of individual acceptances.

The following two sections are devoted to discuss other interesting relations between acceptance and belief. We will first provide an analysis of the shared aspect of collective acceptance expressed in terms of mutual belief. Then, we will briefly consider the problem of the incompatibility between acceptance and belief.

4.3.1 The shared nature of collective acceptance

As emphasized in the philosophical literature [Gilbert, 1989, Tuomela, 1992], a collective acceptance of the agents in a set C must not be confused with (nor reduced to) the sum of the individual acceptance of the agents in C. On the contrary, when the agents in C accept some fact φ to be true *qua* members of a certain institution, it means that every agent in C declares to the other agents of the group C that she/he is willing to accept φ to be true. This aspect of acceptance can be formally derived by supposing the following two principles relating individual beliefs with collective acceptances.

(**PIntrAccept**)	$\mathcal{A}_{C:x}\varphi \rightarrow Bel_i\mathcal{A}_{C:x}\varphi$	if $i \in C$
(**NegIntrAccept**)	$\neg\mathcal{A}_{C:x}\varphi \rightarrow Bel_i\neg\mathcal{A}_{C:x}\varphi$	if $i \in C$

The first principle says that: if the agents in C accept that φ while functioning together as members of the institution x then, every agent in C believes this. The second principle says that: if the agents in C do not accept φ *qua* members of x then every agent in C believes this.

We can easily prove that, under the previous two principles, collective acceptance is always shared so much that the group C accepts φ if and only if the agents in C mutually believe this. More formally:

Proposition 2. *For any $C{:}x \in \Delta$, the following formulas are derivable from the axiom **D** for belief (following from **KD45**$_{\mathbf{Bel}}$), Axiom **FixPoint** and Rule of inference **InductionRule** for mutual belief, and the interaction Principles **PIntrAccept** and **NegIntrAccept** for acceptance and belief.*

(2a)	$\mathcal{A}_{C:x}\varphi \leftrightarrow \mathcal{MB}_C\mathcal{A}_{C:x}\varphi$
(2b)	$\neg\mathcal{A}_{C:x}\varphi \leftrightarrow \mathcal{MB}_C\neg\mathcal{A}_{C:x}\varphi$

According to Proposition 2a, the agents in C accept that φ while functioning together as members of the institution x if and only if there is a mutual belief in C that they accept that φ while functioning together as members of the institution x. According to Proposition 2b, the agents in C do not accept that φ *qua* members of x if and only if there is a mutual belief in C that they do not accept that φ *qua* members of x. Hence, accepting (resp. not accepting) a proposition while functioning as members of an institution is always a *mutually believed* fact (for the members of the group) which is out in the open and that is used by all the members to reason about each other in the institutional context.

4.3.2 Acceptance and belief might be incompatible

Individual belief and individual acceptance are both private mental attitudes but: an individual belief does not depend on context, whilst an individual acceptance is a context-dependent attitude which is entertained by an agent *qua* member of a given institution. Therefore, an agent can privately disbelieve something she/he accepts while functioning as a member of a given institution. Formally: $Bel_i\varphi \wedge \mathcal{A}_{i:x}\neg\varphi$ may be true. In a similar way, as emphasized in [Tuomela, 1992], a collective acceptance that φ by a group of agents C (*qua* members of a given institution) might be compatible with the fact that none of the agents in C believes that φ (and even that every agent in C believes that $\neg\varphi$). The following example, inspired by [Tuomela, 1992, p. 285], illustrates this point.

Example 5. *At the end of the 80s, the Communist Party of Ruritania accepted that capitalist countries will soon perish (but none of its members really believed so).*

This means that the agents in C accept that capitalist countries will perish ($ccwp$) *qua* members of the Communist Party of Ruritania (CPR) but nobody in C (privately) believes this. Thus, formally: $\neg\mathcal{A}_{C:CPR}\bot \wedge \mathcal{A}_{C:CPR}ccwp \wedge \bigwedge_{i\in C}\neg Bel_i ccwp$.

In the following Section 5 we will show how institutional facts can be grounded on agents' acceptances in such a way that the existence of the former depends on the latter.

5 Truth in an institutional context

Recent theories of institutions [Lagerspetz, 2006, Searle, 1995, Tuomela, 2002] share at least the following two theses.

Performativity: the acceptance that a certain fact is true shared by the members of a certain institution may contribute to the truth of this fact within the context of the institution.

Reflexivity: if a certain fact is true within the context of a certain institution, the acceptance of this fact by the members of the institution is present.

More precisely, a certain fact φ is true within the context of an institution x if and only if the fact φ is accepted to be true by the members of the institution x. Therefore, a *necessary* condition for the existence of a fact within the context of an institution is that this fact is accepted to exist by the members of the institution. Moreover, the acceptance of a certain fact by the members of an institution is a *sufficient* condition for the existence of this fact within the context of the institution.

Example 6. *If the agents,* qua *European citizens, accept a certain piece of paper with a certain shape, color,* etc. *as money, then, within the context of EU, this piece of paper is money (performativity). At the same time, if it is true that a certain piece of paper is money within the context of EU, then the agents* qua *European citizens accept the piece of paper as money (reflexivity).*

Our aim here is to represent in $\mathcal{AL}$ those facts that are true within the context of an institution, that is, to define the concept of truth with respect to an institutional context (*institutional truth*) in a way that respects the previous two principles of reflexivity and performativity. We formalize the notion of institutional truth by means of the operator $[x]$. A formula $[x]\,\varphi$ is read "within the institutional context x, it is the case that φ". We take the latter to be synonymous of "for every set of agents C, the agents in C accept that φ while functioning together as members of the institution x". Formally, for every $x \in INST$:

$$[x]\,\varphi \stackrel{def}{=} \bigwedge_{C \in 2^{AGT\star}} \mathcal{A}_{C:x}\varphi$$

According to our definition, a fact φ is true within the context of institution x if and only if, for every group C, the agents in C accept φ, while functioning together as members of x. Hence the performativity and the reflexivity principles mentioned above are guaranteed.

It is worth noting that this formal definition of truth with respect to an institution is perfectly adequate to characterize informal institutions in which there are no specialized agents called legislators empowered to change the institution itself on behalf of everybody else. It is a peculiar property of informal institutions the fact that they are based on the general consensus of all their members [Coleman, 1990], that is, a certain fact φ is true within the context of an informal institution x if and only if all members of x accept φ to be true. In Section 7 we will show how the operator $[x]$ can be appropriately redefined in order to characterize formal (legal) institution and to distinguish them from informal institutions. For the moment, we just suppose that our model only applies to the basic informal institutions of a society in which no legislator is given.

It is straightforward to prove that $[x]$ is a normal modal operator satisfying Axiom K and the necessitation rule.

Theorem 6. *For every $x \in INST$:*

(6a) $\qquad \vdash_{\mathcal{AL}} \quad [x]\,(\varphi \to \psi) \to ([x]\,\varphi \to [x]\,\psi)$

(6b) $\qquad$ *From* $\vdash_{\mathcal{AL}} \varphi$ *infer* $\vdash_{\mathcal{AL}} [x]\,\varphi$

Nevertheless, institutional operators of type $[x]$ fail to satisfy Axiom 4 and Axiom 5. That is, $[x]\,\varphi \wedge \neg\,[x]\,[x]\,\varphi$ and $\neg\,[x]\,\varphi \wedge \neg\,[x]\,\neg\,[x]\,\varphi$ are satisfiable in the logic $\mathcal{AL}$ for any $x \in INST$. This means that for every institution x, the members of x might accept φ while they do not accept that they accept φ and, it might be the case that the members of x do not accept φ, while they do not accept that they do not accept φ. The operator $[x]$ does to satisfy these two properties because of the restriction imposed on Axioms **PAccess** and **NAccess** according to which, the agents in a group B have access to all facts accepted (resp. not accepted) by the agents in another group C, *only if* B is a subgroup of C. Therefore, in the logic $\mathcal{AL}$, a certain fact φ might be accepted by all groups of members of a certain institution x, while some group of members of x does not have access to the fact that all groups of members of x accept φ. (See Section 8.1 for a discussion about a different point of view.)

The following operator $[Univ]$ is defined in order to express facts which are true in all institutions:

$$[Univ]\,\varphi \stackrel{def}{=} \bigwedge_{x \in INST} [x]\,\varphi$$

where $[Univ]\,\varphi$ is meant to stand for "φ is universally accepted as true". The operator $[Univ]$ is also a normal modal operator satisfying Axiom K and the necessitation rule:

Theorem 7.

(7a) $\quad \vdash_{\mathcal{AL}} \quad [Univ]\,(\varphi \rightarrow \psi) \rightarrow ([Univ]\,\varphi \rightarrow [Univ]\,\psi)$

(7b) $\quad$ *From* $\vdash_{\mathcal{AL}} \varphi$ *infer* $\vdash_{\mathcal{AL}} [Univ]\,\varphi$

The operator $[Univ]$ too fails to satisfy Axiom 4 and Axiom 5. Indeed, $[Univ]\,\varphi \wedge \neg\,[Univ]\,[Univ]\,\varphi$ and $\neg\,[Univ]\,\varphi \wedge \neg\,[Univ]\,\neg\,[Univ]\,\varphi$ are satisfiable in the logic $\mathcal{AL}$. This means that: φ might be universally accepted, while it is not universally accepted that φ is universally accepted and; it might be the case that φ is not universally accepted, while it is not universally accepted that φ is not universally accepted.

In the following section operators of institutional truth of type $[x]$ and the operator of universal truth $[Univ]$ will be used to define the concepts of *constitutive rule* and *regulative rule*. These two concepts are indeed fundamental for a theory of institutions.

6 Constitutive rules and regulative rules

According to many philosophers [Rawls, 1955, Alchourrón and Bulygin, 1971] working on social theory and researchers in the field of normative multi-agent systems [Boella and van der Torre, 2004], institutions are based both on regulative and non-regulative components. In particular, institutions are not only defined in terms of sets of permissions, obligations, and prohibitions (*i.e. norms of conduct* [Bulygin, 1992]) but also in terms of rules which specify and create new forms of behavior and concepts. Several terms such as *constitutive rule* [Searle, 1969, Searle, 1995], *conceptual rule* [Bulygin, 1992] or *determinative rule* [Von Wright, 1963] have been used to identify this non-regulative dimension of institutions. According to Searle for instance "(...) regulative rules regulate antecedently or independently existing forms of behavior (...). But constitutive rules do not merely regulate, they create or define new forms of behavior" [Searle, 1969, p. 33]. In Searle's theory of institutions [Searle, 1969, Searle, 1995], constitutive (*i.e.* non-regulative) rules are expressed by means of "counts-as" statements of the form "X counts as Y in context x" where the context x refers to the institution/normative system in which the rule is specified. As emphasized in [Grossi et al., 2006], "counts-as" statements are used to express classifications and subsumption relations between different concepts, that is, they assert just that a concept X is a subconcept of a concept Y. These classifications are fundamental for establishing the relations between "brute" physical facts and objects on the one hand, and institutional facts and objects on the other hand (*e.g.* money, private property, *etc.*). For example, in the institutional context of Europe, a piece of paper with a certain shape, color, *etc.* (a physical object) counts as a five-euro bill (an institutional object).

6.1 Constitutive rules

From the concept of institutional truth presented above, a notion of constitutive rule of the form "φ counts as ψ in the institutional context x" can be defined in the logic $\mathcal{AL}$. We conceive a constitutive rule as a material implication of the form $\varphi \rightarrow \psi$ in the scope of an operator $[x]$. Thus, "φ counts as ψ in the institutional context x" only if every group of members of institution x accepts that φ entails ψ. Furthermore, we suppose that a constitutive rule is *intrinsically contextual*, which means that the rule is not universally valid while it is accepted by the members of a certain institution. More precisely, we exclude situations in which $[Univ](\varphi \rightarrow \psi)$ is true (*i.e.* situations in which it is universally accepted that φ entails ψ).

In this perspective, "counts-as" statements with respect to a certain institutional context x do not just express that the members of institution x classify φ as ψ in virtue of their acceptances, but also that this classification is proper to the institution, *i.e.* it is not universally accepted that φ entails ψ. (See [Grossi et al., 2006] for a similar perspective.) In this sense, the notion of "counts-as" presented here is aimed at capturing the proper meaning of the term "constitutive rule", that is, a rule which constitutes something new within the context of an institution.

Thus, for every $x \in INST$ the following abbreviation is given:

$$\varphi \overset{x}{\rhd} \psi \overset{def}{=} [x](\varphi \rightarrow \psi) \wedge \neg [Univ](\varphi \rightarrow \psi)$$

where $\varphi \overset{x}{\rhd} \psi$ stands for "φ counts as ψ in the institutional context x".

Example 7. *Let consider the institutional context of gestural language. There exists a constitutive rule in this language according to which, the nodding gesture counts as an endorsement of what the speaker is suggesting,* i.e. $nodding \overset{gesture}{\rhd} yes$. *This means that every group of speakers using gestural language accepts that making the nodding gesture entails endorsing what the speaker is suggesting,* i.e. $[gesture](nodding \rightarrow yes)$*, and there are members of other institutions (*e.g. *different cultural contexts in which the same gesture does not express the same fact) who do not accept this,* i.e. $\neg [Univ](nodding \rightarrow yes)$.

Note that a stronger version of the concept of constitutive rule could be given by supposing that "φ counts as ψ in the institutional context x" if and only if φ entails ψ within the institutional context x, *i.e.* $[x](\varphi \rightarrow \psi)$, and for every institution y, if $y \neq x$ then it is not the case that φ entails ψ within the institutional context y, *i.e.* $\bigwedge_{y \in INST, y \neq x} \neg [y](\varphi \rightarrow \psi)$. The latter condition implies the condition $\neg [Univ](\varphi \rightarrow \psi)$ in the definition of the "counts-as" conditional $\varphi \overset{x}{\rhd} \psi$. This stronger version of the concept of constitutive rule is not analyzed in the present paper.

The following two theorems highlight some valid and invalid properties of "counts-as" operators of the form $\overset{x}{\rhd}$. Similar properties of "counts-as" have been isolated in [Jones and Sergot, 1996] and [Grossi et al., 2006].

The invalidities 8a-8e show that operators $\overset{x}{\rhd}$ do not satisfy reflexivity (invalidity 8a), transitivity (invalidity 8b), strengthening of the antecedent (invalidity 8c), weakening of the consequent (invalidity 8d) and cautious monotonicity (invalidity 8e).

On the contrary, operators $\overset{x}{\rhd}$ satisfy the properties of right logical equivalence (Theorem 9a), left logical equivalence (Theorem 9b), conjunction of the consequents (Theorem 9c), disjunction of the antecedents (Theorem 9d), cumulative transitivity (Theorem 9e).

Theorem 8.

(8a) $\nvdash_{\mathcal{AL}} \varphi \overset{x}{\rhd} \varphi$

(8b) $\nvdash_{\mathcal{AL}} ((\varphi_1 \overset{x}{\rhd} \varphi_2) \wedge (\varphi_2 \overset{x}{\rhd} \varphi_3)) \rightarrow (\varphi_1 \overset{x}{\rhd} \varphi_3)$

(8c) $\nvdash_{\mathcal{AL}} (\varphi_1 \overset{x}{\rhd} \varphi_2) \rightarrow ((\varphi_1 \wedge \varphi_3) \overset{x}{\rhd} \varphi_2)$

(8d) $\nvdash_{\mathcal{AL}} (\varphi_1 \overset{x}{\rhd} \varphi_2) \rightarrow (\varphi_1 \overset{x}{\rhd} (\varphi_2 \vee \varphi_3))$

(8e) $\nvdash_{\mathcal{AL}} ((\varphi_1 \overset{x}{\rhd} \varphi_2) \wedge (\varphi_1 \overset{x}{\rhd} \varphi_3)) \rightarrow ((\varphi_1 \wedge \varphi_2) \overset{x}{\rhd} \varphi_3)$

Theorem 9. *For every $x \in INST$:*

(9a) *From* $\vdash_{\mathcal{AL}} (\varphi_2 \leftrightarrow \varphi_3)$ *infer* $\vdash_{\mathcal{AL}} (\varphi_1 \overset{x}{\rhd} \varphi_2) \leftrightarrow (\varphi_1 \overset{x}{\rhd} \varphi_3)$

(9b) *From* $\vdash_{\mathcal{AL}} (\varphi_1 \leftrightarrow \varphi_3)$ *infer* $\vdash_{\mathcal{AL}} (\varphi_1 \overset{x}{\rhd} \varphi_2) \leftrightarrow (\varphi_3 \overset{x}{\rhd} \varphi_2)$

(9c) $\vdash_{\mathcal{AL}} ((\varphi_1 \overset{x}{\rhd} \varphi_2) \wedge (\varphi_1 \overset{x}{\rhd} \varphi_3)) \rightarrow (\varphi_1 \overset{x}{\rhd} (\varphi_2 \wedge \varphi_3))$

(9d) $\vdash_{\mathcal{AL}} ((\varphi_1 \overset{x}{\rhd} \varphi_2) \wedge (\varphi_3 \overset{x}{\rhd} \varphi_2)) \rightarrow ((\varphi_1 \vee \varphi_3) \overset{x}{\rhd} \varphi_2)$

(9e) $\vdash_{\mathcal{AL}} ((\varphi_1 \overset{x}{\rhd} \varphi_2) \wedge ((\varphi_1 \wedge \varphi_2) \overset{x}{\rhd} \varphi_3)) \rightarrow (\varphi_1 \overset{x}{\rhd} \varphi_3)$

The invalidities 8a-8e are due to the local nature of the "counts-as" conditional $\varphi \overset{x}{\rhd} \psi$. For instance, the fact that $\varphi_1 \overset{x}{\rhd} \varphi_2$ and $\varphi_2 \overset{x}{\rhd} \varphi_3$ are constitutive rules of the institution x does not necessarily entail that $\varphi_1 \overset{x}{\rhd} \varphi_3$ is a constitutive rule of x since it does not necessarily entail $\neg [Univ] (\varphi_1 \rightarrow \varphi_3)$. This is the reason why $\overset{x}{\rhd}$ fails to satisfy transitivity.

Example 8. *In the US state of Texas, "to commit a murder counts as to be punishable by the Death Penalty", and "to be punishable by the Death Penalty counts as to be liable to indictment". As the Death Penalty is not universally accepted in all institutions, both these rules are constitutive rules of Texas,* i.e. $murder \overset{Texas}{\rhd} DeathPenalty$ *and* $DeathPenalty \overset{Texas}{\rhd} indictable$. *From this, it does not follow that it is a constitutive rule of Texas that "to commit a murder counts as to be liable to indictment". Indeed,* $\neg(murder \overset{Texas}{\rhd} indictable)$ *is true. This is due to the fact that "to commit a murder counts as to be liable to indictment" in all countries and institutions, and it is not constitutive of Texas,* i.e. $[Univ] (murder \rightarrow indictable)$.

Similarly, $\overset{x}{\rhd}$ fails to satisfy reflexivity. Indeed, all agents in all possible institutions accept the tautology $\varphi \rightarrow \varphi$ so that "φ counts as φ" cannot be intrinsically contextual

with respect to a certain institution. For similar reasons, strengthening of the antecedent is not a valid property of the operator $\overset{x}{\rhd}$.[3] The following example clarifies this aspect.

Example 9. *It is an accepted custom in the US that a person must leave a tip to the waiter that served him/her at a restaurant. That is, it is a constitutive rule of US that "not leaving a tip to the waiter counts as a violation",* i.e. $\neg leaveTip \overset{US}{\rhd} viol$. *From this, it does not follow that it is a constitutive rule of US that "not leaving a tip to the waiter and not paying the bill counts as a violation". Indeed,* $\neg((\neg leaveTip \wedge \neg payBill) \overset{US}{\rhd} viol)$ *is true. This is because "not leaving a tip and not paying the bill counts as a violation" in all countries and institutions, and it is not constitutive of US,* i.e. $[Univ]((\neg leaveTip \wedge \neg payBill) \rightarrow viol)$.

Discussion

The formal analysis of "counts-as" presented in this section is in agreement with the formal analysis of "counts-as" proposed in [Grossi et al., 2006], where a notion of *proper classificatory rule* is introduced. A proper classificatory rule is represented by the construction $\varphi \Rightarrow_x^{cl+} \psi$ which is meant to stand for "φ counts as ψ in the normative system x". Proper classificatory rules are distinguished by Grossi et al. from (non-proper) *classificatory rules* of type $\varphi \Rightarrow_x^{cl} \psi$. In a way similar to our concept of constitutive rule, *proper classificatory rules* have the specific property of not being universally valid (*i.e.* valid in all institutional contexts). That is, differently from non-proper *classificatory rules*, *proper classificatory rules* are rules which would not hold without the normative system/institution stating them.[4]

Non-proper classificatory rules could be expressed in our logical framework by constructions of the form $[x](\varphi \rightarrow \psi)$, that is, by removing the condition $\neg[Univ](\varphi \rightarrow \psi)$ from the definition of $\varphi \overset{x}{\rhd} \psi$. In agreement with Grossi et al., we would be able to prove that, differently from constitutive rules of the form $\varphi \overset{x}{\rhd} \psi$, such a kind of rules satisfy reflexivity, transitivity, strengthening of the antecedent, weakening of the consequent, and cautious monotonicity. Indeed, the following formulas are all theorems of our logic $\mathcal{AL}$:

Theorem 10.

(10a) $\vdash_{\mathcal{AL}} [x](\varphi \rightarrow \varphi)$

(10b) $\vdash_{\mathcal{AL}} ([x](\varphi_1 \rightarrow \varphi_2) \wedge [x](\varphi_2 \rightarrow \varphi_3)) \rightarrow [x](\varphi_1 \rightarrow \varphi_3)$

(10c) $\vdash_{\mathcal{AL}} [x](\varphi_1 \rightarrow \varphi_2) \rightarrow [x]((\varphi_1 \wedge \varphi_3) \rightarrow \varphi_2)$

(10d) $\vdash_{\mathcal{AL}} [x](\varphi_1 \rightarrow \varphi_2) \rightarrow [x](\varphi_1 \rightarrow (\varphi_2 \vee \varphi_3))$

(10e) $\vdash_{\mathcal{AL}} ([x](\varphi_1 \rightarrow \varphi_2) \wedge [x](\varphi_1 \rightarrow \varphi_3)) \rightarrow [x]((\varphi_1 \wedge \varphi_2) \rightarrow \varphi_3)$

In Section 8.1 a more elaborate and detailed analysis of the logic presented in [Grossi et al., 2006] will be provided and its formal relationships with our logic of acceptance will be studied.

[3] Other authors have defended the idea that strengthening of the antecedent and transitivity should not be valid properties of "counts-as" conditionals (*e.g.* [Gelati et al., 2004]).

[4] See [Grossi et al., 2008] for a refinement of this typology of rules.

Another important aspect to be discussed about our formalization of "counts-as" is the problem of contraposition. Indeed, at the present stage, $\varphi \overset{x}{\rhd} \psi$ is logically equivalent to $\neg\psi \overset{x}{\rhd} \neg\varphi$ which can be counterintuitive in some situations. However, the problem of contraposition could be solved by distinguishing in the language of the logic $\mathcal{AL}$ formulas denoting "brute" physical facts from formulas denoting institutional facts and by imposing that the consequent ψ of a "counts-as" conditional $\varphi \overset{x}{\rhd} \psi$ is always a formula denoting an institutional fact. Under this assumption, if the negation of the antecedent in the "counts-as" conditional is not an institutional fact (*i.e.* formula $\neg\varphi$ does not denote an institutional fact), contraposition is not allowed. That is, $\varphi \overset{x}{\rhd} \psi$ does not imply $\neg\psi \overset{x}{\rhd} \neg\varphi$.[5] It is worth noting that, this distinction between formulas denoting "brute" physical facts and formulas denoting institutional facts would enable us to account for an aspect of "counts-as" that our current formalization is not able to capture, namely: the function of "counts-as" statements of establishing the relations between physical facts and objects on the one hand (the antecedent of the "counts-as"), and institutional facts and objects on the other hand (the consequent of the "counts-as"), *e.g.* a certain piece of paper counts as a five-euro bill.

6.2 Regulative rules

Constitutive rules as defined in the previous Section 6.1 are still not sufficient for a characterization of institutional reality. An institution is indeed connected to a deontic dimension that up to now is still missing in our analysis. This deontic dimension consists in several concepts such as obligation, permission, prohibition, *etc.* which are aimed at regulating agents' behaviors and social interactions within the context of the institution.

In order to capture this deontic dimension of institutions, our logic $\mathcal{AL}$ can be appropriately extended by introducing a *violation* atom $viol$ as in Anderson's reduction of deontic logic to alethic logic [Anderson, 1958] and in dynamic deontic logic [Meyer, 1988]. A similar approach has been recently taken in [Grossi, 2008]. By means of the new formal construct $viol$ we can specify the concepts of obligation and that of permission in a way that respects their being also a kind of attitude-dependent facts holding in a specific institutional context.

As far as obligations are concerned, we introduce operators of the form O_x which are used to specify what is obligatory in the context of a certain institution x:

$$O_x\varphi \overset{def}{=} \neg\varphi \overset{x}{\rhd} viol$$

According to this definition, "φ is obligatory within the institutional context x" if and only if "$\neg\varphi$ counts as a violation within the institutional context x".

Example 10. *The formula* $(driveCar \wedge RightSide) \overset{UK}{\rhd} viol$ *which is equivalent to* $O_{UK}(driveCar \rightarrow \neg RightSide)$ *expresses that in the UK it is obligatory to drive on*

[5] See also [Grossi, 2008] for a different solution on how to solve the problem of contraposition in a normal modal logic of "counts-as".

*the left side of the street (*i.e. *"driving a car on the right side of the street counts as violation in UK").*

As the following theorem highlights, our O_x operators satisfy axiom K (Theorem 11a) and do not allow obligations about tautologies (Theorem 11b).

Theorem 11. *For every $x \in INST$:*

(11a) $\vdash_{\mathcal{AL}} \quad O_x(\varphi \rightarrow \psi) \rightarrow (O_x\varphi \rightarrow O_x\psi)$

(11b) $\vdash_{\mathcal{AL}} \neg O_x \top$

On the contrary, obligation operators do not satisfy the necessitation rule. This is due to the negative condition $\neg\,[Univ]\,(\neg\varphi \rightarrow viol)$ in the definition of $O_x\varphi$. Indeed, in order to have a normal modal operator for obligation, it is sufficient to remove the negative condition $\neg\,[Univ]\,(\varphi \rightarrow \psi)$ from the definition of the "counts-as" conditional $\varphi \overset{x}{\triangleright} \psi$ given in Section 6.1. The following theorem highlights other interesting invalidities of the obligation operators O_x.

Theorem 12.

(12a) $\nvdash_{\mathcal{AL}} \neg O_x \bot$

(12b) $\nvdash_{\mathcal{AL}} O_x\varphi \rightarrow O_x(\varphi \vee \psi)$

(12c) $\nvdash_{\mathcal{AL}} O_x(\varphi \wedge \psi) \rightarrow O_x\varphi$

According to the invalidity 12a, obligation operators do not satisfy the axiom D of Standard Deontic Logic (SDL) [Åqvist, 2002]. For instance, in the logic $\mathcal{AL}$ institutions might be empty, that is, for every $C \in 2^{AGT\star}$, $\mathcal{A}_{C:x}\bot$. If institution x is empty, it does not have any obligation (*i.e.* $O_x\bot$). According to the other two invalidities we have that: if φ is obligatory within the context of institution x then, it is not necessarily the case that φ or ψ is obligatory within the context of the same institution (invalidity 12b) and if φ and ψ are obligatory within the context of institution x then, it is not necessarily the case that φ is obligatory within the context of the same institution (invalidity 12c). Thus, our obligation operators O_x do not incur two classical problems of Standard Deontic Logic which are commonly referred to as "Ross paradox" and "Good Samaritan paradox" [Carmo and Jones, 2002]. On the one hand, it seems rather odd to say that the *obligation to mail a certain letter* entails an *obligation to mail the letter or to burn it* which can be fulfilled simply by burning the letter (something presumably forbidden) ("Ross paradox"). On the other hand, it seems rather odd to say that if *it is obligatory that Mary helps John who has had an accident*, then *it is obligatory that John has an accident* ("Good Samaritan paradox"). Here we do not consider other well-known paradoxes of deontic logic (such as Chisholm paradox for instance) which require an elaborate and detailed analysis of contrary-to-duty obligations and defeasible conditional obligations (on this see [Prakken and Sergot, 1997, Hansen et al., 2007] for instance). Indeed, this issue goes beyond the objectives of the present work.

As far as permissions are concerned we say that "φ is permitted within the institutional context x" (noted $P_x\varphi$) if and only if $\neg\varphi$ is not obligatory within the institutional context x. Formally:

$$P_x\varphi \overset{def}{=} \neg O_x \neg\varphi$$

That is, we define the permission operator in the standard way as the dual of the obligation operator.[6]

Before concluding this section, it is important to stress again that in our approach regulative rules of type $O_x\varphi$ and $P_x\varphi$ as well as constitutive rules of type $\varphi \overset{x}{\rhd} \psi$ of a certain institution are attitude-dependent facts which are grounded on the acceptances of the members of a certain institution.

7 Towards legal institutions

In Section 5 we have supposed that φ is true within the context of institution x if and only if all members of this institution accept φ to be true. At this point, it might be objected that there are facts which are true in an institutional context but only "special" members of the institution are aware of them. For instance, there are laws in every country that are known only by the specialists of the domain (lawyers, judges, members of the Parliament, *etc.*). Aren't these facts true notwithstanding that many members of the institution are not aware of them?

In order to resist to this objection recall that until now our model applied to the basic informal institutions of a society, that is, *rule-governed social practices* [Tuomela, 2002] in which no member with "special" powers is introduced.

It is a peculiar property of informal institutions to be based on general consensus [Coleman, 1990], that is, a certain fact φ is true within the context of an informal institution x if and only if all members of x accept φ to be true. Relative to this restriction, the assumption made in Section 5 is justified because, with respect to informal institutions, there are no specialized agents called legislators empowered to change the institution itself on behalf of everybody else. For instance, in the informal institution of common language, nobody has the power to change the rules for promising. (See [Searle, 1969] for more details.) On the contrary, it is a specificity of legal (formal) institutions to have such specialized agents with special powers to interpret and modify the institution itself. This distinction between informal and formal (legal) institutions has been stressed by many authors working in the field of social and legal theory [Castelfranchi, 2003, North, 1990, Lorini and Longin, 2008, Von Wright, 1963]. Consider for instance the following quotation from Von Wright where the terms *prescription* and *custom* respectively correspond to the terms *formal institution* and *informal institution* used here: "(...) Prescriptions are *given* or *issued* by someone. They 'flow' from or have their 'source' in the will of norm-giver (...) Customs, first of all, are not *given* by any authority to subjects. If we can speak of an authority behind the customs at all this authority would be the community itself" [Von Wright, 1963, p. 7–9].

In the rest of this section we will show how the logic $\mathcal{AL}$ can be appropriately refined in order to move beyond informal institutions and to capture some essential

[6] We do not consider here the classical distinction between *weak permission* and *strong permission* [Alchourrón and Bulygin, 1971, Raz, 1975, Von Wright, 1963]. According to legal theory, a weak permission corresponds to the absence in a normative system of a norm prohibiting φ (this is represented by our permission operator P_x). A strong permission corresponds to the existence in the normative system of an explicit norm, issued by the legislators, according to which φ is permitted. For a logical analysis of the distinction between weak and strong permission see our related work [Lorini and Longin, 2008].

properties of formal (legal) institutions in which legislators are introduced. We will discuss some general principles which seem adequate for a formal characterization of legal institutions. For the sake of simplicity and readability of the article, these principles will not be included in the axiomatization of the logic $\mathcal{AL}$ and their semantic counterparts will not be studied.

In order to distinguish formal from informal institutions, we introduce a total function Leg which assigns a (possibly empty) set of agents to every institution x:

$$Leg : INST \longrightarrow 2^{AGT}$$

$Leg(x)$ denotes the set of legislators of institution x, that is, the set of agents legally responsible over institution x and which are entitled to modify its structure. The function Leg allows distinguishing formal from informal institutions in a simple way. It is indeed reasonable to suppose that informal institutions are those institutions that do not have legislators, that is, x is an informal institution if and only if $Leg(x) = \emptyset$. On the contrary, if $Leg(x) \neq \emptyset$, x is a legal or formal institution. In this sense, the cardinality of $Leg(x)$ provides an important property: it allows us to distinguish between legal institutions and informal institutions.

It seems reasonable to suppose that the legislators of a certain legal institution x must function together as members of institution x. This assumption is expressed by the following principle. For any $x \in INST$ such that $Leg(x) \neq \emptyset$:

$$\neg \mathcal{A}_{Leg(x):x} \bot$$

As emphasized in Section 5, legislators are "special" agents who have the power to affect the acceptances of the other members of the institution. In legal institutions, all facts that are accepted by the legislators must be universally accepted by all members of the institution. In this perspective, legal institutions are characterized by the following principle which explains how the collective acceptance of a set C of members of institution x is affected by the acceptance of the legislators of the institution. For every $C \in 2^{AGT^\star}$ and $x \in INST$ such that $Leg(x) \neq \emptyset$:

(**Legislators**) $$\mathcal{A}_{C:x}(\bigwedge_{i \in Leg(x)} \mathcal{A}_{i:x}\varphi \rightarrow \varphi)$$

According to **Legislators**, for every group of agents C, while functioning together as members of the institution x, the agents in C accept that if the legislators of x accept that φ, then φ is the case. As emphasized in Section 4.1, the Principle **Legislators** can be conceived as an additional specification of how collective acceptances of groups of agents are built within the context of an institution. It is worth noting that **Legislators** is perfectly compatible with the general principle of unanimity of the logic $\mathcal{AL}$ described by Axiom **Unanim** (and the related Theorems 3e, 3f). Indeed, we can reasonably suppose that the members of an institution might accept certain things on the basis of a criterion of unanimity and, at the same time, accept what the legislators accept and decide.[7]

[7]Note that a further principle which seems reasonable for legal institutions is a majority principle for

We conclude by showing how the concept of institutional truth proposed in Section 5 can be appropriately refined in order to deal with legal institutions. Differently from informal institutions, legal institutions do not necessarily depend on the general consensus of all their members. More precisely, if a certain fact φ is true within the context of the legal institution x then, it is not necessarily the case that for every set of agents C, the agents in C accept φ while functioning together as members of the legal institution x. In a legal institution it is sufficient that the legislators accept φ to be true to make it true for the institution. This means that the notion of institutional truth for legal institutions should be defined as follows. For any $x \in INST$ such that $Leg(x) \neq \emptyset$:

$$[x]^L \varphi \stackrel{def}{=} \mathcal{A}_{Leg(x):x}\varphi$$

This means that "within the context of the legal institution x it is the case that φ" if and only if "the legislators of institution x accept that φ".

From the principles of $\mathcal{AL}$ and the definition of the function $Leg()$, it follows that the operators $[x]^L$ are also normal. Moreover, differently from the $[x]$ operators, which adequately characterize the notion of institutional truth for informal institutions, $[x]^L$ operators satisfy axioms 4 and 5 of modal logic, that is: if the legislators of institution x accept φ then, they accept that they accept φ (Theorem 13c); if the legislators of an institution x do not accept φ, then they accept that they do not accept φ (Theorem 13d).[8]

Theorem 13. *For every $x \in INST$:*

(13a) $\quad \vdash_{\mathcal{AL}} \quad [x]^L (\varphi \rightarrow \psi) \rightarrow ([x]^L \varphi \rightarrow [x]^L \psi)$

(13b) $\quad$ *From* $\vdash_{\mathcal{AL}} \varphi$ *infer* $\vdash_{\mathcal{AL}} [x]^L \varphi$

(13c) $\quad \vdash_{\mathcal{AL}} \quad [x]^L \varphi \rightarrow [x]^L [x]^L \varphi$

(13d) $\quad \vdash_{\mathcal{AL}} \quad \neg [x]^L \varphi \rightarrow [x]^L \neg [x]^L \varphi$

It is worth noting that the analysis of constitutive rules and regulative rules proposed in Sections 6.1 and 6.2 could be refined in the light of this distinction between informal and legal institutions. In particular, a new form of "counts-as" and two related concepts of obligation and permission could be defined in terms of the previous operator $[x]^L$. This is in order to characterize a notion of constitutive rule and a notion of regulative rule which apply straightforwardly to the context of legal institutions, and which go beyond the notions of constitutive rule and regulative rule for informal institutions given in Sections 6.1 and 6.2 and based on the operator $[x]$. We postpone this kind of analysis to future works.

---

legislators: the legislators of a certain legal institution x accept that if the majority of them accept φ, then φ is true. This should be conceived as a particular case of the majority principle discussed in Section 4.1. Formally, for any $x \in INST$ such that $Leg(x) \neq \emptyset$, if $B \subseteq Leg(x)$ and $|Leg(x) \setminus B| < |B|$ (*i.e.* B represents the majority of the legislators of the institution x) then: $\mathcal{A}_{Leg(x):x}(\bigwedge_{i \in B} \mathcal{A}_{i:x}\varphi \rightarrow \varphi)$.

[8] Note that the operator $[x]$ is stronger than the operator $[x]^L$, that is, $[x]\,\varphi$ implies $[x]^L\,\varphi$.

8 Comparison with other logical approaches to normative systems

In the following two sections our logic $\mathcal{AL}$ will be compared with two approaches to normative systems and institutions which have been recently proposed in the multi-agent system domain.

8.1 Embedding Grossi et al.'s logic of "counts-as" into $\mathcal{AL}$

Because of the interesting formal similarities, we will first compare $\mathcal{AL}$ with the modal logic of normative systems proposed in [Grossi et al., 2006], henceforth abbreviated $\mathcal{GMD}$ logic.

In the $\mathcal{GMD}$ logic a set of contexts CXT denoting normative systems is introduced. $\mathcal{GMD}$ logic is based on a set of modal operators $[\![x]\!]$ (one for every context x in CXT). Operators $[\![x]\!]$ are similar to our operators $[x]$ defined in Section 5.[9] A formula $[\![x]\!]\,\varphi$ approximately stands for "in the institutional context/normative system x it is the case that φ". It is supposed that CXT contains a special context $Univ$, where the operator $[\![Univ]\!]$ is used for denoting facts which universally hold. We note $CXT_0 = CXT \setminus \{Univ\}$. The language of the $\mathcal{GMD}$ logic is given by the following BNF:

$$\varphi \ ::= \ \bot \mid p \mid \neg\varphi \mid \varphi \vee \varphi \mid [\![x]\!]\,\varphi$$

where p ranges over ATM and x ranges over CXT. $\wedge$, $\rightarrow$, $\leftrightarrow$ and $\top$ are defined from $\vee$, $\neg$ and $\bot$ in the usual manner.

As noted in Section 6.1, operators $[\![x]\!]$ and $[\![Univ]\!]$ are exploited in Grossi et al.'s logic to define contextual conditionals called *proper classificatory rules*, noted $\varphi \Rightarrow_x^{cl+} \psi$, which are an abbreviation of $[\![x]\!]\,(\varphi \rightarrow \psi) \wedge \neg\,[\![Univ]\!]\,(\varphi \rightarrow \psi)$ and which read "φ counts as ψ in the normative system x". The construction $\varphi \Rightarrow_x^{cl+} \psi$ is similar to our $\varphi \overset{x}{\rhd} \psi$.

The most striking difference between our logic of acceptance $\mathcal{AL}$ and the $\mathcal{GMD}$ logic is that in the logic $\mathcal{AL}$ the contextual operators $[x]$ are built on the notion of collective acceptance, whereas in the $\mathcal{GMD}$ logic the contextual operators $[\![x]\!]$ are given as primitive operators.

Frames of the $\mathcal{GMD}$ logic are called *multi-context frames*. A multi-context frame has the following form:

$$\mathcal{F}^{\mathcal{GMD}} = \langle S, \{S_x\}_{x \in CXT_0} \rangle$$

where:

- S is a set of possible worlds;
- $\{S_x\}_{x \in CXT_0}$ is a family of subsets of S, one for every institutional context $x \in CXT_0$.

A *multi-context model* is a tuple

$$\mathcal{M}^{\mathcal{GMD}} = \langle \mathcal{F}^{\mathcal{GMD}}, \pi \rangle$$

where:

[9] Here we use the notation $[\![x]\!]$ in order to distinguish their operators from ours.

- $\mathcal{F}^{\mathcal{GMD}}$ is a multi-context frame;
- $\pi : ATM \rightarrow 2^S$ is a valuation function associating a set of possible worlds $\pi(p) \subseteq S$ to each atomic formula p of ATM.

The truth conditions for formulas of the $\mathcal{GMD}$ logic are just standard for contradiction, atomic propositions, negation and disjunction. The following are the truth conditions for $[\![x]\!]\,\varphi$ and $[\![Univ]\!]\,\varphi$.

- $\mathcal{M}^{\mathcal{GMD}}, w \models [\![x]\!]\,\varphi$ iff $\mathcal{M}, w' \models \varphi$ for all $w' \in S_x$;
- $\mathcal{M}^{\mathcal{GMD}}, w \models [\![Univ]\!]\,\varphi$ iff $\mathcal{M}, w' \models \varphi$ for all $w' \in S$.

A formula φ is *true in a $\mathcal{GMD}$ model* $\mathcal{M}^{\mathcal{GMD}}$ iff $\mathcal{M}^{\mathcal{GMD}}, w \models \varphi$ for every world w in $\mathcal{M}^{\mathcal{GMD}}$. φ is $\mathcal{GMD}$ *valid* (noted $\models_{\mathcal{GMD}} \varphi$) if and only if φ is true in all $\mathcal{GMD}$ models. φ is $\mathcal{GMD}$ *satisfiable* iff $\neg\varphi$ is not $\mathcal{GMD}$ valid.

The $\mathcal{GMD}$ logic is axiomatized by the following principles, where x and y denote elements of the set CXT_0 :

(**ProTau**)	All principles of propositional calculus
($\mathbf{K}_{[\![x]\!]}$)	$[\![x]\!]\,(\varphi \rightarrow \psi) \rightarrow ([\![x]\!]\,\varphi \rightarrow [\![x]\!]\,\psi)$
($\mathbf{K}_{[\![Univ]\!]}$)	$[\![Univ]\!]\,(\varphi \rightarrow \psi) \rightarrow ([\![Univ]\!]\,\varphi \rightarrow [\![Univ]\!]\,\psi)$
($\mathbf{4}_{[\![x]\!],[\![y]\!]}$)	$[\![x]\!]\,\varphi \rightarrow [\![y]\!]\,[\![x]\!]\,\varphi$
($\mathbf{5}_{[\![x]\!],[\![y]\!]}$)	$\neg\,[\![x]\!]\,\varphi \rightarrow [\![y]\!]\,\neg\,[\![x]\!]\,\varphi$
($\mathbf{4}_{[\![Univ]\!]}$)	$[\![Univ]\!]\,\varphi \rightarrow [\![Univ]\!]\,[\![Univ]\!]\,\varphi$
($\mathbf{5}_{[\![Univ]\!]}$)	$\neg\,[\![Univ]\!]\,\varphi \rightarrow [\![Univ]\!]\,\neg\,[\![Univ]\!]\,\varphi$
($\mathbf{T}_{[\![Univ]\!]}$)	$[\![Univ]\!]\,\varphi \rightarrow \varphi$
($\subseteq_{[\![Univ]\!],[\![x]\!]}$)	$[\![Univ]\!]\,\varphi \rightarrow [\![x]\!]\,\varphi$
(**MP**)	From $\vdash \varphi$ and $\vdash \varphi \rightarrow \psi$ infer $\vdash \psi$
($\mathbf{Nec}_{[\![x]\!]}$)	From $\vdash \varphi$ infer $\vdash [\![x]\!]\,\psi$
($\mathbf{Nec}_{[\![Univ]\!]}$)	From $\vdash \varphi$ infer $\vdash [\![Univ]\!]\,\psi$

We write $\vdash_{\mathcal{GMD}} \varphi$ if formula φ is a theorem of $\mathcal{GMD}$.

Axiom $\mathbf{K}_{[\![x]\!]}$ and Rule $\mathbf{Nec}_{[\![x]\!]}$ express that the operators $[\![x]\!]$ are normal modal operators. Axioms $\mathbf{K}_{[\![Univ]\!]}$, $\mathbf{4}_{[\![Univ]\!]}$, $\mathbf{5}_{[\![Univ]\!]}$, $\mathbf{T}_{[\![Univ]\!]}$ and the rule of inference $\mathbf{Nec}_{[\![Univ]\!]}$ express that the universal modality $[\![Univ]\!]$ is defined in the modal logic system S5. According to the Axioms $\mathbf{4}_{[\![x]\!],[\![y]\!]}$ and $\mathbf{5}_{[\![x]\!],[\![y]\!]}$, truth and falsehood in institutional contexts/normative systems are absolute because they remain invariant even if they are evaluated from another institutional context/normative system. This means that every normative system y has full access to all facts which are true in a different normative system x. In our view, these two principles are criticizable because they rely on a strong assumption of perfect information, *i.e.* a normative system has perfect information about the facts that are true in the other normative systems. Axiom $\subseteq_{[\![Univ]\!],[\![x]\!]}$ expresses the relationship between the universal modality and the contextual modalities.

In [Grossi et al., 2006] it is proved that the $\mathcal{GMD}$ logic is sound and complete with respect to the class of $\mathcal{GMD}$ frames.

It is easy to show that the principles of the acceptance logic $\mathcal{AL}$ given in Section 3 are not sufficient to derive the principles of the $\mathcal{GMD}$ logic. In particular, Axioms $\mathbf{4}_{[\![x]\!],[\![y]\!]}$, $\mathbf{5}_{[\![x]\!],[\![y]\!]}$, $\mathbf{4}_{[\![Univ]\!]}$, $\mathbf{5}_{[\![Univ]\!]}$ and $\mathbf{T}_{[\![Univ]\!]}$ are not derivable in $\mathcal{AL}$.

In order to embed $\mathcal{GMD}$ we need to slightly modify the properties of the logic $\mathcal{AL}$. On the one hand, we need to generalize Axioms **PAccess** and **NAccess** by supposing that they **also** hold for the case $B \not\subseteq C$. This is in order to infer the formulas $[x]\,\varphi \rightarrow [y]\,[x]\,\varphi$ and $\neg\,[x]\,\varphi \rightarrow [y]\,\neg\,[x]\,\varphi$ in the augmented logic $\mathcal{AL}$. Thus, we need to assume that, given two arbitrary sets of agents B and C, the agents in B have access to all facts that the agents in C accept (do not accept), while functioning together as members of a certain institution x. On the other hand, we need to add the principle $[Univ]\,\varphi \rightarrow \varphi$ to the logic $\mathcal{AL}$. The way to embed the $\mathcal{GMD}$ logic into our logic $\mathcal{AL}$ is illustrated in the following paragraph.

An embedding of $\mathcal{GMD}$ logic. Let us slightly modify the logic of acceptance $\mathcal{AL}$ in order to provide a correct embedding of $\mathcal{GMD}$. We call $\mathcal{AL}^+$ the modified logic of acceptance.

$\mathcal{AL}^+$ has the same language as $\mathcal{AL}$ (see Section 3.1). $\mathcal{AL}^+$ frames are tuples $\mathcal{F} = \langle W, \mathscr{A} \rangle$ where W and $\mathscr{A}$ are defined as for $\mathcal{AL}$ frames, except that the constraints **S.1** and **S.2** given in Section 3.2 are supposed to hold also for the case $B \not\subseteq C$ and the following additional constraint **S.6** is imposed. That is, for any world $w \in W$, institutional context $x \in INST$, and sets of agents $C, B \in 2^{AGT\star}$ we suppose:

(**S.1'**) if $w' \in \mathscr{A}_{B:y}(w)$ then $\mathscr{A}_{C:x}(w') \subseteq \mathscr{A}_{C:x}(w)$

(**S.2'**) if $w' \in \mathscr{A}_{B:y}(w)$ then $\mathscr{A}_{C:x}(w) \subseteq \mathscr{A}_{C:x}(w')$

Furthermore, for any world $w \in W$ we suppose:

(**S.6**) $\exists C \in 2^{AGT\star}$, $\exists x \in INST$ such that $w \in \mathscr{A}_{C:x}(w)$

The axiomatization of $\mathcal{AL}^+$ is given by the axiom schemes and rules of inference of $\mathcal{AL}$, except that an Axiom corresponding to the Axiom $\mathbf{T}_{[\![Univ]\!]}$ of the $\mathcal{GMD}$ logic is added, and the Axioms **PAccess** and **NAccess** of the logic $\mathcal{AL}$ are generalized in such a way that they also for hold for the case $B \not\subseteq C$. That is, for any sets of agents $C, B \in 2^{AGT\star}$, we suppose:

(**PAccess**$^+$) $\mathcal{A}_{C:x}\varphi \rightarrow \mathcal{A}_{B:y}\mathcal{A}_{C:x}\varphi$

(**NAccess**$^+$) $\mathcal{A}_{C:x}\varphi \rightarrow \mathcal{A}_{B:y}\neg\mathcal{A}_{C:x}\varphi$

Furthermore, we suppose:

($\mathbf{T}_{Univ}$) $[Univ]\,\varphi \rightarrow \varphi$

Axioms **PAccess**$^+$ and **NAccess**$^+$ respectively correspond to the semantic constraints **S.1'** and **S.2'**, whilst Axiom $\mathbf{T}_{Univ}$ corresponds to the semantic constraint **S.6**.

The definitions of validity and satisfiability in $\mathcal{AL}^+$ are given accordingly. We write $\models_{\mathcal{AL}^+} \varphi$ if formula φ is *valid* in all $\mathcal{AL}^+$ models satisfying the semantic constraints **S.3**, **S.4**, **S.5** given in Section 3.2 and the constraints **S.1'**, **S.2'**, **S.6** given here. We call $\mathcal{AL}^+$ the logic axiomatized by Axiom $\mathbf{T}_{Univ}$ and the principles of the logic $\mathcal{AL}$ (Section 3.4), where Axioms **PAccess** and **NAccess** are generalized to **PAccess**$^+$ and **NAccess**$^+$. We write $\vdash_{\mathcal{AL}^+} \varphi$ if formula φ is a theorem of $\mathcal{AL}^+$.

We can prove that $\mathcal{AL}^+$ as well is sound and complete. More precisely:

Theorem 14. *$\vdash_{\mathcal{AL}^+} \varphi$ if and only if $\models_{\mathcal{AL}^+} \varphi$.*

Consider the following translation tr from $\mathcal{GMD}$ to the new logic $\mathcal{AL}^+$:

- $tr(\bot) = \bot$
- $tr(p) = p$
- $tr(\neg\varphi) = \neg tr(\varphi)$
- $tr(\varphi \vee \psi) = tr(\varphi) \vee tr(\psi)$
- $tr([\![x]\!]\,\varphi) = [x]\,tr(\varphi)$
- $tr([\![Univ]\!]\,\varphi) = [Univ]\,tr(\varphi)$,

As the following Theorem 15 shows, tr is a correct embedding of the $\mathcal{GMD}$ logic.

Theorem 15. *Let $INST = CXT$ and φ be a formula of the $\mathcal{GMD}$ logic. Then, φ is $\mathcal{GMD}$ satisfiable if and only if $tr(\varphi)$ is $\mathcal{AL}^+$ satisfiable.*

REMARK. It is worth noting that $\mathcal{GMD}$ logic can also be embedded into the variant of $\mathcal{AL}$ with legislators presented in Section 7 by the translations $tr([\![x]\!]\,\varphi) = [x]^L\,\varphi$ and $tr([\![Univ]\!]\,\varphi) = [Univ]^L\,\varphi$, after defining

$$[Univ]^L\,\varphi \stackrel{def}{=} \bigwedge_{x \in INST} [x]^L\,\varphi.$$

To obtain a correct embedding of the $\mathcal{GMD}$ logic, it is sufficient to add to $\mathcal{AL}$ the three axioms $[x]^L\,\varphi \rightarrow [y]^L\,[x]^L\,\varphi$, $\neg\,[x]^L\,\varphi \rightarrow [y]^L\,\neg\,[x]^L\,\varphi$ and $[Univ]^L\,\varphi \rightarrow \varphi$ and the two corresponding semantic constraints over $\mathcal{AL}$ frames:

$$\text{if } w' \in \mathscr{A}_{Leg(y):y}(w) \text{ then } \mathscr{A}_{Leg(x):x}(w') = \mathscr{A}_{Leg(x):x}(w), \text{ and}$$
$$\exists x \in INST \text{ such that } w \in \mathscr{A}_{Leg(x):x}(w).$$

8.2 A conceptual comparison with Boella & van der Torre's model

The formal approach to institutions and normative systems proposed by Boella & van der Torre [Boella and van der Torre 2004, 2007] is similar in some respect to ours. Here we just provide a *conceptual* comparison between the two approaches. We are not able to provide a more *technical* comparison. Indeed, our formalism based on modal logic and their formalism based on input-output logic [Makinson and van der Torre, 2000] are too different to be compared in the fashion followed in Section 8.1.

Boella & van der Torre emphasize the relevance of the concept of acceptance for a formal model of institutions. In their model, individual agents *accept* a norm, together

with its associated sanctions and rewards, when they recognize that this norm serves to achieve their desires and believe that the other agents will conform to it. According to them, for a norm to be really effective it must be respected due to its acceptance, and not only to the fear of sanctions. Although they take the concept of acceptance into consideration, they do not analyze it in detail. In particular, in their model there is no distinction between individual acceptance and collective acceptance. On the contrary, this distinction is fundamental in our $\mathcal{AL}$ logic in which we clarify the relationships between individual acceptances and collective acceptances and we provide an explanation of how the collective acceptance of a group of agents C is built from the individual acceptances of the agents in C.

Moreover, in Boella & van der Torre's approach, normative systems and institutions are conceived as agents and mental attitudes such as beliefs and goals are ascribed to them. Differently from them, we do not claim that institutions can be conceived as agents. In our approach, we only defend the idea that the institutional reality is built on the top of the agents' attitudes. In particular, we claim that institutions are grounded on the individual and collective *acceptances* of their members and groups of members, and their dynamics depend on the dynamics of these acceptances.

9 Conclusion

We have presented in this article a logic of acceptance and applied it to the analysis of institutions. Our logic of acceptance allows to express that agents accept something to be true *qua* members of a certain institution. Given the properties of this demystified notion of acceptance, we have provided an analysis of the kind of attitude-dependent facts which are typical of institutions. We have formalized the concept of constitutive rule expressed by statements of the form "X counts as Y in the context of institution x". Then, we have introduced a notion of obligation and a notion permission with respect to an institutional context (*i.e.* so-called regulative rules). While constitutive rules and regulative rules are usually defined from the external perspective of a normative system or institution, in the present work we have anchored these rules in the agents' acceptances.

Directions for future research are manifold. For instance, future works will be devoted to integrate modalities expressing agents' goals and preferences, such as the ones provided in [Cohen and Levesque, 1990], into the logical framework presented in this paper. This is in order to investigate the decision to join (resp. not to join) a given institution and the related decision to accept (resp. not to accept) the norms of the institution with its associated sanctions and rewards. These kinds of decisions are indeed influenced by the inconsistency between the agent's goals and the current norms and rules of the institution. For instance, if the agent's goals conflict with the norms proclaimed by the legislators then, the agent will probably decide not to join the institution.

Another interesting topic to be investigated in future works is the dynamics of individual and collective acceptances in institutional contexts. We have already started to study this topic in a recent work [Herzig et al., 2008]. The idea is to extend the logic of acceptance $\mathcal{AL}$ by events of type $x!\varphi$ and corresponding dynamic operators

of the form $[x!\varphi]$. A formula $[x!\varphi]\psi$, means that ψ is true after every announcement of formula φ in the context of institution x. Operators of type $[x!\varphi]$, which are similar to the operators of announcements in dynamic epistemic logic [Baltag et al., 1998, Gerbrandy and Groeneveld, 1997, van Ditmarsch et al., 2007], express that the members of an institution x learn that φ is true in that institution in such a way that their acceptances, *qua* members of institution x, are updated. Such operators can also be used to describe how the acceptances of the members of institution x change, after that a certain norm (e.g. obligation, permission) is *issued* or *promulgated* within the context of this institution.

References

[Ågotnes et al., 2007] Ågotnes, T., van der Hoek, W., Rodriguez-Aguilar, J., Sierra, C., and Wooldridge, M. (2007). On the logic of normative systems. In *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence (IJCAI'07)*, pages 1181–1186. AAAI Press.

[Alchourrón and Bulygin, 1971] Alchourrón, C. and Bulygin, E. (1971). *Normative systems*. Springer, New York.

[Anderson, 1958] Anderson, A. (1958). A reduction of deontic logic to alethic modal logic. *Mind*, 22:100–103.

[Åqvist, 2002] Åqvist, L. (2002). Deontic Logic. In Gabbay, D. and Guenthner, F., editors, *Handbook of Philosophical Logic*, volume 8, pages 147–264. Kluwer Academic Publishers, 2nd edition.

[Baltag et al., 1998] Baltag, A., Moss, L., and Solecki, S. (1998). The logic of public announcements, common knowledge and private suspicions. In *Proceedings of the Seventh Conference on Theoretical Aspects of Rationality and Knowledge (TARK'98)*, pages 43–56, San Francisco, CA. Morgan Kaufmann Publishers Inc.

[van Benthem, 2001] van Benthem, J. (2001). Correspondence theory. In Gabbay, D. and Guenthner, F., editors, *Handbook of Philosophical Logic*, volume 3, pages 325–408. Kluwer Academic Publishers, 2nd edition.

[Blackburn et al., 2001] Blackburn, P., de Rijke, M., and Venema, Y. (2001). *Modal Logic*. Cambridge University Press, Cambridge.

[Boella and van der Torre, 2004] Boella, G. and van der Torre, L. (2004b). Regulative and constitutive norms in normative multiagent systems. In *Proceedings of the 9th International Conference on Principles on Principles of Knowledge Representation and Reasoning(KR 2004)*, pages 255–266. AAAI Press.

[Boella and van der Torre, 2007] Boella, G. and van der Torre, L. (2007). Norm negotiation in multiagent systems. *International Journal of Cooperative Information Systems*, 16(1):97–122.

[Bratman, 1992] Bratman, M. E. (1992). Practical reasoning and acceptance in context. *Mind*, 101(401):1–15.

[Bulygin, 1992] Bulygin, E. (1992). On norms of competence. *Law and Philosophy*, 11(3):201–216.

[Carmo and Jones, 2002] Carmo, J. and Jones, A. (2002). Deontic logic and contrary-to-duties. In Gabbay, D. and Guenthner, F., editors, *Handbook of Philosophical Logic*, volume 8, pages 265–343. Kluwer Academic Publishers, 2nd edition.

[Castelfranchi, 2003] Castelfranchi, C. (2003). Formalising the informal? Dynamic social order, bottom-up social control, and spontaneous normative relations. *Journal of Applied Logic*, 1(1-2):47–92.

[Chellas, 1980] Chellas, B. F. (1980). *Modal Logic: an Introduction*. Cambridge University Press, Cambridge.

[Clarke, 1994] Clarke, D. (1994). Does acceptance entail belief? *American Philosophical Quarterly*, 31(2):145–155.

[Cohen, 1992] Cohen, L. J. (1992). *An essay on belief and acceptance*. Oxford University Press, New York, USA.

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

[Coleman, 1990] Coleman, J. (1990). *Foundations of Social Theory*. Harvard University Press, Cambridge.

[Conte et al., 1998] Conte, R., Castelfranchi, C., and Dignum, F. (1998). Autonomous norm acceptance. In *Intelligent Agents V (ATAL'98)*, volume 1555 of *LNCS*, pages 99–112, Berlin. Springer Verlag.

[Dignum and Dignum, 2001] Dignum, V. and Dignum, F. (2001). Modelling agent societies: Coordination frameworks and institutions. In Brazdil, P. and Jorge, A., editors, *Proceedings of the Tenth Portuguese Conference in Artificial Intelligence (EPIA'01)*, volume 2258 of *LNAI*, pages 191–204, Berlin. Springer-Verlag.

[van Ditmarsch et al., 2007] van Ditmarsch, H., van der Hoek, W., and Kooi, B. (2007). *Dynamic Epistemic Logic*, volume 337 of *Synthese Library Series*. Springer.

[Durkheim, 1982] Durkheim, E. (1982). *The rules of Sociological Method*. Free Press, New York. first published in French in 1895.

[Engel, 1998] Engel, P. (1998). Believing, holding true, and accepting. *Philosophical Explorations*, 1(2):140–151.

[Esteva et al., 2001] Esteva, M., Padget, J., and Sierra, C. (2001). Formalizing a language for institutions and norms. In *Intelligent Agents VIII (ATAL'01)*, volume 2333 of *LNAI*, pages 348–366, Berlin. Springer Verlag.

[Fagin et al., 1995] Fagin, R., Halpern, J., Moses, Y., and Vardi, M. (1995). *Reasoning about Knowledge*. MIT Press, Cambridge.

[Gaudou et al., 2008] Gaudou, B., Longin, D., Lorini, E., and Tummolini, L. (2008). Anchoring Institutions in Agents' Attitudes: Towards a Logical Framework for Autonomous MAS. In Padgham, L. and Parkes, D. C., editors, *Proceedings of the Seventh International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS'08)*, pages 728–735. ACM Press.

[Gelati et al., 2004] Gelati, J., Rotolo, A., Sartor, G., and Governatori, G. (2004). Normative autonomy and normative co-ordination: Declarative power, representation, and mandate. *Artificial Intelligence and Law*, 12(1-2):53–81.

[Gerbrandy and Groeneveld, 1997] Gerbrandy, J. and Groeneveld, W. (1997). Reasoning about information change. *Journal of Logic, Language, and Information*, 6:147–196.

[Gilbert, 1987] Gilbert, M. (1987). Modelling collective belief. *Synthese*, 73(1):185–204.

[Gilbert, 1989] Gilbert, M. (1989). *On Social Facts*. Routledge, London and New York.

[Goldblatt, 1992] Goldblatt, R. (1992). *Logics of Time and Computation, 2nd edition*. CSI Lecture Notes, Stanford, California.

[Grossi, 2008] Grossi, D. (2008). Pushing Anderson's envelope: the modal logic of ascription. In *Proceedings of the 9th International Conference on Deontic Logic in Computer Science (DEON'08)*, number 5076 in LNAI, pages 263–277. Springer Verlag.

[Grossi et al., 2006] Grossi, D., Meyer, J.-J. C., and Dignum, F. (2006). Classificatory aspects of counts-as: An analysis in modal logic. *Journal of Logic and Computation*, 16(5):613–643.

[Grossi et al., 2008] Grossi, D., Meyer, J.-J. C., and Dignum, F. (2008). The many faces of counts-as: A formal analysis of constitutive rules. *Journal of Applied Logic*, 6(2):192–217.

[Hakli, 2006] Hakli, P. (2006). Group beliefs and the distinction between belief and acceptance. *Cognitive Systems Research*, 7:286–297.

[Hansen et al., 2007] Hansen, J., Pigozzi, G., and van der Torre, L. (2007). Ten philosophical problems in deontic logic. In Boella, G., van der Torre, L., and Verhagen, H., editors, *Normative Multi-agent Systems*, number 07122 in Dagstuhl Seminar Proceedings. Internationales Begegnungs- und Forschungszentrum fuer Informatik (IBFI), Schloss Dagstuhl, Germany.

[Hart, 1992] Hart, H. L. A. (1992). *The concept of law*. Clarendon Press, Oxford. new edition.

[Herzig et al., 2008] Herzig, A., de Lima, T., and Lorini, E. (2008). What do we accept after an announcement? In Meyer, J.-J. and Broersen, J., editors, *Proceedings of the First Workshop on Knowledge Representation for Agents and Multi-Agent Systems (KRAMAS 2008)*, pages 81–94.

[Hintikka, 1962] Hintikka, J. (1962). *Knowledge and Belief: An Introduction to the Logic of the Two Notions*. Cornell University Press, Ithaca.

[Jones and Sergot, 1996] Jones, A. and Sergot, M. J. (1996). A formal characterization of institutionalised power. *Journal of the IGPL*, 4:429–445.

[Lagerspetz, 2006] Lagerspetz, E. (2006). Institutional facts, performativity and false beliefs. *Cognitive Systems Research*, 7(2-3):298–306.

[Lewis, 1969] Lewis, D. K. (1969). *Convention: a philosophical study*. Harvard University Press, Cambridge.

[Lopez y Lopez et al., 2004] Lopez y Lopez, F., Luck, M., and d'Inverno, M. (2004). Normative agent reasoning in dynamic societies. In *Proceedings of the Third International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS'04)*, pages 732–739. ACM Press.

[Lorini and Longin, 2008] Lorini, E. and Longin, D. (2008). A logical account of institutions: from acceptances to norms via legislators. In Brewka, G. and Lang, J., editors, *Proceedings of the Eleventh International Conference on Principles on Principles of Knowledge Representation and Reasoning (KR 2008)*, pages 38–48. AAAI Press.

[Makinson and van der Torre, 2000] Makinson, D. and van der Torre, L. (2000). Input-output logics. *Journal of Philosophical Logic*, 29:383–408.

[Mantzavinos et al., 2004] Mantzavinos, C., North, D., and Shariq, S. (2004). Learning, institutions, and economic performance. *Perspectives on Politics*, 2:75–84.

[Meyer, 1988] Meyer, J. J. (1988). A different approach to deontic logic: Deontic logic viewed as a variant of dynamic logic. *Notre Dame Journal of Formal Logic*, 29(1):109–136.

[North, 1990] North, D. (1990). *Institutions, Institutional Change, and Economic Performance*. Cambridge University Press, Cambridge.

[Pettit, 2001] Pettit, P. (2001). Deliberative democracy and the discursive dilemma. *Philosophical Issues*, 11:268–99.

[Prakken and Sergot, 1997] Prakken, H. and Sergot, M. (1997). Dyadic deontic logic and contrary-to-duty obligations. In Nute, D., editor, *Defeasible Deontic Logic*, pages 223–262. Synthese Library.

[Rawls, 1955] Rawls, J. (1955). Two concepts of rules. *The Philosophical Review*, 64:3–32.

[Raz, 1975] Raz, J. (1975). *Practical reason and norms*. Hutchinson, London.

[Sahlqvist, 1975] Sahlqvist, H. (1975). Completeness and correspondence in the first and second order semantics for modal logics. In Kanger, S., editor, *Proceedings of the 3rd Scandinavian Logic Symposium*, volume 82 of *Studies in Logic*.

[Searle, 1969] Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language*. Cambridge University Press, New York.

[Searle, 1995] Searle, J. R. (1995). *The Construction of Social Reality*. The Free Press, New York.

[Stalnaker, 1984] Stalnaker, R. (1984). *Inquiry*. MIT Press, Cambridge.

[Tollefsen, 2002] Tollefsen, D. P. (2002). Challenging epistemic individualism. *Protosociology*, 16:86–117.

[Tollefsen, 2003] Tollefsen, D. P. (2003). Rejecting rejectionism. *Protosociology*, 18–19:389–405.

[Tuomela, 1992] Tuomela, R. (1992). Group beliefs. *Synthese*, 91:285–318.

[Tuomela, 2000] Tuomela, R. (2000). Belief versus Acceptance. *Philosophical Explorations*, 2:122–137.

[Tuomela, 2002] Tuomela, R. (2002). *The Philosophy of Social Practices: A Collective Acceptance View*. Cambridge University Press, Cambridge.

[Von Wright, 1963] Von Wright, G. H. (1963). *Norm and Action*. Routledge and Kegan, London.

A Annex: proofs of some theorems

This Annex contains some selected proofs of the theorems presented in the paper.

Proof of Theorem 1

Axiom **K** and rule of inference **Nec** define a minimal normal modal logic. Thus, they do not have an associated semantic constraint. It is a routine task to check that the Axioms **PAccess**, **NAccess**, **Inc**, **Unanim** and **Mon** of the logic $\mathcal{AL}$ correspond to their semantic counterparts **S.1**-**S.5** over $\mathcal{AL}$ models. In particular, the following correspondences exist between the axioms of the logic $\mathcal{AL}$ and the semantic constraints over $\mathcal{AL}$ frames.

- Axioms **PAccess** corresponds to the constraint **S.1**.
- Axiom **NAccess** corresponds to the constraint **S.2**.
- Axiom **Inc** corresponds to the constraint **S.3**.
- Axiom **Unanim** corresponds to the constraint **S.4**.
- Axiom **Mon** corresponds to the constraint **S.5**.

It is a routine, too, to check that all of axioms of the logic $\mathcal{AL}$ are in the Sahlqvist class, for which a general completeness result exists.
(See [Sahlqvist, 1975, Blackburn et al., 2001].)

Proof of Theorem 2

For notational convenience, we will use the following abbreviation in the proof:

$$\widehat{\mathcal{A}}_{C:x}\varphi \stackrel{def}{=} \neg\mathcal{A}_{C:x}\neg\varphi$$

We have to prove that if φ is $\mathcal{AL}$ *satisfiable* then it is satisfiable in a finite $\mathcal{AL}$ model.

Suppose that $\mathcal{M} = \langle W, \mathscr{A}, \mathscr{V}\rangle$ is a $\mathcal{AL}$ model which satisfies φ. Our aim is to build a finite $\mathcal{AL}$ model which satisfies φ. To do this, we use a filtration method [Blackburn et al., 2001, Goldblatt, 1992].

Let us introduce the following definition.

Definition 1. *A set of formulas Σ is closed under subformulas (cus) if for all formulas φ, φ': if $\varphi \vee \varphi' \in \Sigma$ then so are φ and φ'; if $\neg\varphi' \in \Sigma$ then so is φ; for any $x \in INST$ and $C \in 2^{AGT\star}$ if $\mathcal{A}_{C:x}\varphi \in \Sigma$ then $\varphi \in \Sigma$.*

Let us now consider an arbitrary finite set of formulas Σ_φ which is *closed under subformulas* and which contains φ. From Σ_φ we define the set Σ^+_φ as follows.

Σ^+_φ is defined as the smallest superset of Σ_φ such that:

1. for all $x, y \in INST$ and $C, B \in 2^{AGT\star}$, if $\mathcal{A}_{C:x}\varphi \in \Sigma^+_\varphi$ then $\mathcal{A}_{B:y}\varphi \in \Sigma^+_\varphi$;

2. for all $x \in INST$ and $C \in 2^{AGT\star}$, if $\mathcal{A}_{C:x}\varphi \in \Sigma_\varphi^+$ then $\neg\mathcal{A}_{C:x}\varphi \in \Sigma_\varphi^+$;

3. for all $x \in INST$ and $C \in 2^{AGT\star}$, $\mathcal{A}_{C:x}\bot \in \Sigma_\varphi^+$;

4. $\bot \in \Sigma_\varphi^+$.

The following proposition follows straightforwardly due to the fact that the sets AGT and $INST$ are supposed to be finite.

Proposition 3. *Σ_φ^+ is finite and closed under subformulas.*

We define the relation $\leftrightsquigarrow$ between the worlds in W of the model $\mathcal{M}$. For every two worlds $w, v \in W$:

- $w \leftrightsquigarrow v$ iff for all $\varphi \in \Sigma_\varphi^+$, $\mathcal{M}, w \models \varphi$ iff $\mathcal{M}, v \models \varphi$.

For every world $w \in W$, we note $|w|$ the equivalence class of world w of $\mathcal{M}$ with respect to $\leftrightsquigarrow$. Moreover, let $W_{\Sigma_\varphi^+} = \{|w| \mid w \in W\}$.

Now, we have to build a filtrated model $\mathcal{M}^f = \langle W^f, \mathscr{A}^f, \mathscr{V}^f \rangle$ of the model $\mathcal{M}$.

Definition 2. *We define $\mathcal{M}^f$ as follows.*

A. $W^f = W_{\Sigma_\varphi^+}$*;*

B. *for every $B \in 2^{AGT\star}$ and $x \in INST$, $|v| \in \mathscr{A}^f_{B:x}(|w|)$ if and only if:*

1. $\forall \mathcal{A}_{B:x}\varphi \in \Sigma_\varphi^+$*, if* $\mathcal{M}, w \models \mathcal{A}_{B:x}\varphi$ *then* $\mathcal{M}, v \models \varphi$*;*

2. $\forall y \in INST$ *and* $\forall C \in 2^{AGT\star}$*, if* $B \subseteq C$ *then:*
$\forall \mathcal{A}_{C:y}\varphi \in \Sigma_\varphi^+$*, if* $\mathcal{M}, w \models \mathcal{A}_{C:y}\varphi$ *then* $\mathcal{M}, v \models \mathcal{A}_{C:y}\varphi$*;*

3. $\forall y \in INST$ *and* $\forall C \in 2^{AGT\star}$*, if* $B \subseteq C$ *then:*
$\forall \widehat{\mathcal{A}}_{C:y}\varphi \in \Sigma_\varphi^+$*, if* $\mathcal{M}, w \models \widehat{\mathcal{A}}_{C:y}\varphi$ *then* $\mathcal{M}, v \models \widehat{\mathcal{A}}_{C:y}\varphi$*;*

4. $\forall C \in 2^{AGT\star}$*, if* $B \subseteq C$ *then:*
$\forall \mathcal{A}_{C:x}\varphi, \widehat{\mathcal{A}}_{C:x}\top \in \Sigma_\varphi^+$*, if* $\mathcal{M}, w \models \widehat{\mathcal{A}}_{C:x}\top \wedge \mathcal{A}_{C:x}\varphi$ *then* $\mathcal{M}, v \models \varphi$*;*

5. $\exists i \in B$ *such that* $\forall \mathcal{A}_{i:x}\varphi \in \Sigma_\varphi^+$*, if* $\mathcal{M}, v \models \mathcal{A}_{i:x}\varphi$ *then* $\mathcal{M}, v \models \varphi$*.*

C. $\mathscr{V}^f(p) = \{|w| \mid \mathcal{M}, w \models p\}$*, for all propositional atoms in* Σ_φ^+*.*

It is straightforward to prove that the model $\mathcal{M}^f$ is indeed a filtration of $\mathcal{M}$ through Σ_φ^+.

Lemma 1. *$\mathcal{M}^f$ is a filtration of $\mathcal{M}$ through Σ_φ^+.*

The next step consists in proving that $\mathcal{M}^f$ is a $\mathcal{AL}$ model.

Lemma 2. *$\mathcal{M}^f$ is a $\mathcal{AL}$ model.*

Proof. We have to prove that the model $\mathcal{M}^f$ satisfies the five semantic constraints **S.1**-**S.5** over $\mathcal{AL}$ models.

Let us start with constraint **S.1**. We have to prove that the following condition holds in $\mathcal{M}^f$ for any $x, y \in INST$ and $C, B \in 2^{AGT\star}$ such that $B \subseteq C$:

- if $|w'| \in \mathscr{A}^f_{B:x}(|w|)$ and $|w''| \in \mathscr{A}^f_{C:y}(|w'|)$ then $|w''| \in \mathscr{A}^f_{C:y}(|w|)$.

Suppose $|w'| \in \mathscr{A}^f_{B:x}(|w|)$ and $w'' \in \mathscr{A}^f_{C:y}(|w'|)$, where $B \subseteq C$. We have to prove that $|w''| \in \mathscr{A}^f_{C:y}(|w|)$. By Definition 2, the latter is equivalent to:

1. $\forall \mathcal{A}_{C:y}\varphi \in \Sigma^+_\varphi$, if $\mathcal{M}, w \models \mathcal{A}_{C:y}\varphi$ then $\mathcal{M}, w'' \models \varphi$;

2. $\forall z \in INST$ and $\forall D \in 2^{AGT\star}$, if $C \subseteq D$ then:
$\forall \mathcal{A}_{D:z}\varphi \in \Sigma^+_\varphi$, if $\mathcal{M}, w \models \mathcal{A}_{D:z}\varphi$ then $\mathcal{M}, w'' \models \mathcal{A}_{D:z}\varphi$;

3. $\forall z \in INST$ and $\forall D \in 2^{AGT\star}$, if $C \subseteq D$ then:
$\forall \widehat{\mathcal{A}}_{D:z}\varphi \in \Sigma^+_\varphi$, if $\mathcal{M}, w \models \widehat{\mathcal{A}}_{D:z}\varphi$ then $\mathcal{M}, w'' \models \widehat{\mathcal{A}}_{D:z}\varphi$;

4. $\forall D \in 2^{AGT\star}$, if $C \subseteq D$ then:
$\forall \mathcal{A}_{D:y}\varphi, \widehat{\mathcal{A}}_{D:y}\top \in \Sigma^+_\varphi$, if $\mathcal{M}, w \models \widehat{\mathcal{A}}_{D:y}\top \wedge \mathcal{A}_{D:y}\varphi$ then $\mathcal{M}, w'' \models \varphi$;

5. $\exists i \in C$ such that $\forall \mathcal{A}_{i:y}\varphi \in \Sigma^+_\varphi$, if $\mathcal{M}, w'' \models \mathcal{A}_{i:y}\varphi$ then $\mathcal{M}, w'' \models \varphi$.

So, to prove **S.1** we just need to prove that the previous items **1**-**5** are consequences of $|w'| \in \mathscr{A}^f_{B:x}(|w|)$ and $|w''| \in \mathscr{A}^f_{C:y}(|w'|)$ when $B \subseteq C$.

Item 1. Suppose $\mathcal{A}_{C:y}\varphi \in \Sigma^+_\varphi$ and $\mathcal{M}, w \models \mathcal{A}_{C:y}\varphi$. As $B \subseteq C$ and $|w'| \in \mathscr{A}^f_{B:x}(|w|)$, it follows that $\mathcal{M}, w'' \models \varphi$.

Item 2. Take an arbitrary D such that $C \subseteq D$ and an arbitrary $z \in INST$. As $B \subseteq C$, we have $B \subseteq D$. Suppose $\mathcal{A}_{D:z}\varphi \in \Sigma^+_\varphi$ and $\mathcal{M}, w \models \mathcal{A}_{D:z}\varphi$. As $|w'| \in \mathscr{A}^f_{B:x}(|w|)$, it follows that $\mathcal{M}, w' \models \mathcal{A}_{D:z}\varphi$. As $w'' \in \mathscr{A}^f_{C:y}(|w'|)$, we conclude that $\mathcal{M}, w'' \models \mathcal{A}_{D:z}\varphi$.

Item 3. Take an arbitrary D such that $C \subseteq D$ and an arbitrary $z \in INST$. As $B \subseteq C$, we have $B \subseteq D$. Suppose $\widehat{\mathcal{A}}_{D:z}\varphi \in \Sigma^+_\varphi$ and $\mathcal{M}, w \models \widehat{\mathcal{A}}_{D:z}\varphi$. As $|w'| \in \mathscr{A}^f_{B:x}(|w|)$, it follows that $\mathcal{M}, w' \models \widehat{\mathcal{A}}_{D:z}\varphi$. As $w'' \in \mathscr{A}^f_{C:y}(|w'|)$, we conclude that $\mathcal{M}, w'' \models \widehat{\mathcal{A}}_{D:z}\varphi$.

Item 4. Take an arbitrary D such that $C \subseteq D$. As $B \subseteq C$, we have $B \subseteq D$. Suppose $\mathcal{A}_{D:y}\varphi, \widehat{\mathcal{A}}_{D:y}\top \in \Sigma^+_\varphi$ and $\mathcal{M}, w \models \mathcal{A}_{D:y}\varphi \wedge \widehat{\mathcal{A}}_{D:y}\top$. As $|w'| \in \mathscr{A}^f_{B:x}(|w|)$, it follows that $\mathcal{M}, w' \models \mathcal{A}_{D:y}\varphi \wedge \widehat{\mathcal{A}}_{D:y}\top$. As $w'' \in \mathscr{A}^f_{C:y}(|w'|)$, we conclude that $\mathcal{M}, w'' \models \varphi$.

Item 5. This item follows straightforwardly from the fact $w'' \in \mathscr{A}^f_{C:y}(|w'|)$.

This proves that **S.1** holds.

Let us now consider constraint **S.2**. We have to prove that the following condition holds in $\mathcal{M}^f$ for any $x, y \in INST$ and $C, B \in 2^{AGT\star}$ such that $B \subseteq C$:

- if $|w'| \in \mathscr{A}^f_{B:x}(|w|)$ and $w'' \in \mathscr{A}^f_{C:y}(|w|)$ then $w'' \in \mathscr{A}^f_{C:y}(|w'|)$.

Suppose $|w'| \in \mathscr{A}^f_{B:x}(|w|)$ and $w'' \in \mathscr{A}^f_{C:y}(|w|)$, where $B \subseteq C$. We have to prove that $|w''| \in \mathscr{A}^f_{C:y}(|w'|)$. By Definition 2, the latter is equivalent to:

1. $\forall \mathcal{A}_{C:y}\varphi \in \Sigma^+_\varphi$, if $\mathcal{M}, w' \models \mathcal{A}_{C:y}\varphi$ then $\mathcal{M}, w'' \models \varphi$;

2. $\forall z \in INST$ and $\forall D \in 2^{AGT\star}$, if $C \subseteq D$ then:
$\forall \mathcal{A}_{D:z}\varphi \in \Sigma^+_\varphi$, if $\mathcal{M}, w' \models \mathcal{A}_{D:z}\varphi$ then $\mathcal{M}, w'' \models \mathcal{A}_{D:z}\varphi$;

3. $\forall z \in INST$ and $\forall D \in 2^{AGT\star}$, if $C \subseteq D$ then:
$\forall \widehat{\mathcal{A}}_{D:z}\varphi \in \Sigma^+_\varphi$, if $\mathcal{M}, w' \models \widehat{\mathcal{A}}_{D:z}\varphi$ then $\mathcal{M}, w'' \models \widehat{\mathcal{A}}_{D:z}\varphi$;

4. $\forall D \in 2^{AGT\star}$, if $C \subseteq D$ then:
$\forall \mathcal{A}_{D:y}\varphi, \widehat{\mathcal{A}}_{D:y}\top \in \Sigma^+_\varphi$, if $\mathcal{M}, w' \models \widehat{\mathcal{A}}_{D:y}\top \wedge \mathcal{A}_{D:y}\varphi$ then $\mathcal{M}, w'' \models \varphi$;

5. $\exists i \in C$ such that $\forall \mathcal{A}_{i:y}\varphi \in \Sigma^+_\varphi$, if $\mathcal{M}, w'' \models \mathcal{A}_{i:y}\varphi$ then $\mathcal{M}, w'' \models \varphi$.

So, to prove **S.2** we just need to prove that items **1**-**5** are consequences of $|w'| \in \mathscr{A}^f_{B:x}(|w|)$ and $w'' \in \mathscr{A}^f_{C:y}(|w|)$.

Item 1. Suppose $\mathcal{A}_{C:y}\varphi \in \Sigma^+_\varphi$ and $\mathcal{M}, w' \models \mathcal{A}_{C:y}\varphi$. By construction of Σ^+_φ we have $\widehat{\mathcal{A}}_{C:y}\neg\varphi \in \Sigma^+_\varphi$. As $|w'| \in \mathscr{A}^f_{B:x}(|w|)$ and $B \subseteq C$, it follows that $\mathcal{M}, w \models \mathcal{A}_{C:y}\varphi$. As $|w''| \in \mathscr{A}^f_{C:y}(|w|)$, we conclude that $\mathcal{M}, w'' \models \varphi$.

Item 2. Take an arbitrary D such that $C \subseteq D$ and an arbitrary $z \in INST$. As $B \subseteq C$, we have $B \subseteq D$. Suppose $\mathcal{A}_{D:z}\varphi \in \Sigma^+_\varphi$ and $\mathcal{M}, w' \models \mathcal{A}_{D:z}\varphi$. By construction of Σ^+_φ we have $\widehat{\mathcal{A}}_{D:z}\neg\varphi \in \Sigma^+_\varphi$. As $|w'| \in \mathscr{A}^f_{B:x}(|w|)$ and $B \subseteq D$, it follows that $\mathcal{M}, w \models \mathcal{A}_{D:z}\varphi$. As $|w''| \in \mathscr{A}^f_{C:y}(|w|)$, we conclude that $\mathcal{M}, w'' \models \mathcal{A}_{D:z}\varphi$.

Item 3. Take an arbitrary D such that $C \subseteq D$ and an arbitrary $z \in INST$. As $B \subseteq C$, we have $B \subseteq D$. Suppose $\widehat{\mathcal{A}}_{D:z}\varphi \in \Sigma^+_\varphi$ and $\mathcal{M}, w' \models \widehat{\mathcal{A}}_{D:z}\varphi$. By construction of Σ^+_φ we have $\mathcal{A}_{D:z}\neg\varphi \in \Sigma^+_\varphi$. As $|w'| \in \mathscr{A}^f_{B:x}(|w|)$ and $C \subseteq D$, it follows that $\mathcal{M}, w \models \widehat{\mathcal{A}}_{D:z}\varphi$. As $|w''| \in \mathscr{A}^f_{C:y}(|w|)$, we conclude that $\mathcal{M}, w'' \models \widehat{\mathcal{A}}_{D:z}\varphi$.

Item 4. Take an arbitrary D such that $C \subseteq D$. As $B \subseteq C$, we have $B \subseteq D$. Suppose $\mathcal{A}_{D:y}\varphi, \widehat{\mathcal{A}}_{D:y}\top \in \Sigma^+_\varphi$ and $\mathcal{M}, w' \models \mathcal{A}_{D:y}\varphi \wedge \widehat{\mathcal{A}}_{D:y}\top$. By construction of Σ^+_φ we have $\mathcal{A}_{D:y}\bot, \widehat{\mathcal{A}}_{D:y}\neg\varphi \in \Sigma^+_\varphi$. As $|w'| \in \mathscr{A}^f_{B:x}(|w|)$ and $C \subseteq D$, it follows that $\mathcal{M}, w \models \mathcal{A}_{D:y}\varphi \wedge \widehat{\mathcal{A}}_{D:y}\top$. As $|w''| \in \mathscr{A}^f_{C:y}(|w|)$, we conclude that $\mathcal{M}, w'' \models \varphi$.

Item 5. This item follows straightforwardly from the fact $w'' \in \mathscr{A}^f_{C:y}(|w|)$.

This proves that **S.2** holds.

As a next step we have to prove the model $\mathcal{M}^f$ satisfies the semantic condition **S.3**. That is, we have to prove that for any $x \in INST$ and $C, B \in 2^{AGT\star}$ such that $B \subseteq C$:

- if $\mathscr{A}^f_{C:x}(|w|) \neq \emptyset$ then $\mathscr{A}^f_{B:x}(|w|) \subseteq \mathscr{A}^f_{C:x}(|w|)$.

The following proposition is needed to prove that $\mathcal{M}^f$ satisfies the condition **S.3**.

Proposition 4. *For every $x \in INST$ and $C \in 2^{AGT\star}$, if $\mathscr{A}^f_{C:x}(|w|) \neq \emptyset$ then $\exists w \in |w|$ such that $\mathcal{M}, w \models \widehat{\mathcal{A}}_{C:x}\top$.*

Proof. Let us suppose that $\mathscr{A}^f_{C:x}(|w|) \neq \emptyset$, and $\mathcal{M}, w \models \mathcal{A}_{C:x}\bot$ for all $w \in |w|$. We are going to show that the two facts are inconsistent.

Condition $\mathscr{A}^f_{C:x}(|w|) \neq \emptyset$ implies that $\exists |w'| \in W^f$ such that: if $\mathcal{A}_{C:x}\varphi \in \Sigma^+_\varphi$ then, if $\mathcal{M}, w \models \mathcal{A}_{C:x}\varphi$ then $\mathcal{M}, w' \models \varphi$. As we have $\mathcal{A}_{C:x}\bot \in \Sigma^+_\varphi$ (by construction of Σ^+_φ) and we have supposed $\mathcal{M}, w \models \mathcal{A}_{C:x}\bot$, we can infer that $\mathcal{M}, w' \models \bot$. $\square$

Let us now prove that $\mathcal{M}^f$ satisfies the condition **S.3**. Consider an arbitrary $x \in INST$ and $C, B \in 2^{AGT\star}$ such that $B \subseteq C$. Suppose that $\mathscr{A}^f_{C:x}(|w|) \neq \emptyset$ and $w' \in \mathscr{A}^f_{B:x}(|w|)$. We have to prove that $w' \in \mathscr{A}^f_{C:x}(|w|)$. By Definition 2, the latter is equivalent to:

1. $\forall \mathcal{A}_{C:x}\varphi \in \Sigma^+_\varphi$, if $\mathcal{M}, w \models \mathcal{A}_{C:x}\varphi$ then $\mathcal{M}, w' \models \varphi$;

2. $\forall z \in INST$ and $\forall D \in 2^{AGT\star}$, if $C \subseteq D$ then:
$\forall \mathcal{A}_{D:z}\varphi \in \Sigma^+_\varphi$, if $\mathcal{M}, w \models \mathcal{A}_{D:z}\varphi$ then $\mathcal{M}, w' \models \mathcal{A}_{D:z}\varphi$;

3. $\forall z \in INST$ and $\forall D \in 2^{AGT\star}$, if $C \subseteq D$ then:
$\forall \widehat{\mathcal{A}}_{D:z}\varphi \in \Sigma^+_\varphi$, if $\mathcal{M}, w \models \widehat{\mathcal{A}}_{D:z}\varphi$ then $\mathcal{M}, w' \models \widehat{\mathcal{A}}_{D:z}\varphi$;

4. $\forall D \in 2^{AGT\star}$, if $C \subseteq D$ then:
$\forall \mathcal{A}_{D:x}\varphi, \widehat{\mathcal{A}}_{D:x}\top \in \Sigma^+_\varphi$, if $\mathcal{M}, w \models \widehat{\mathcal{A}}_{D:x}\top \wedge \mathcal{A}_{D:x}\varphi$ then $\mathcal{M}, w' \models \varphi$;

5. $\exists i \in C$ such that $\forall \mathcal{A}_{i:x}\varphi \in \Sigma^+_\varphi$, if $\mathcal{M}, w' \models \mathcal{A}_{i:x}\varphi$ then $\mathcal{M}, w' \models \varphi$.

So, to prove that $\mathcal{M}^f$ satisfies the condition **S.3** we just need to prove that items **1**-**5** are consequences of $\mathscr{A}^f_{C:x}(|w|) \neq \emptyset$ and $|w'| \in \mathscr{A}^f_{B:x}(|w|)$.

Item 1. Suppose $\mathcal{A}_{C:x}\varphi \in \Sigma^+_\varphi$ and $\mathcal{M}, w \models \mathcal{A}_{C:x}\varphi$. By construction of Σ^+_φ, we have $\widehat{\mathcal{A}}_{C:x}\top \in \Sigma^+_\varphi$. From $\mathscr{A}^f_{C:x}(|w|) \neq \emptyset$ it follows that $\exists w \in |w|$ such that $\mathcal{M}, w \models \widehat{\mathcal{A}}_{C:x}\top$ (by Proposition 4). Thus, by definition of $|w|$, we can conclude that $\forall w \in |w|$ it holds that $\mathcal{M}, w \models \widehat{\mathcal{A}}_{C:x}\top$. Then, in particular, $\mathcal{M}, w \models \widehat{\mathcal{A}}_{C:x}\top$. As $\mathcal{M}, w \models \mathcal{A}_{C:x}\varphi$ and $B \subseteq C$, from the latter it follows that $\mathcal{M}, w \models \mathcal{A}_{B:x}\varphi$ (by Axiom **Inc** of the logic $\mathcal{AL}$). As $|w'| \in \mathscr{A}^f_{B:x}(|w|)$ and $\mathcal{A}_{B:x}\varphi \in \Sigma^+_\varphi$ (from $\mathcal{A}_{C:x}\varphi \in \Sigma^+_\varphi$, by construction of Σ^+_φ), from the latter we conclude $\mathcal{M}, w' \models \varphi$.

Item 2. Take an arbitrary D such that $C \subseteq D$ and an arbitrary $z \in INST$. As $B \subseteq C$, we have $B \subseteq D$. Moreover, suppose $\mathcal{A}_{D:z}\varphi \in \Sigma^+_\varphi$ and $\mathcal{M}, w \models \mathcal{A}_{D:z}\varphi$. As $|w'| \in \mathscr{A}^f_{B:x}(|w|)$, we conclude that $\mathcal{M}, w' \models \mathcal{A}_{D:z}\varphi$.

Item 3. Take an arbitrary D such that $C \subseteq D$ and an arbitrary $z \in INST$. As $B \subseteq C$, we have $B \subseteq D$. Moreover, suppose $\widehat{\mathcal{A}}_{D:z}\varphi \in \Sigma^+_\varphi$ and $\mathcal{M}, w \models \widehat{\mathcal{A}}_{D:z}\varphi$. As $|w'| \in \mathscr{A}^f_{B:x}(|w|)$, we conclude that $\mathcal{M}, w' \models \widehat{\mathcal{A}}_{D:z}\varphi$.

Item 4. Take an arbitrary D such that $C \subseteq D$. As $B \subseteq C$, we have $B \subseteq D$. Moreover, suppose $\mathcal{A}_{D:x}\varphi, \widehat{\mathcal{A}}_{D:x}\top \in \Sigma^+_\varphi$ and $\mathcal{M}, w \models \mathcal{A}_{D:x}\varphi \wedge \widehat{\mathcal{A}}_{D:x}\top$. As $|w'| \in \mathscr{A}^f_{B:x}(|w|)$, we conclude that $\mathcal{M}, w' \models \varphi$.

Item 5. From $w' \in \mathscr{A}^f_{B:x}(|w|)$, it follows that $\exists i \in B$ such that $\forall \mathcal{A}_{i:x}\varphi \in \Sigma^+_\varphi$, if $\mathcal{M}, w' \models \mathcal{A}_{i:x}\varphi$ then $\mathcal{M}, w' \models \varphi$. As $B \subseteq C$, the latter implies that $\exists i \in C$ such that $\forall \mathcal{A}_{i:x}\varphi \in \Sigma^+_\varphi$, if $\mathcal{M}, w' \models \mathcal{A}_{i:x}\varphi$ then $\mathcal{M}, w' \models \varphi$.

This proves that **S.3** holds.

Now, we prove that the model $\mathcal{M}^f$ satisfies the semantic condition **S.4**. That is, we prove that for any $x \in INST$ and $C \in 2^{AGT\star}$:

- if $|w'| \in \mathscr{A}^f_{C:x}(|w|)$ then $|w'| \in \bigcup_{i \in C} \mathscr{A}_{i:x}(|w'|)$.

Suppose $|w'| \in \mathscr{A}^f_{C:x}(|w|)$. We have to prove that $|w'| \in \bigcup_{i\in C} \mathscr{A}_{i:x}(|w'|)$. By Definition 2, the latter is equivalent to the fact that $\exists i \in C$ such that:

1. $\forall \mathcal{A}_{i:x}\varphi \in \Sigma^+_\varphi$, if $\mathcal{M}, w' \models \mathcal{A}_{i:x}\varphi$ then $\mathcal{M}, w' \models \varphi$;

2. $\forall z \in INST$ and $\forall D \in 2^{AGT\star}$, if $i \in D$ then:
$\forall \mathcal{A}_{D:z}\varphi \in \Sigma^+_\varphi$, if $\mathcal{M}, w' \models \mathcal{A}_{D:z}\varphi$ then $\mathcal{M}, w' \models \mathcal{A}_{D:z}\varphi$;

3. $\forall z \in INST$ and $\forall D \in 2^{AGT\star}$, if $i \in D$ then:
$\forall \widehat{\mathcal{A}}_{D:z}\varphi \in \Sigma^+_\varphi$, if $\mathcal{M}, w' \models \widehat{\mathcal{A}}_{D:z}\varphi$ then $\mathcal{M}, w' \models \widehat{\mathcal{A}}_{D:z}\varphi$;

4. $\forall D \in 2^{AGT\star}$, if $i \in D$ then:
$\forall \mathcal{A}_{D:x}\varphi, \widehat{\mathcal{A}}_{D:x}\top \in \Sigma^+_\varphi$, if $\mathcal{M}, w' \models \widehat{\mathcal{A}}_{D:x}\top \wedge \mathcal{A}_{D:x}\varphi$ then $\mathcal{M}, w' \models \varphi$;

5. $\forall \mathcal{A}_{i:x}\varphi \in \Sigma^+_\varphi$, if $\mathcal{M}, w' \models \mathcal{A}_{i:x}\varphi$ then $\mathcal{M}, w' \models \varphi$.

Thus, we have to suppose $|w'| \in \mathscr{A}^f_{C:x}(|w|)$ and prove that $\exists i \in C$ which satisfies items **1**-**5**. Items **2** and **3** trivially hold for all $\exists i \in C$. Moreover, items **1** and **5** are the same condition. Therefore, we just need to prove that $|w'| \in \mathscr{A}^f_{C:x}(|w|)$ implies that $\exists i \in C$ which satisfies items **1** and **4**.

From $|w'| \in \mathscr{A}^f_{C:x}(|w|)$, we can infer that $\exists i \in C$ such that $\forall \mathcal{A}_{i:x}\varphi \in \Sigma^+_\varphi$, if $\mathcal{M}, w' \models \mathcal{A}_{i:x}\varphi$ then $\mathcal{M}, w' \models \varphi$.

By Axiom **Inc** of the logic $\mathcal{AL}$ and by construction of Σ^+_φ the following property holds for all $i \in C$. For all $D \in 2^{AGT\star}$, if $i \in D$ then: $\forall \mathcal{A}_{D:x}\varphi, \widehat{\mathcal{A}}_{D:x}\top \in \Sigma^+_\varphi$, if $\mathcal{M}, w' \models \widehat{\mathcal{A}}_{D:x}\top \wedge \mathcal{A}_{D:x}\varphi$ then $\mathcal{M}, w' \models \mathcal{A}_{i:x}\varphi$ and $\mathcal{A}_{i:x}\varphi \in \Sigma^+_\varphi$.

From the previous two facts, we conclude that $\exists i \in C$ such that: $\forall \mathcal{A}_{i:x}\varphi \in \Sigma^+_\varphi$, if $\mathcal{M}, w' \models \mathcal{A}_{i:x}\varphi$ then $\mathcal{M}, w' \models \varphi$; $\forall D \in 2^{AGT\star}$, if $i \in D$ then: $\forall \mathcal{A}_{D:x}\varphi, \widehat{\mathcal{A}}_{D:x}\top \in \Sigma^+_\varphi$, if $\mathcal{M}, w' \models \widehat{\mathcal{A}}_{D:x}\top \wedge \mathcal{A}_{D:x}\varphi$ then $\mathcal{M}, w' \models \varphi$.

This proves that **S.4** holds.

It remains to be proved that the model $\mathcal{M}^f$ satisfies the semantic condition **S.5**. That is, we have to prove that for any $x \in INST$ and $C, B \in 2^{AGT\star}$ such that $B \subseteq C$:

- if $\mathscr{A}^f_{C:x}(|w|) \neq \emptyset$ then $\mathscr{A}^f_{B:x}(|w|) \neq \emptyset$.

In order to prove this, we prove first that $\mathscr{A}^f_{C:x}(|w|) \neq \emptyset$ implies $\mathcal{M}, w \models \widehat{\mathcal{A}}_{B:x}\top$, when $B \subseteq C$.

Let us suppose that $\mathscr{A}^f_{C:x}(|w|) \neq \emptyset$ and $\mathcal{M}, w \models \mathcal{A}_{B:x}\bot$ with $B \subseteq C$. We show that these facts are inconsistent.

From $\mathcal{M}, w \models \mathcal{A}_{B:x}\bot$ we infer $\mathcal{M}, w \models \mathcal{A}_{C:x}\bot$ (by Axiom **Mon** of the logic $\mathcal{AL}$ and the fact that $B \subseteq C$). From Definition 2 and $\mathscr{A}^f_{C:x}(|w|) \neq \emptyset$, we can infer that $\exists |w'|$ such that $\forall \mathcal{A}_{C:x}\varphi \in \Sigma^+_\varphi$, if $\mathcal{M}, w \models \mathcal{A}_{C:x}\varphi$ then $\mathcal{M}, w' \models \varphi$. By construction of Σ^+_φ we have that $\mathcal{A}_{C:x}\bot \in \Sigma^+_\varphi$. Thus, as we have $\mathcal{M}, w \models \mathcal{A}_{C:x}\bot$, we conclude that $\exists |w'|$ such that $\mathcal{M}, w' \models \bot$.

This proves that $\mathscr{A}^f_{C:x}(|w|) \neq \emptyset$ implies $\mathcal{M}, w \models \widehat{\mathcal{A}}_{B:x}\top$, when $B \subseteq C$.

Now, we have to show that $\mathcal{M}, w \models \widehat{\mathcal{A}}_{B:x}\top$ implies $\mathscr{A}^f_{B:x}(|w|) \neq \emptyset$.

$\mathcal{M}, w \models \widehat{\mathcal{A}}_{B:x}\top$ implies that $\exists w'$ such that $w' \in \mathscr{A}_{B:x}(w)$. As $\mathcal{M}^f$ is a filtration of $\mathcal{M}$ (Lemma 1), from the latter we conclude that $\exists |w'|$ such that $|w'| \in \mathscr{A}^f_{B:x}(|w|)$.

This proves that **S.5** holds.

□

Lemma 3. *The model $\mathcal{M}^f$ contains at most 2^n worlds where n denotes the size of Σ^+_φ.*

Proof. From Lemma 1 and Proposition 2.38 given in [Blackburn et al., 2001, p. 79]. □

Lemma 4. *$\mathcal{M}^f$ is a finite model.*

Proof. From Lemma 3 and Proposition 3. □

Lemma 5. *Formula φ is satisfiable in $\mathcal{M}^f$.*

Proof. From Lemma 1 and Proposition 3, the fact that φ is satisfiable in $\mathcal{M}$, the fact that $\varphi \in \Sigma^+_\varphi$ and the Filtration Theorem given in [Blackburn et al., 2001, p. 79]. □

Lemma 6. *The logic $\mathcal{AL}$ has the finite model property.*

Proof. We have started with an arbitrary formula φ which is satisfiable in a $\mathcal{AL}$ model $\mathcal{M}$. We have built a model $\mathcal{M}^f$ and proved that $\mathcal{M}^f$ is a finite $\mathcal{AL}$ model (Lemma 4). Finally, we have proved that φ is satisfiable in $\mathcal{M}^f$ (Lemma 5). Thus, we can conclude that for every formula φ, if φ is $\mathcal{AL}$ satisfiable then, φ is satisfiable in a finite $\mathcal{AL}$ model. □

Theorem 2 is a direct consequence of Lemma 6.

Proof of Theorem 14

As for the logic $\mathcal{AL}$, it is a routine to prove soundness, whereas completeness is again obtained by Sahlqvist completeness theorem. Indeed, all axioms of $\mathcal{AL}^+$ are in the Sahlqvist class, for which a general completeness result exists [Sahlqvist, 1975, Blackburn et al., 2001].

Proof of Theorem 15

In order to prove Theorem 15, it is sufficient to prove that if $INST = CXT$ and φ is a formula of the $\mathcal{GMD}$ logic then: if φ is a theorem of $\mathcal{GMD}$ then $tr(\varphi)$ is a theorem of $\mathcal{AL}^+$ and, if φ is $\mathcal{GMD}$ satisfiable then $tr(\varphi)$ is $\mathcal{AL}^+$ satisfiable.

Proposition 5. *Suppose that $INST = CXT$ and φ is a formula of the logic $\mathcal{GMD}$ then: if $\vdash_{\mathcal{GMD}} \varphi$ then $\vdash_{\mathcal{AL}^+} tr(\varphi)$.*

Proof. We only need to prove that the translations of the axioms of the $\mathcal{GMD}$ logic are theorems of $\mathcal{AL}^+$ and that the translated rules of inference of $\mathcal{GMD}$ preserves validity.

It is straightforward to show that the translation of the rules of inference $\mathbf{Nec}_{[\![x]\!]}$, $\mathbf{Nec}_{[\![Univ]\!]}$ and **MP** preserve validity. As the $\mathcal{AL}^+$ operators $[x]$ and $[Univ]$ are normal, it is a routine to verify that the translation of the $\mathcal{GMD}$ Axioms $\mathbf{K}_{[\![x]\!]}$ and $\mathbf{K}_{[\![Univ]\!]}$ are

theorems of $\mathcal{AL}^+$. Furthermore, by the definitions of $[x]\,\varphi$ and $[Univ]\,\varphi$, it is just trivial to prove that that the translation of the $\mathcal{GMD}$ Axiom $\subseteq_{[\![Univ]\!],[\![x]\!]}$ is a theorem of $\mathcal{AL}^+$. The translation of the $\mathcal{GMD}$ Axiom $\mathbf{T}_{[\![Univ]\!]}$ is a theorem of $\mathcal{AL}^+$ as well. Indeed, this corresponds to the Axiom $\mathbf{T}_{Univ}$ of the logic $\mathcal{AL}^+$. By Axioms **PAccess**$^+$ and **NAccess**$^+$ we can prove that the translations of the $\mathcal{GMD}$ Axioms $\mathbf{4}_{[\![x]\!],[\![y]\!]}$ and $\mathbf{5}_{[\![x]\!],[\![y]\!]}$ are theorems of $\mathcal{AL}^+$. By the same principles, we can prove that the translations of the $\mathcal{GMD}$ Axioms $\mathbf{4}_{[\![Univ]\!]}$ and $\mathbf{5}_{[\![Univ]\!]}$ are theorems of $\mathcal{AL}^+$. □

Proposition 6. *Suppose that $INST = CXT$ and φ is a formula of the logic $\mathcal{GMD}$ then: if φ is $\mathcal{GMD}$ satisfiable then $tr(\varphi)$ is $\mathcal{AL}^+$ satisfiable.*

Proof. Suppose that φ is $\mathcal{GMD}$ satisfiable. Thus, there exists a $\mathcal{GMD}$ model $\mathcal{M}^{\mathcal{GMD}} = \langle S, \{S_x\}_{x \in CXT_0}, \pi \rangle$ which satisfies φ. We prove that we can build a $\mathcal{AL}^+$ model $\mathcal{M}$ which satisfies the same formulas as $\mathcal{M}^{\mathcal{GMD}}$.

As we have supposed $INST = CXT$, the $\mathcal{AL}^+$ model $\mathcal{M}$ associated with the $\mathcal{GMD}$ model $\mathcal{M}^{\mathcal{GMD}}$ can be defined as follows.

- $W = S$;
- $\forall w \in W, \forall x \in CXT_0, \forall C \in 2^{AGT\star}$, $\mathscr{A}_{C:x}(w) = S_x$;
- $\forall w \in W, \forall C \in 2^{AGT\star}$, $\mathscr{A}_{C:Univ}(w) = S$;
- $\forall w \in W, \forall p \in ATM, w \in \pi(p)$ if and only if $w \in \mathscr{V}(p)$.

It is a routine to verify that the previous conditions ensure that the model $\mathcal{M}$ is indeed a $\mathcal{AL}^+$ model. By structural induction on φ, it is also a routine to prove that the previous $\mathcal{AL}^+$ model satisfies the same formulas as the $\mathcal{GMD}$ model it is associated. That is, $\mathcal{M}^{\mathcal{GMD}}, w \models \varphi$ if and only if $\mathcal{M}, w \models tr(\varphi)$.

Theorem 15 is an immediate corollary of Proposition 5 and Proposition 6.

□

Proof of Theorems 3, 4, 5, 6 and 7

Theorems 3 and 4 can be syntactically proved using $\mathcal{AL}$ logic axiomatization. Theorem 5 proof is based on Theorem 3. As every $\mathcal{A}_{C:x}$ operator is normal, Theorems 6 and 7 can be proved by iteration of the Axiom (**K**) and the Rule of Necessitation (**Nec**) for every group of $2^{AGT\star}$ and every institution of $INST$. We provide in the sequel only the complete proof for Theorems (3a) and (3e).

Proof. Theorem (3a):

(1) $\vdash_{\mathcal{AL}} \mathcal{A}_{C:x}\bot \vee \neg\mathcal{A}_{C:x}\bot$, by (**ProTau**)

(2) $\vdash_{\mathcal{AL}} \neg\mathcal{A}_{C:x}\bot \rightarrow \mathcal{A}_{C:x}\neg\mathcal{A}_{C:x}\bot$, by (**NAccess**)

(3) $\vdash_{\mathcal{AL}} \mathcal{A}_{C:x}\bot \rightarrow \mathcal{A}_{C:x}\neg\mathcal{A}_{C:x}\bot$, by (**ProTau**), (**Nec**) and (**K**)

(4) $\vdash_{\mathcal{AL}} \mathcal{A}_{C:x}\neg\mathcal{A}_{C:x}\bot$, from (1), (2) and (3) by (**ProTau**)

□

Proof. Theorem (3e):

(1) $\vdash_{\mathcal{AL}} \mathcal{A}_{C:x}((\neg\mathcal{A}_{C:x}\bot \wedge \mathcal{A}_{C:x}\varphi) \rightarrow \mathcal{A}_{i:x}\varphi)$, for every $i \in C$, from Axiom (**Inc**) by inference rule (**Nec**),

(2) $\vdash_{\mathcal{AL}} \mathcal{A}_{C:x}((\neg\mathcal{A}_{C:x}\bot \wedge \mathcal{A}_{C:x}\varphi) \rightarrow \bigwedge_{i\in C}\mathcal{A}_{i:x}\varphi)$, from (1), by K principles

(3) $\vdash_{\mathcal{AL}} \mathcal{A}_{C:x}(\bigwedge_{i\in C}\mathcal{A}_{i:x}\varphi \rightarrow \varphi)$, from (**Unanim**)

(4) $\vdash_{\mathcal{AL}} \mathcal{A}_{C:x}((\neg\mathcal{A}_{C:x}\bot \wedge \mathcal{A}_{C:x}\varphi) \rightarrow \varphi)$, from (2) and (3) by (**ProTau**) and (**K**)

(5) $\vdash_{\mathcal{AL}} \mathcal{A}_{C:x}\neg\mathcal{A}_{C:x}\bot \rightarrow \mathcal{A}_{C:x}(\mathcal{A}_{C:x}\varphi \rightarrow \varphi)$, from (4) by (**ProTau**) and (**K**)

(6) $\vdash_{\mathcal{AL}} \neg\mathcal{A}_{C:x}\bot \rightarrow \mathcal{A}_{C:x}\neg\mathcal{A}_{C:x}\bot$, by (**NAccess**)

(7) $\vdash_{\mathcal{AL}} \neg\mathcal{A}_{C:x}\bot \rightarrow \mathcal{A}_{C:x}(\mathcal{A}_{C:x}\varphi \rightarrow \varphi)$, from (5) and (6) by (**ProTau**)

(8) $\vdash_{\mathcal{AL}} \mathcal{A}_{C:x}\bot \rightarrow \mathcal{A}_{C:x}(\mathcal{A}_{C:x}\varphi \rightarrow \varphi)$, by (**ProTau**), (**Nec**) and (**K**)

(9) $\vdash_{\mathcal{AL}} \mathcal{A}_{C:x}(\mathcal{A}_{C:x}\varphi \rightarrow \varphi)$, from (7) and (8) by (**ProTau**)

□

Proof of Proposition 1

Proof. Let suppose that the majority Principle (**Majority**) holds for any sets of agents C and B such that $B \subseteq C$ and $|C \setminus B| < |B|$. We will prove by induction on the set C_n, that there exists a set C_n such that:

$$(P_n) \qquad (\neg\mathcal{A}_{AGT:x}\bot \wedge \mathcal{A}_{AGT:x}\mathcal{A}_{\{i,j\}:x}\varphi) \rightarrow \mathcal{A}_{C_n:x}\varphi$$

where $i, j \in AGT$, $C_n \subseteq AGT$, $|C_n| = n$ and $n \geq 2$.

We begin by showing that (P_2) holds.

(1) $\vdash_{\mathcal{AL}} \neg\mathcal{A}_{AGT:x}\bot \wedge \mathcal{A}_{AGT:x}\mathcal{A}_{\{i,j\}:x}\varphi \rightarrow \mathcal{A}_{\{i,j\}:x}\mathcal{A}_{\{i,j\}:x}\varphi$, by (**Inc**).

(2) $\vdash_{\mathcal{AL}} \neg\mathcal{A}_{AGT:x}\bot \wedge \mathcal{A}_{AGT:x}\mathcal{A}_{\{i,j\}:x}\varphi \rightarrow \mathcal{A}_{\{i,j\}:x}\varphi$, from (1) by Theorem (3c)

(2) entails that (P_2) holds.

We suppose that (P_n) holds for any n such that $C_n \subset AGT$. Under this hypothesis we will show that (P_{n+1}) holds. We suppose that C_{n+1} is defined as: $C_{n+1} = C_n \cup \{i\}$, with $i \in AGT$ and $i \notin C_n$ (thus $C_n \subset C_{n+1}$).

(3) $\vdash_{\mathcal{AL}} (\neg\mathcal{A}_{AGT:x}\bot \wedge \mathcal{A}_{AGT:x}\mathcal{A}_{\{i,j\}:x}\varphi) \rightarrow \mathcal{A}_{C_n:x}\varphi$, by induction hypothesis (P_n)

(4) $\vdash_{\mathcal{AL}} (\mathcal{A}_{C_{n+1}:x}\neg\mathcal{A}_{AGT:x}\bot \wedge \mathcal{A}_{C_{n+1}:x}\mathcal{A}_{AGT:x}\mathcal{A}_{\{i,j\}:x}\varphi) \rightarrow \mathcal{A}_{C_{n+1}:x}\mathcal{A}_{C_n:x}\varphi$, from (3) by (**Nec**), (**K**) and standard properties of normal modal operator $\mathcal{A}_{C_{n+1}:x}$

(5) $\vdash_{\mathcal{AL}} (\neg\mathcal{A}_{AGT:x}\bot \wedge \mathcal{A}_{AGT:x}\mathcal{A}_{\{i,j\}:x}\varphi) \rightarrow \mathcal{A}_{C_{n+1}:x}\mathcal{A}_{C_n:x}\varphi \wedge \neg\mathcal{A}_{C_{n+1}:x}\bot$, from (4) by (**PAccess**), (**NAccess**), (**Mon**) and (**ProTau**)

(6) $\vdash_{\mathcal{AL}} \neg\mathcal{A}_{C_{n+1}:x}\bot \rightarrow \neg\mathcal{A}_{C_n:x}\bot$, by (**Mon**)

(7) $\vdash_{\mathcal{AL}} \mathcal{A}_{C_{n+1}:x}\neg\mathcal{A}_{C_{n+1}:x}\bot \rightarrow \mathcal{A}_{C_{n+1}:x}\neg\mathcal{A}_{C_n:x}\bot$, from (6) by (**Nec**), (**K**)

(8) $\vdash_{\mathcal{AL}} \neg\mathcal{A}_{C_{n+1}:x}\bot \rightarrow \mathcal{A}_{C_{n+1}:x}\neg\mathcal{A}_{C_n:x}\bot$, from (7) by (**NAccess**) and (**ProTau**)

(9) $\vdash_{\mathcal{AL}} (\neg\mathcal{A}_{AGT:x}\bot \wedge \mathcal{A}_{AGT:x}\mathcal{A}_{\{i,j\}:x}\varphi) \rightarrow \mathcal{A}_{C_{n+1}:x}(\mathcal{A}_{C_n:x}\varphi \wedge \neg\mathcal{A}_{C_n:x}\bot)$, from (5) and (8) by (**ProTau**) and standard properties of normal modal operator $\mathcal{A}_{C_{n+1}:x}$

(10) $\vdash_{\mathcal{AL}} (\neg\mathcal{A}_{AGT:x}\bot \wedge \mathcal{A}_{AGT:x}\mathcal{A}_{\{i,j\}:x}\varphi) \rightarrow \mathcal{A}_{C_{n+1}:x}(\bigwedge_{k\in C_n}\mathcal{A}_{k:x}\varphi)$, from (9) by (**Inc**), (**K**), (**Nec**) and (**ProTau**)

(11) $\vdash_{\mathcal{AL}} (\neg\mathcal{A}_{AGT:x}\bot \wedge \mathcal{A}_{AGT:x}\mathcal{A}_{\{i,j\}:x}\varphi) \rightarrow \mathcal{A}_{C_{n+1}:x}\varphi$, from (10) by (**Majority**), (**K**), (**Nec**) and (**ProTau**)

Thus (11) entails (P_{n+1}).

As (P_2) holds and from (P_n) we can infer that (P_{n+1}) for $n < |AGT|$, we can thus deduce by induction that (P_n) holds for $n \leq |AGT|$. In particular, we can deduce from the extension of the Principles (**Majority**) for every set of agents C and B such that $B \subseteq C$ and $|C \setminus B| < |B|$, that the following counterintuitive formula holds:

$$(\mathcal{A}_{AGT:x}\mathcal{A}_{\{i,j\}:x}\varphi \wedge \neg\mathcal{A}_{AGT:x}\bot) \rightarrow \mathcal{A}_{AGT:x}\varphi$$

□

Proof of Proposition 2

Lemma 7.

(7a) $\mathcal{A}_{C:x}\varphi \leftrightarrow Bel_i\mathcal{A}_{C:x}\varphi$ *if* $i \in C$

(7b) $\neg\mathcal{A}_{C:x}\varphi \leftrightarrow Bel_i\neg\mathcal{A}_{C:x}\varphi$ *if* $i \in C$

Proof. Lemma (7a) and (7b):

(1) $\neg\mathcal{A}_{C:x}\varphi \rightarrow Bel_i\neg\mathcal{A}_{C:x}\varphi$, by (**NegIntrAccept**), for $i \in C$

(2) $Bel_i\neg\mathcal{A}_{C:x}\varphi \rightarrow \neg Bel_i\mathcal{A}_{C:x}\varphi$, by Axiom (**D**) for Bel_i

(3) $Bel_i\mathcal{A}_{C:x}\varphi \rightarrow \mathcal{A}_{C:x}\varphi$, from (1), (2), and (**ProTau**), for $i \in C$

The proof of Lemma (7b) is similar to the one of Lemma (7a), we only use Axiom (**PIntrAccept**) instead of Axiom (**NegIntrAccept**). □

Proof. Propositions (2a) and (2b):

(1) $\mathcal{MB}_C\mathcal{A}_{C:x}\varphi \rightarrow \bigwedge_{i\in C} Bel_i(\mathcal{A}_{C:x}\varphi \wedge \mathcal{MB}_C\mathcal{A}_{C:x}\varphi)$, by (**FixPoint**)

(2) $\bigwedge_{i\in C} Bel_i(\mathcal{A}_{C:x}\varphi \wedge \mathcal{MB}_C\mathcal{A}_{C:x}\varphi) \rightarrow \bigwedge_{i\in C} Bel_i\mathcal{A}_{C:x}\varphi$, because Bel_i are normal modal operators

(3) $\bigwedge_{i\in C} Bel_i\mathcal{A}_{C:x}\varphi \rightarrow \mathcal{A}_{C:x}\varphi$, by Lemma (7a)

(4) $\mathcal{MB}_C\mathcal{A}_{C:x}\varphi \rightarrow \mathcal{A}_{C:x}\varphi$, from (1), (2), (3) by (**ProTau**)

(5) $\mathcal{A}_{C:x}\varphi \rightarrow Bel_i\mathcal{A}_{C:x}\varphi$, by (**PIntrAccept**), for every $i \in C$

(6) $\mathcal{A}_{C:x}\varphi \rightarrow E_C(\mathcal{A}_{C:x}\varphi \wedge \mathcal{A}_{C:x}\varphi)$, from (5), by (**ProTau**) and definition of E_C

(7) $\mathcal{A}_{C:x}\varphi \rightarrow \mathcal{MB}_C\mathcal{A}_{C:x}\varphi$, from (6) by inference rule (**InductionRule**) (left to right direction of Theorem (2a))

(8) $\mathcal{A}_{C:x}\varphi \leftrightarrow \mathcal{MB}_C\mathcal{A}_{C:x}\varphi$, from (4) and (7)

The proof of Proposition (2b) is similar to the one of Proposition (2a), we only use Lemma (7b) instead of Lemma (7a) and (**NegIntrAccept**) instead of (**PIntrAccept**). □

Proof of Theorem 8

To prove that these formulas are not valid in $\mathcal{AL}$, we only have to exhibit a model where there is a world where these formulas are false. We give the complete proof only for Theorem (8b), the others are very similar.

Proof. Theorem (8b):
We will build a $\mathcal{AL}$ model $\mathcal{M}$ in which there is a world w is which the formula is false, *i.e.*: $\mathcal{M}, w \vDash (\varphi_1 \overset{x}{\rhd} \varphi_2) \wedge (\varphi_2 \overset{x}{\rhd} \varphi_3) \wedge \neg(\varphi_1 \overset{x}{\rhd} \varphi_3)$. Let $ATM = \{\varphi_1, \varphi_2, \varphi_3\}$, $AGT = \{i\}$, $INST = \{x, y, z\}$ and $W = \{w, w_x, w_y, w_z\}$. We build the valuation function $\mathscr{V}$: $\mathscr{V}(\varphi_1) = \{w_y\}$, $\mathscr{V}(\varphi_2) = \{w_z\}$ and $\mathscr{V}(\varphi_3) = \{w_y\}$, and the relation $\mathscr{A}$: $\mathscr{A}_{\{i\}:x}(w) = \{w_x\}$, $\mathscr{A}_{\{i\}:y}(w) = \{w_y\}$ and $\mathscr{A}_{\{i\}:z}(w) = \{w_z\}$.

As we want $\mathcal{M}$ to be a $\mathcal{AL}$ model, we ensure that it satisfies the constraints **S.1**-**S.5**.

- In order to satisfy (**S.1**) and (**S.2**) we impose: $\langle w_y, w_x\rangle \in \mathscr{A}_{\{i\}:x}$, $\langle w_z, w_x\rangle \in \mathscr{A}_{\{i\}:x}$, $\langle w_x, w_y\rangle \in \mathscr{A}_{\{i\}:y}$, $\langle w_z, w_y\rangle \in \mathscr{A}_{\{i\}:y}$, $\langle w_x, w_z\rangle \in \mathscr{A}_{\{i\}:z}$ and $\langle w_y, w_z\rangle \in \mathscr{A}_{\{i\}:z}$;
- as there is only one agent in our model, (**S.3**) and (**S.5**) are satisfied;
- in order to satisfy (**S.4**) we impose that: $\langle w_x, w_x\rangle \in \mathscr{A}_{\{i\}:x}$, $\langle w_y, w_y\rangle \in \mathscr{A}_{\{i\}:y}$ and $\langle w_z, w_z\rangle \in \mathscr{A}_{\{i\}:z}$;

In this model $\mathcal{M}$:

- $\mathcal{M}, w \vDash [x](\varphi_1 \rightarrow \varphi_2) \wedge [x](\varphi_2 \rightarrow \varphi_3)$
- $\mathcal{M}, w \vDash \neg[y](\varphi_1 \rightarrow \varphi_2)$ and thus $\mathcal{M}, w \vDash \neg[Univ](\varphi_1 \rightarrow \varphi_2)$
- $\mathcal{M}, w \vDash \neg[z](\varphi_2 \rightarrow \varphi_3)$ and thus $\mathcal{M}, w \vDash \neg[Univ](\varphi_2 \rightarrow \varphi_3)$
- $\mathcal{M}, w \vDash [x](\varphi_1 \rightarrow \varphi_3) \wedge [y](\varphi_1 \rightarrow \varphi_3) \wedge [z](\varphi_1 \rightarrow \varphi_3)$, *i.e.* $\mathcal{M}, w \vDash [Univ](\varphi_1 \rightarrow \varphi_3)$

We have built a $\mathcal{AL}$ model which satisfies the formula $(\varphi_1 \overset{x}{\rhd} \varphi_2) \wedge (\varphi_2 \overset{x}{\rhd} \varphi_3) \wedge \neg(\varphi_1 \overset{x}{\rhd} \varphi_3)$. Thus, $(\varphi_1 \overset{x}{\rhd} \varphi_2) \wedge (\varphi_2 \overset{x}{\rhd} \varphi_3) \rightarrow (\varphi_1 \overset{x}{\rhd} \varphi_3)$ is not valid in $\mathcal{AL}$. By Theorem 1, we conclude that $(\varphi_1 \overset{x}{\rhd} \varphi_2) \wedge (\varphi_2 \overset{x}{\rhd} \varphi_3) \rightarrow (\varphi_1 \overset{x}{\rhd} \varphi_3)$ is not a theorem of $\mathcal{AL}$. □

Proof of Theorem 9

Proof. Theorems (9a) and (9b):
Since $[x]$ and $[Univ]$ are normal modal operators, they satisfy the rule of equivalence RE [Chellas, 1980]. Theorems (9a) and (9b) follow straightforwardly from RE. □

Proof. Theorem (9c):

(1) $\vdash_{\mathcal{AL}} ((\varphi_1 \rightarrow \varphi_2) \wedge (\varphi_1 \rightarrow \varphi_3)) \leftrightarrow (\varphi_1 \rightarrow (\varphi_2 \wedge \varphi_3))$, by (**ProTau**)

(2) $\vdash_{\mathcal{AL}} ([x](\varphi_1 \rightarrow \varphi_2) \wedge [x](\varphi_1 \rightarrow \varphi_3)) \rightarrow [x](\varphi_1 \rightarrow (\varphi_2 \wedge \varphi_3))$, from (1) by Theorem (6b)

(3) $\vdash_{\mathcal{AL}} (\neg[Univ](\varphi_1 \rightarrow \varphi_2) \wedge \neg[Univ](\varphi_1 \rightarrow \varphi_3)) \rightarrow (\neg[Univ](\varphi_1 \rightarrow \varphi_2) \vee \neg[Univ](\varphi_1 \rightarrow \varphi_3))$, by (**ProTau**)

(4) $\vdash_{\mathcal{AL}} (\neg[Univ](\varphi_1 \rightarrow \varphi_2) \vee \neg[Univ](\varphi_1 \rightarrow \varphi_3)) \rightarrow \neg[Univ]((\varphi_1 \rightarrow \varphi_2) \wedge (\varphi_1 \rightarrow \varphi_3))$, by standard properties of normal modal operator $[Univ]$

(5) $\vdash_{\mathcal{AL}} (\neg[Univ](\varphi_1 \rightarrow \varphi_2) \wedge \neg[Univ](\varphi_1 \rightarrow \varphi_3)) \rightarrow \neg[Univ](\varphi_1 \rightarrow (\varphi_2 \wedge \varphi_3))$, from (3) and (4) by (**ProTau**)

(6) $((\varphi_1 \overset{x}{\triangleright} \varphi_2) \wedge (\varphi_1 \overset{x}{\triangleright} \varphi_3)) \rightarrow (\varphi_1 \overset{x}{\triangleright} (\varphi_2 \wedge \varphi_3))$, from (2) and (5) and (**ProTau**)

□

Proof. Theorems (9d) and (9e):
The proofs of Theorems (9d) and (9e) are very similar to the previous one. Both apply (**Nec**), (**K**) and propositional tautologies.

□

Proof of Theorem 10

All these theorems follow from the necessitation rule (**Nec**) and logical tautologies. Proofs also need Axiom (**K**) and theorems (M) and (C)[10] [Chellas, 1980] for the distribution over conjunction. We give the complete proof of Theorem (10b) as an example.

Proof. Theorem (10b):

(1) $\vdash_{\mathcal{AL}} ((\varphi_1 \rightarrow \varphi_2) \wedge (\varphi_2 \rightarrow \varphi_3)) \rightarrow (\varphi_1 \rightarrow \varphi_3)$, by (**ProTau**)

(2) $\vdash_{\mathcal{AL}} [x](((\varphi_1 \rightarrow \varphi_2) \wedge (\varphi_2 \rightarrow \varphi_3)) \rightarrow (\varphi_1 \rightarrow \varphi_3))$, from (1) by (**Nec**)

(3) $\vdash_{\mathcal{AL}} [x]((\varphi_1 \rightarrow \varphi_2) \wedge (\varphi_2 \rightarrow \varphi_3)) \rightarrow [x](\varphi_1 \rightarrow \varphi_3)$, from (2) by (**K**) and (C).

□

[10]The conjunction of both (M) and (C) give the equivalence: $[x](\varphi_1 \wedge \varphi_2) \leftrightarrow ([x]\varphi_1 \wedge [x]\varphi_2)$.

Proof of Theorem 11

Proof. Theorem (11a):

This theorem comes straightforwardly from Theorems 6a, 6b, 7a, 7b, and the following propositional tautologies:

(1) $\vdash_{\mathcal{AL}} (((\varphi \wedge \neg\psi) \rightarrow \mathit{viol}) \wedge (\neg\varphi \rightarrow \mathit{viol})) \rightarrow (\neg\psi \rightarrow \mathit{viol})$, by (**ProTau**)

(2) $\vdash_{\mathcal{AL}} (\neg\psi \rightarrow \mathit{viol}) \rightarrow ((\neg\psi \wedge \varphi) \rightarrow \mathit{viol})$, by (**ProTau**)

□

Proof. Theorem (11b):

(1) $\vdash_{\mathcal{AL}} O_x\top \rightarrow \neg\,[\mathit{Univ}]\,(\bot \rightarrow \mathit{viol})$, by the definition of $O_x\top$

(2) $\vdash_{\mathcal{AL}} \neg\,[\mathit{Univ}]\,(\bot \rightarrow \mathit{viol}) \rightarrow \neg\,[\mathit{Univ}]\,\top$, by (**ProTau**)

(3) $\vdash_{\mathcal{AL}} \neg\,[\mathit{Univ}]\,\top \rightarrow \bot$, by standard properties of normal modal operator $[\mathit{Univ}]$

(4) $\vdash_{\mathcal{AL}} \top \rightarrow \neg O_x\top$, from (1), (2) and (3) by (**ProTau**)

□

Social connectedness improves co-ordination on individually costly, efficient outcomes *

Giuseppe Attanasi† Astrid Hopfensitz‡ Emiliano Lorini§
Frédéric Moisan¶

Abstract

We study the impact of social ties on behavior in two types of asymmetric coordination games. Social ties are varied by making players interact with partners from different in-groups (fellow members of their own sports team, members of their sports club, students of their university). Subjective social ties are further measured by direct questionnaires.

We find that smaller and more salient in-groups lead to significantly more group beneficial choices. The same effect is observed for players that report high values of their subjective social ties. We discuss the implication of these results for theories assuming that socially tied individuals follow some group beneficial reasoning.

JEL classification numbers: C72, C91, C92.
Keywords: Social ties, group identity, coordination, experiment.

*We thank seminar participants at Norms, Actions and Games (London), the Thurgau Experimental Economics Meeting, the ESA meeting in New York for comments, the Institute for Advanced Studies Toulouse (IAST) for support. Funding through the ANR 2010 JCJC 1803 01 TIES is gratefully acknowledged.

†BETA, Université de Strasbourg
‡Toulouse School of Economics, astrid.hopfensitz@tse-fr.eu
§IRIT, Université Paul Sabatier
¶University of Cambridge

1 Introduction

This paper analyzes behavior in asymmetric coordination games for different levels of social ties between players. Behavior in interactions requiring coordination is intrinsically difficult to predict. While participants in coordination problems clearly prefer to coordinate, reaching this state in the absence of communication is not trivial. Focal points help in symmetric games but might not be strong enough when games are asymmetric (Crawford et al., 2008). Incidentally human interactions are never completely void of information about interaction partners. When we interact with others, we take into account people's nationality, gender, political preferences, or favorite sports team. Group membership of others leads us to anticipate certain behaviors or influences our concerns for their welfare. These effects have, over the last years, received increased attention in economics and psychology. Specifically, recent experimental evidence has investigated the importance of a joint social identity on coordination among multiple Pareto ranked equilibria (Chen and Li, 2009).

The importance of *social identity* in economics has been pointed out by the seminal work of Akerlof and Kranton (2000). Social identity theory is based on the assumption that an individual is not characterized by one unique 'personal self', but rather by many 'selves' that correspond to overlapping circles of group identities. Different cues might trigger the individuals to act and feel on their personal, family or national 'level of self' (Turner et al., 1987). While social identity theory clearly considers the importance of different levels or 'strength' of social identity, the main experimental approach has focused on identifying a '*minimal group*' level, which allows us to observe discriminatory behavior.[1] Since the seminal work by Tajfel et al. (1971), results support the idea that even minimal group membership enhances behavior beneficial for group members, sometimes at the expenses of the out-group (i.e., people who are not members of the group).

[1]Primarily based on this 'minimal group paradigm', a large body of evidence has been collected in psychology (e.g. Brewer, 1979, 1999; Tajfel and Turner, 1979) and in economics (e.g. Bernhard et al., 2006; Buchan et al., 2006; Chavanne et al., 2011; Goette et al., 2012) on in-group vs out-group behavior.

A reasonable assumption is that the *strength of a social tie* between two individuals can be described through the level with which each of them commonly identifies with the same group(s). This interpretation relies on a gradual type of group identification that extends the binary in-group/out-group classification presented above (one may instead identify with a given group up to a certain degree). In this case, the question is how behavior is changed when we interact with a person we know very well compared to an acquaintance, or a person we are only minimally tied with (e.g. a perfect stranger). What to expect is not immediately obvious. Stronger ties with others might lead to better predictions of others' intentions. For example, a stronger tie between two individuals might influence their concerns for each other's outcomes according to existing theories of social preferences (e.g. Charness and Rabin, 2002; Fehr and Schmidt, 1999), which all assume individual intentions (what should I do?) to maximize the overall benefit of the group. Alternatively, stronger connectedness might induce another type of reasoning that focuses on collective intentions (what should *we* do?) to reach the same group beneficial outcome (e.g. Bacharach, 1999; Sugden, 2000, 2003). Finally, a stronger social tie between two individuals might enforce some commonly known external norms such that the cost for deviating from them increases (Goette et al., 2012): not conforming to an in-group norm may lead one to identify more as an out-group member and consequently deteriorate the quality of the social relationship.

Recent experimental evidence has observed that, for the case of *symmetric games*, an 'enhanced' minimal group paradigm enables coordination among multiple Pareto ranked equilibria (Chen and Chen, 2011). Similarly Gaechter et al. (2012) observe that players scoring higher on a psychological measure of shared identity ('one'-ness) are more likely to coordinate on high effort levels in a weakest link game. Furthermore, Charness et al. (2007) have shown that coordination can increase in a battle of the sexes game when a 'host' player (in-group member) interacts with a 'guest' player (out-group member). In this case, making these roles salient to the players (when they are both observed by the host's in-group members) leads coordination to favor an 'aggressive'

host who faces a more 'accommodating' guest.

Asymmetric games have however not been investigated for the case of different levels of connectedness among interacting players. Yet stronger connectedness might favor coordination on outcomes that are considered as better for the group. For example, take a battle of the sexes game that does not offer symmetric payoffs dependent on which outcome the players agree on. Consider the classical example of Ann and Bob who have to choose between going to the opera or to a football game. Though Ann might prefer the opera and Bob the football game, Ann might have a higher utility from the football game than Bob from the opera. Thus the overall efficiency for the couple is higher when they coordinate on the football game.

In this paper, we study with an experiment how the level of connectedness with others influences coordination in asymmetric battle of the sexes games, where coordination comes at a cost for the individual. We investigate this effect in two games: an *asymmetric battle of the sexes game* (baseline game) and an extension of this game where, due to the presence of an *outside option*, one player has to make a conscious choice to enter the battle of the sexes game (entrance game). The entrance game enables us to investigate how social connectedness influences the interpretation of moves by the other.

Objective social ties are varied by making players interact with partners from different in-groups: fellow members of their own sports team, members of their sports club, students of their university. Sports team are exogenously assigned, according to gender and skill at playing volleyball. In this regard, our experimental design is different from other studies on social ties where groups are formed endogenously Leider et al. (2009); Goeree et al. (2010), while instead it shares features with studies where groups are randomly assigned Goette et al. (2012) or differ in terms of member characteristics Fershtman and Gneezy (2001). *Subjective social ties* are further measured by direct questionnaires. These questionnaires are aimed at eliciting a subject's perceived self-connectedness to the group and perceived ties between other team members.

Our results show that even in asymmetric games where one player has

to accept an individual cost, coordination on a group beneficial outcome is increased with both stronger objective ties and stronger subjective ties. Higher social connectedness indeed enhances the focal value of such group beneficial outcomes.

The rest of the article is organized as follows. Section 2 clarifies the concept of a social tie that we consider. Section 3 presents the two versions of the asymmetric battle of the sexes game (the baseline and entrance game) studied in this paper. We further discuss how social ties are measured. Specifically we distinguish between objective ties (which refer to the type of partners a subject interacts with) and subjective ties (which correspond to a subject's own perception about social relationships within a group). Section 4 gives the procedures of the experiment and Section 5 presents results from both coordination games, depending on both types of social ties (objective and subjective). Section 6 discusses the theoretical implications of our results.

2 Defining social ties

The concept of social ties that we consider relies on *social identity theory* (Tajfel and Turner, 1979; Hogg, 2002), according to which an individual's social identity is built upon a set of social features, each referring to a salient characteristic that can be shared by individuals in a particular context (e.g., one may identify himself as a student of a specific university, a member of a particular sport club, a member of a political party, etc).

However, departing from this theory, we assert that the minimal criterion for the existence of a social tie between two individuals is for them to commonly believe that they share the same social features that define their social identities (Attanasi et al., 2014). Such a requirement clearly distinguishes *social ties* from *unilateral ties* where a person's perceived closeness with another may not be reciprocated (feeling close to someone, e.g., a political figure, is insufficient to be described as a social tie).

Beyond this basic definition, an important property of social ties lies in their quantitative aspect: two individuals can indeed be more or less socially

connected with one another. In this case, we argue that the strength of a tie can be determined based on the quantity and importance of the social features that are shared by the two persons. Intuitively, sharing more social features of high importance (as commonly perceived by the two individuals) leads to a stronger social tie. This reasonable statement characterizes an *informational dimension* of social ties, in the sense that two individuals can be more socially connected simply by commonly acquiring relevant information about each other's social identities.

However, we believe that there exists another important aspect influencing the strength of a social tie between two individuals: the quantity and quality of past interactions between them. Indeed, a tie between two persons can reasonably become stronger if they interact frequently and/or share positive meaningful experiences (e.g., exchanging ideas, opinions, emotions, etc.) with each other. This characterizes an *experiential dimension* of social ties.

In this study, we aim at testing the effects of those two factors by (1) exogeneously manipulating individuals' knowledge about each other's social identities (*objective social ties*), and (2) measuring their own perception of social closeness with one another (*subjective social ties*).

3 Experimental Design

In the following, we will introduce two coordination games to study social ties: the first is a variant of the battle of the sexes game that we call the '**baseline game**'; the second extends the previous game by adding an outside option and is named the '**entrance game**'.

3.1 The baseline game

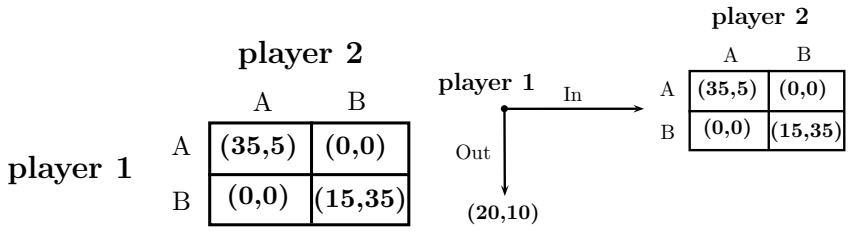
The coordination game that we consider as baseline is an **asymmetric battle of the sexes game**. It is a simultaneous move game with two players (1 for row and 2 for column), each of which has to choose between two available actions A and B. The corresponding payoff matrix is represented in Figure

1(a). As in the traditional battle of the sexes, the worst scenario for both players is to miscoordinate (i.e., playing A while the other plays B or vice versa). Furthermore, the players have diverging preferences regarding the best outcome for themselves: player 1 prefers coordination on (A,A) while player 2 prefers coordination on (B,B). However, unlike the classical battle of the sexes game, the lowest payoff is different in the two coordination outcomes: outcome (B,B) is worth more to player 1 than outcome (A,A) is worth to player 2.

In spite of this difference, the game theoretic properties of this asymmetric battle of the sexes game remain as in the classical case: both (A,A) and (B,B) are the only pure-strategy Nash equilibria, which also are the only Pareto optimal outcomes.[2]

The main feature of this game lies on the impact of group preferences on players' behavior. As in the classical battle of the sexes, being self-interested is not sufficient to guarantee coordination success. However, in our asymmetric game, one can notice the existence of a unique group beneficial outcome for the group, which is not present in the classical battle of the sexes game: out of the two pure-strategy Nash equilibria, the outcome (B,B) seems better for the group. Whether one considers the sum, the average, the difference, or the minimum value among the individual payoffs as a measure of the group's utility, this outcome always outperforms every other solution. In fact, the asymmetry in the players' payoffs provides a clear possibility of favoring the group as a whole, which can allow them to also maximize their self interest (any coordination is always better than miscoordination). Both players may then consider this salient solution as a coordination device. However, one should note that, as the corresponding solution (B,B) favors player 2 more than it favors player 1 (what is best for the group is also best for player 2), coordination is not guaranteed. We will investigate whether participants in the role of player 1 detect and follow the salient outcome (B,B), and which factors weaken or strengthen the focus on it.

[2]There also exists a Nash equilibrium in mixed strategies, which consists of playing A with probability 7/8 for player 1 and playing B with probability 7/10 for player 2 (in this case, the respective expected payoffs are 10.5 for player 1, and 4.4 for player 2).



(a) The baseline game (asymmetric battle of the sexes game).

(b) The entrance game (asymmetric battle of the sexes game with outside option).

Figure 1: Baseline and entrance game.

3.2 The entrance game

We extend the baseline game presented above to the 'entrance game' by adding an outside option (see Figure 1(b)), i.e. **asymmetric battle of the sexes game with outside option**. In this two-player game, prior to playing the coordination game itself, player 1 is offered the possibility of a fixed outside option. If he chooses to enter the coordination game (play 'In') both players play the asymmetric battle of the sexes game as described in the previous section. If he takes the outside option (play 'Out'), the game ends with player 1 earning 20 and player 2 earning 10.

The outside option makes participation in the coordination game a voluntary decision by player 1. Entering the game can therefore be interpreted as a signal to play a certain strategy. How this signal is interpreted will depend on player 2's beliefs about player 1's motivations: specifically if player 1 is expected to be self-interested or to take the group interest into account.

Before expanding this forward induction argument, let us notice that the entrance game contains three Nash equilibria in pure strategies: ((In,A),A), ((Out,A),B), ((Out,B),B).[3] Considering subgame perfect Nash equilibria by backward induction allows us to rule out the solution ((Out,A),B).[4]

[3]Moreover, the entrance game also has Nash equilibria in mixed strategies, which consist of player 1 always playing Out (i.e., selecting either strategy (Out,A) or strategy (Out,B) with probability 1) and player 2 playing B with probability 3/7.

[4]There also exists a Nash equilibrium in behavioral strategies, which consists of player 1 always choosing Out first and playing B with probability 1/8 in the subgame; while player 2 plays B with probability 7/10.

Forward induction then allows us to restrict the set of subgame perfect Nash equilibria to those solutions that survive the iterated elimination of weakly dominated strategies. Initially player 1's strategy (In,B) is weakly (and strictly) dominated by any strategy involving Out. Then player 2's strategy B becomes weakly dominated by A. Thus player 1's strategies (Out,A) and (Out,B) are both weakly (and strictly) dominated by (In,A). Therefore, the unique forward induction solution is ((In,A),A).

Indeed, assuming common knowledge that both players are fully rational and motivated by their own self interest, this solution should be played. When playing In, player 1 signals that he intends to play A in the subgame (if he intended to play B, he would have been better off playing Out in the first place). Therefore, as a best response, player 2's unique rational move is to play A. Finally, since outcome ((In,A),A) is better for player 1 than selecting Out, he chooses (In,A).[5] This therefore suggests that such a game introduces some 'first mover' advantage, assuming there is common knowledge that both players are self-interested agents.[6] Let us also point out that such a forward induction argument has already received wide experimental support in the literature (e.g. Brandts and Holt, 1989; Cooper et al., 1992, 1993; Van Huyck et al., 1993; Balkenborg, 1994; Brandts and Holt, 1995; Cachon and Camerer, 1996; Shahriar, 2009).

However, if players focus on some collective goals and expect others to do the same, entering the subgame will be associated with a choice of B. As a result, stronger social ties with a group might lead to either effect: a stronger belief in individual rationality of partners that are more identifiable (group members) or a stronger belief in collective rationality by group members. In the former case, ((In,A),A) will be played. In the latter case the outcome will be ((In,B),B).

Specifically, the entrance game will enable us to observe if players linked by stronger social ties are more likely to expect coordination in the subgame and

[5]No Nash equilibrium in mixed/behavioral strategies does resist this forward induction argument.

[6]The forward induction argument holds also if player 2 is not a self-interested agent, provided that he prefers outcome ((In,A),A) to outcome ((In,A),B).

therefore more likely to enter the second stage of the game (when acting as player 1). In turn, reactions by player 2 will allow us to investigate whether and how – via coordination on either (A,A) or (B,B) – this intention is understood. The baseline game will serve us as control to see whether the first stage is always needed to signal intentions.

3.3 Varying social ties

We vary the strength of social ties by considering partners that come from more or less strongly linked 'in-groups'. This manipulation aims at controlling for the *informational dimension* of social ties introduced in Section 2. More precisely, we investigate three levels of 'in-groups' (*objective ties*).

The weakest level of social ties concerns our treatment *university*. In this treatment, participants know that they will interact with a fellow student from their own university. Note that this is the default in most laboratory experiments and therefore the possibility of social ties between such participants is assumed to be minimal.

Our strongest level of social ties concerns our treatment *team* in which two players from the same volleyball team interact. Teams consist of 7 to 9 players and meet at least once per week for a two-hour training session. Note that interactions were anonymous in the sense that no participant could identify his interaction partner from the game. However it was common knowledge that both participants were members of the same team.

A third treatment gives some intermediate level of social ties: *club*. Here both participants were members of the same volleyball club, but not playing in the same team. The club had around 70 members and members might interact before and after training with players that were not from their own team.

We further elicited through questionnaires how well the players saw themselves linked to their teammates and how they considered the relationship between their teammates. We use these subjective measures to determine social ties through their *experiential dimension* as presented in Section 2. More specifically, participants responded to two types of scales (*subjective ties*).

The first scale (direct scale) asked with respect to each team member 'how do you think this person feels about you?' (see Figure 7 in the Appendix for an example). Participants could choose between 'likes me a lot', 'likes me', 'dislikes me' and 'is indifferent'.[7] Answers to this measure allow us to determine how well the individual feels 'liked' and thus connected to his team (index of subjective connectedness of the self to the group). Specifically, for every participant i, the corresponding coefficient of i's belief about *self* connectedness to the group G ($i \in G$) is determined by:

$$k_i^S = \frac{N_i}{|G| - 1}$$

where $|G|$ denotes the size of the team and N_i defines the number of individuals in G that participant i believes to strongly like him. Specifically, N_i indicates how many times the answer 'the other likes me a lot' was selected by i in the questionnaire. Note that k_i^S simply stands for the probability that individual i interacts with a person he believes to strongly like him.

Let us also define the average self connectedness K^S within the group G as follows:

$$K^S = \frac{\sum_{i \in G} k_i^S}{|G|}$$

We will use this measure later to determine whether an individual scores more or less high concerning beliefs about his own popularity compared to his team mates.

The second scale (indirect scale) aimed at eliciting ties between team members as perceived by the participant. To do so, each participant i was asked to indicate for any two members of his team whether he considered them to be 'friends' (for an example, see Figure 8 in the Appendix). The scale was presented in a visual intuitive form with all team members' photographs arranged in a circle, where participants were asked to indicate by a line any two

[7] Participants also answered for each team member whether they 'liked a lot', 'liked' or 'disliked' this person. Answers were strongly correlated with the indirect question.

members they thought to be friends (excluding themselves).[8] In the example from Figure 8, individual C responds to the questionnaire and indicates her belief that F is friend with A and G, that G is also friend with E, and that D and B are friends.

Based on answers to this measure, we construct an index of the individual belief about the groups connectedness k_i^G. Specifically, we hypothesize that in our game, behavior does not only depend on the individual's closeness to every other member, but also on the belief about every other member's closeness to each other. To illustrate this assumption, imagine a group of four individuals (Alice, Bob, Carol, and Daniel) and suppose it is common knowledge that Alice is equally close to Bob, Carol, and Daniel, while these three characters are not tied with each other. In the case where every individual is equally likely to interact with any other group member, Alice is indifferent between interacting with the three others (she is sure to interact with someone she is tied with). However, Bob, Carol, and Daniel are not indifferent: they all prefer to interact with Alice, which turns out to be a rather unlikely event with probability $p = 1/3$. As a result, Bob, Carol, and Daniel can be seen as weakly tied with the group. Concerning coordination, Alice thus needs to take this into account and should act as if she is a weekly tied participant (if she does not, she exposes herself to the risk of performing some group-regarding behavior that is not reciprocated).

For every participant i, the corresponding coefficient of i's belief about the *group* connectedness is calculated as follows:

$$k_i^G = \frac{N_{-i}}{M}$$

where N_{-i} represent the estimated number of links in the group G (according to i's beliefs) that do not involve i,[9] and M corresponds to the maximum number of individual links that are possible in the group without individual i:

[8] During the experiment, participants were notified that any link that would involve themselves in this question would simply be ignored.

[9] A link not involving player i is a connection between two players j and h, where j and h are different from i.

$$M = \binom{|G|-1}{2} = \frac{(|G|-1)\cdot(|G|-2)}{2}$$

Note that k_i^G resembles the concept of a *local clustering coefficient*, which characterizes the probability that two randomly selected neighbors of i are tied with each other (Watts and Strogatz, 1998; Newman, 2003). As an illustrative example from Figure 8 in the Appendix, assuming the corresponding answer was made by individual C, we would obtain $k_C^G = \frac{4}{21}$.

Let us also define the average group connectedness K^G within some group G as follows:

$$K^G = \frac{\sum_{i \in G} k_i^G}{|G|}$$

4 Experimental Procedure

Participants in our experiments were students from the University of Toulouse (Capitole) who were also members of the main university volleyball club. During a preliminary meeting, every active member of the club was proposed to participate in our study. Upon acceptance, every participant was then photographed for later use in the questionnaire (see Figures 7 and 8 in the Appendix for examples).

The experiment was run in November 2011 during two training sessions. In total, 70 subjects participated (37 men and 33 women). At the beginning of the academic year (September 2011), volleyball players within the club were divided by the coach into 9 single-sex teams: 5 male teams and 4 female teams. Each team had between 7 and 9 members. Students of the same gender were ranked according to their initial skills at playing volleyball: the best players were assigned to team 1, the next best players to team 2, and so on. The coach enforced assignment to teams such that no switch between members of different teams was allowed along the academic year. For instance, any player's request to join a particular team because a friend belonged to that team would be declined.

Another 43 students were recruited from the same university as partners for the game played with another university student. Data from these observations are not discussed in this paper.

The experiment was run by paper and pencil during training sessions. Subjects first filled out a demographic questionnaire and answered to the direct and indirect scales for social ties. Social ties were elicited before the games were played to prevent an impact of game behavior on social tie measures. Presenting the questionnaire before the game further ensures that participants were aware of their social ties to the team and that they were aware that other participants had also been asked the same questions. This common knowledge assumption is indeed part of the minimum criterion for the existence of a social tie, as defined in Section 2.

Every participant was then asked to report strategies for the baseline game and the entrance game according to three different types of reference groups. All treatment comparisons are therefore on a within-subject level. Indeed, within-subject comparison seems necessary for our research question, since social ties are necessarily individual characteristics. To control for order effects between the baseline game and the entrance game, the order of games was counterbalanced across subjects. The detailed instructions of both games are described in Sections B.2 and B.4 of the Appendix.

The three in-group treatments (team, club and university) were played by every subject and the order was inverted for half of the sessions. However, since answers were given by paper and pencil, participants were free to answer these questions in any order they wished. Participants responded by meta-strategy method for each possible treatment, i.e., all subjects had to indicate their decision if assigned the role of player 1, as well as their decision if assigned the role of player 2. This was made for both the baseline and the entrance game, and for each possible treatment, i.e. if playing with a university student, a club member, or a teammate (12 decisions as a whole).[10] Participants were

[10]We acknowledge that our strategy elicitation method might lead to an experimental demand effect: a subject being asked similar questions that only differ for the interaction partner – team, club, or university member –, he could report different strategies for different partners. However, the fact that the strategy elicitation is monetarily incentivized should

informed that only upon completion of the questionnaire, their role, the game, and the treatment selected for payout would be randomly determined.

The experiment lasted approximately one hour. Earnings were payed out during the next training sessions in December 2011. The payment method, which was specified to all subjects in the instructions (see Section B.1 from the Appendix), consisted of randomly drawing one role (i.e., player 1 or player 2), one game (i.e., entrance game or baseline game), one treatment (i.e., university, club, or team), and one co-player (depending on the treatment). A subject's payoff was therefore defined according to his choice made as the selected player in the selected situation (which corresponds to the selected treatment in the selected game), and the selected co-player's choice in the same situation. Each effective payment was made individually and anonymously through random draws taken in front of the concerned participant.[11] Earnings included a 5 euros show-up fee. Mean earnings were about 19 euros[12] (standard deviation of 12 euros, with a maximum of 40 euros and a minimum of 5 euros).

5 Results

We will start the analysis by considering differences across the three treatments (determining different types of partners). In addition to this exogenous variation of ties to the interaction partner, we will in a second part consider whether similar patterns can be observed when considering the subjective measures of social ties as defined above (through coefficients k_i^S and k_i^G). Since we observe no significant effect of team gender and team rank over subjects' behavior in all role-game-treatment situations, we will in the following pool data from the nine different teams.

mitigate this problem. Furthermore, the experimental demand in our study is not easy to detect, and does not necessarily require different strategies if matched with different partners. For instance, subjects with low subjective ties (e.g., group connectedness) should not choose strategy (In, B) in the entrance game independently of the treatment (see Figure 5 in Section 5).

[11]The random selection of the co-player was made through a random code name to preserve anonymity between subjects.

[12]Approximately 25 US dollars at the time of the experiment (1 euro = 1.4 US dollars).

5.1 Objective social ties

We first present descriptive statistics concerning the players' behavior in both the baseline game and the entrance game, for the three treatment scenarios (i.e., team, club and university). Note that in this case, the social ties are considered objective as their strength is exogenously controlled by changing the type of a participant's interaction partner. Since we observe no order effect regarding which game or treatment was presented first, we will in the following pool data from the different sessions.

5.1.1 Baseline game

We present choices in the baseline game, depending on whether the corresponding co-player is a teammate, a club member, or a fellow university student in Figure 2 (detailed results can be found in Table 4 of the Appendix).

Let us recall the predictions concerning the impact of social ties for player 1 and player 2. Specifically, for player 2, the own payoff maximizing outcome coincides with the outcome that is best for the group (i.e. (B,B)). Meanwhile player 1 faces a choice between the own payoff maximizing outcome (A,A) and the outcome that is considered as best for the group (B,B). Increasing social ties is therefore expected to increase the percentage of players 1 choosing option B. This is indeed what we observe. As we see in Figure 2, an increasing percentage of players 1 select option B when the social tie with the interaction group increases. We can reject the null hypothesis that player 1 is making the same choice when paired with a university student as when interacting with a teammate (Wilcoxon signed rank test, $p = 0.002$). For the intermediate level of the social tie (i.e. the club treatment), behavior is situated between the two extremes.

When acting in the role of player 2, subjects clearly favor option B in all types of interactions.[13] Varying the strength of the social tie has no impact on choices by player 2. While this is in line with the prediction that self interest

[13]Note that in this case, player 2's average behavior is close to the optimal mixed strategy i.e., playing B with probability 7/10.

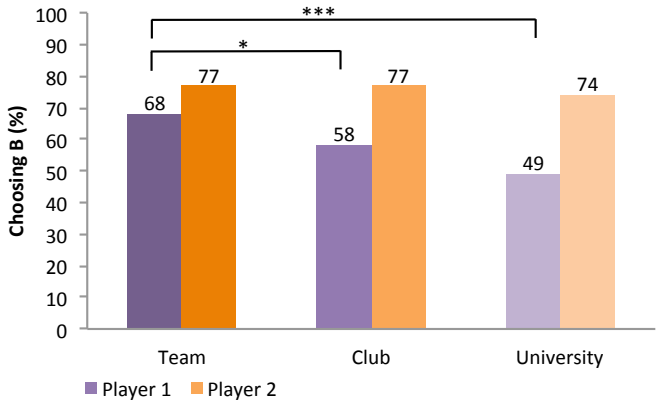


Figure 2: Behavior in the baseline game for all players in each type of matching. Significance levels based on Wilcoxon signed rank tests: $p < 0.01$ (***); $p < 0.1$ (*). Data recorded by meta-strategy method thus each bar consists of 69 observations.

and group interest are not at conflict for player 2, this also implies that they do not seem to anticipate player 1 being influenced by the strength of the social tie. We will next use our observations from the entrance game to see whether player 2 is more likely to take the treatment difference into account when he knows that player 1 has to make an active choice to participate in the coordination game.

5.1.2 Entrance game

In the context of the entrance game, our first observation is that participants interacting with a team member in contrast to a university student are significantly more likely to enter the second stage of the entrance game (Wilcoxon signed rank test: $p = 0.004$). Recall that agents will only enter the second stage of the game if they believe that this will lead to an outcome that is on some dimension preferable to the outside option. Under the assumption that others will maximize own income and that others expect the agent to do the same, this might lead to the forward induction reasoning that results in choosing A in the subgame. If however agents focus on some collective goals

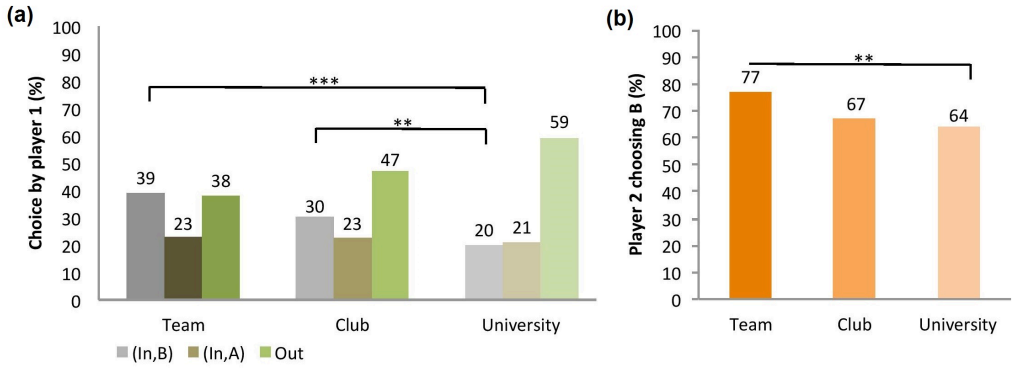


Figure 3: Behavior in the entrance game. Significance levels based on Wilcoxon signed rank tests: $p < 0.01$ (***); $p < 0.05$ (**). Data recorded by meta-strategy method thus 69 observations per treatment and player role.

and expect others to do the same, entering the subgame will be associated with a choice of B. As a result, stronger social ties with a team might lead to either effect: a stronger belief in individual rationality of partners that are more identifiable (teammates) or a stronger belief in collective rationality by team members.

We present choices concerning both roles in the entrance game, depending on whether interacting with a teammate, a club member, or a fellow university student in Figure 3. For player 1, we focus on strategies (In,A), (In,B), and Out.

Figure 3(a) allows us to reject the hypothesis that social ties promote forward induction focused on individual rationality. Indeed, the proportion of subjects selecting strategy (In,A) is similar in all treatments. On the other hand, Figure 3(a) shows that subjects are significantly more likely to choose (In,B) when interacting at the team level than at the university level (39% vs 20%, Wilcoxon signed rank test: $p = 0.003$). This shows a significant fraction of participants that switch from selecting Out when interacting with a fellow university student, to selecting (In,B) when interacting with an individual from their team.

We further observe no significant difference between player 1's behavior in

the second stage of the entrance game (i.e., after choosing In) and choices in the baseline game from Section 5.1.1. Among the subjects who played In in the first stage, we observe that option B is selected for: 63% (team), 56% (club) and 48% (university) of participants. As this behavior is very similar to that in the baseline game from Figure 2 (Wilcoxon signed rank tests: $p = 0.405$ in team treatment, $p = 0.527$ in club treatment, $p = 0.257$ in university treatment), we conclude that the outside option of the entrance game has only a negligible effect on player 1's behavior in the coordination game. In other words, right after playing In, player 1 tends to consider the subgame as a new independent game. We will get back to this observation in Section 5.3 when discussing the joint meta-strategy behind the veil of ignorance whether the agent will act as player 1 or 2.

We now turn to the question of whether choices in the role of player 2 are also unaffected by the outside option. Matched with fellow university students, we observe that choices as player 2 are indeed influenced by the outside option as forward induction would assume. Specifically 64% of players 2 choose B in the entrance game, while 74% choose this option in the baseline game (Wilcoxon signed rank test: $p = 0.07$). We further observe from Figure 3(b) that, when playing with a team member, player 2 chooses B significantly more often than when interacting with a university student (Wilcoxon signed rank test: $p = 0.049$). These results therefore confirm the hypothesis that social ties help people to coordinate on the most group beneficial outcome ((B,B) in the baseline game, (In,B,B) in the entrance game).

5.2 Subjective social ties

We will now extend our analysis to include the subjective measures of social ties as defined in Section 3.3. As discussed previously, every subject in our experiment was asked to provide subjective information about whether they believed their teammates to like them and how much they considered their teammates to be friends with each other. Using these answers, we calculate two subjective measures of social ties for each individual i: k_i^S and k_i^G.

	Category	Condition	N
(a)	H^S	$k_i^S \geq K^S$	38
	L^S	$k_i^S < K^S$	31
(b)	H^G	$k_i^G \geq K^G$	30
	L^G	$k_i^G < K^G$	39

Table 1: Classifications based on subjective reports relative to average in group concerning (a) self connectedness (K^S), and (b) group connectedness (K^G).

To analyze the relation between these measures and behavior in our games, we categorize participants as ranking either above (H^S and H^G) or below (L^S and L^G) the average answers in their own team (i.e., K^S and K^G respectively). By doing so, we avoid the possible confound that some teams might be more closely tied than others and focus on the relative part of the measure [14].

Table 1 summarizes the classification with respect to the group average concerning (a) self connectedness (K^S) and (b) group connectedness (K^G). The two measures show no statistically significant correlation (Pearson's chi-squared test: χ^2=1.464, $p = 0.226$). We therefore consider these two types of measures separately throughout the following analysis.

The objective of the next sections is to identify subjective measures that allow us to replicate the previous observed results for the objective variation of social connectedness. While the previous section focused on a comparison between participants paired with team members, club members, or fellow university students, this section will compare behavior at the *team level* for participants scoring either high or low on the different subjective measures.

5.2.1 Baseline game

We start our analysis with observations from the baseline game. Recall that in Section 5.1.1 we found a significant treatment effect for choices of player 1. We will thus focus, in the following, on choices by player 1 when interacting with another team member. Our comparison will be between individuals cat-

[14]We find that high-rank teams score higher on average on the self-connectedness scale than low-rank teams.

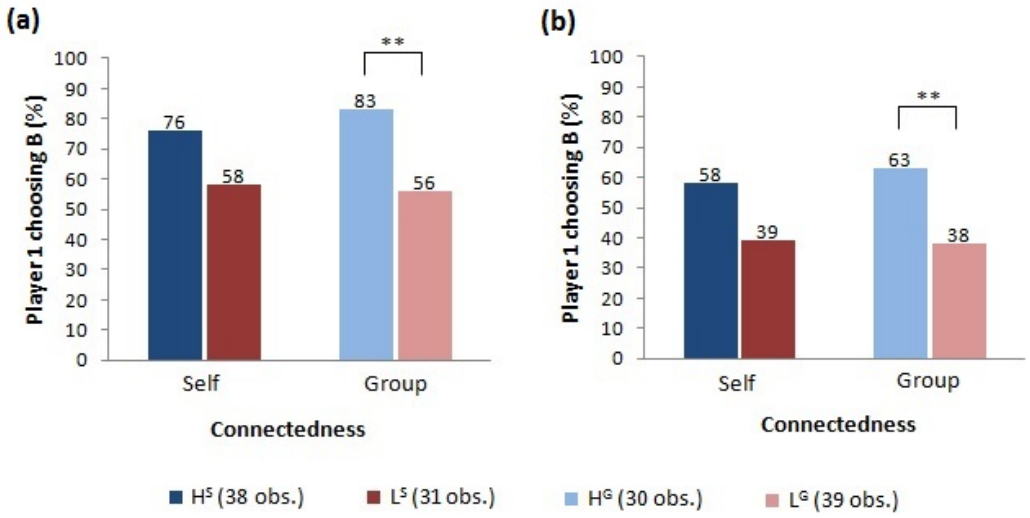


Figure 4: Player 1's behavior in the baseline game. (a) Team treatment; (b) University treatment. Significance levels based on Mann-Whitney tests: $p < 0.05$ (**). Data recorded by meta-strategy method thus 69 observations per treatment.

egorized as either 'high' or 'low' for either our measure of self connectedness (H^S and L^S respectively) or group connectedness (H^G and L^G respectively). As a control, we will also present results concerning choices when interacting with a fellow university student. Recall that both connectedness measures were recorded with respect to team members. A good measure of the social connectedness at the team level should therefore show no effect on choices when interacting with an individual at the university level.

Figure 4(a) summarizes choices for self and group connectedness at the team level (see also Tables 5 and 6 in the Appendix). We observe an effect for both self and group connectedness, marginal in the first case (Mann-Whitney test comparing H^S vs L^S: $p = 0.109$) and significant in the latter (Mann-Whitney test comparing H^G vs L^G: $p = 0.018$). In both cases, participants reporting high values concerning connectedness are more likely to choose the group beneficial outcome B.

Looking at choices by the same groups of participants in the university treatment (see Figure 4(b)), we can now disentangle whether the effect is due

to an increased concern for others in the team or whether the measures also pick up some other effect. Indeed in both cases, the same tendency observed in the team treatment is also observed in the university treatment (Mann-Whitney test in the university treatment: H^S vs L^S : $p = 0.115$; H^G vs L^G: $p = 0.042$). Both measures thus seem to be related to individual characteristics that lead agents to make more group beneficial choices in general. One possible interpretation could be that self/group connectedness among our participants may be correlated with other-regarding traits. An alternative interpretation is that these individuals can more easily detect (B,B) as the unique group beneficial outcome that might help solve the coordination problem. Following this interpretation, they play B to maximize the welfare of the group in the 'team' treatment and to maximize their own individual payoff in the 'university' treatment. Indeed, it is worth recalling that, besides being the fairest outcome, (B,B) is also a Nash equilibrium in the baseline game. To distinguish between these two hypotheses, we will next analyze the results from the entrance game.

5.2.2 Entrance game

Recall that in the entrance game, a treatment effect was observed as well for player 1 as for player 2 when interacting with a team member (see Section 5.1.2). The focus by both players on action B is striking given that the outcome ((In,B),B) is not a Nash equilibrium.

As in the previous section, we now present behavior by player 1 and player 2 in the team treatment in Figures 5(a) and 6(a) respectively (see also Tables 7-12 in the Appendix). We observe from Figure 5(a) that both the group and the individual connectedness measures are correlated with player 1's decision to select strategy (In,B) (self connectedness: 50% vs 26%; Mann-Whitney test: $p = 0.042$; group connectedness: 57% vs 26%; Mann-Whitney test: $p = 0.009$). However, when we turn to behavior as player 2, we observe a difference between the two measures (see Figure 6(a)). Recall that player 2 in this game has to understand the possible reasons that might lead player 1 to enter the game. Behavior as player 2 is very similar no matter the reports of

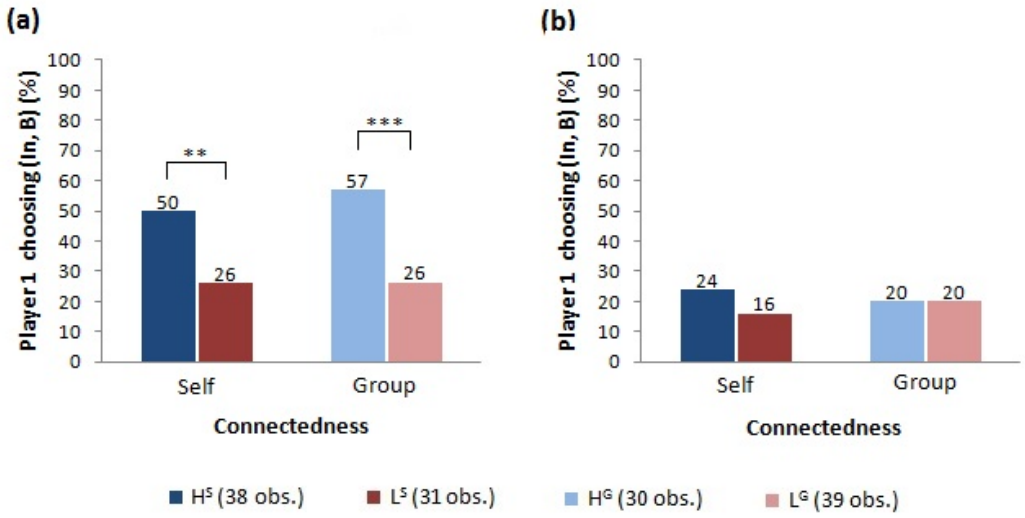


Figure 5: Player 1's behavior in the entrance game. (a) Team treatment; (b) University treatment. Significance levels based on Mann-Whitney tests: $p < 0.05$ (**); $p < 0.01$ (***). Data recorded by meta-strategy method thus 69 observations per treatment.

the individuals' self connectedness (proportion of selecting B: 76% vs 77%). Thus believing that one is liked by many other players is not sufficient to conclude that these other players would in an anonymous interaction take a risky choice (B vs A) with *any* person from the team. Meanwhile, we observe that the group connectedness measure is significantly correlated with player 2's choice (proportion of selecting B: 90% vs 67%; Mann-Whitney test: $p = 0.024$). Thus players that believe in a high interconnectedness in their team are more likely to believe that player 1 enters the game to play the group beneficial outcome.

As before, we can also compare these results to behavior when interacting with a fellow university student. We observe from Figures 5(b) and 6(b) that there exists no significant correlation between the connectedness measures and choices at the university level. This clearly indicates that subjective beliefs concerning self and group connectedness are only related to behavior (as player 1 and player 2) when interacting with a team member. It thus seems that subjective beliefs about self and group connectedness are rather correlated

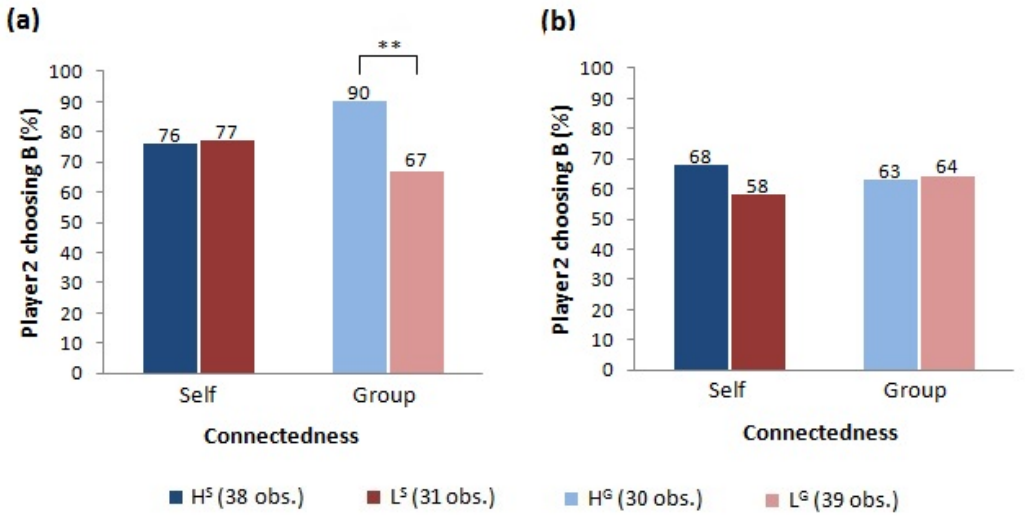


Figure 6: Player 2's behavior in the entrance game. (a) Team treatment; (b) University treatment. Significance levels based on Mann-Whitney tests: $p < 0.05$ (**). Data recorded by meta-strategy method thus 69 observations per treatment.

with an ability to identify the focal nature of the (B,B) outcome and not to a higher level of fairness (see hypotheses at the end of Section 5.2.1).

Results in Figures 5 and 6 further show that subjects who perceive low social connectedness within their team behave similarly with a teammate and a random university student in the entrance game. This observation clearly indicates that increased coordination on the group beneficial outcome in the team treatment is not driven by some notion of group identity, but instead results from the subjective perception of social closeness (and in particular group connectedness). In other words, sharing the same social features that characterize the players' social identities (informational dimension of social ties) is alone insufficient to induce effective coordination in this context. Sharing some positive experience is also required.

Strategies	treatment		
	team	**club**	**university**
(A,A)	13%	13%	12%
(A,B)	19%	29%	**39**%
(B,A)	10%	10%	14%
(B,B)	**58**%	**48**%	35%

Table 2: Meta-strategies in the original position of the baseline game across treatments (69 observations per treatment).

5.3 Behind the veil of ignorance

Recall that in our experiment, strategies were elicited by meta-strategy method for the case of being selected as either player 1 or player 2. This allows us to add to the previous discussion, an analysis of behavior in the 'original position' of the meta-game before actual roles were assigned (Rawls, 1971).

To analyze behavior in the meta-game, we need to consider equilibria for the higher order symmetric game. We will denote strategies for this game as (x_1, x_2), where x_1 indicates the choice when assigned to the role of player 1 and x_2 the choice for the role of player 2.

In the baseline game, four distinct strategies exist: (A,A), (B,B), (A,B) and (B,A). The payoff matrix concerning expected earnings from the transformed game can be found in Table 13 of the Appendix. We can easily see that there exist three different pure-strategy Nash equilibria: (1) both players selecting (A,A); (2) both players selecting (B,B); and (3) one player selecting (A,B) while the other chooses (B,A). Note that the third solution is not consistent with making a decision in Rawls' original position, since the latter implies to select the same strategy that one expects others to select. Furthermore, of all the above strategies, only (A,A) and (B,B) are evolutionary stable.

Observed behavior of meta-strategies for the baseline game is shown in Table 2. When interacting with another university student, 39% of participants select strategy (A,B) in this game. Note that this is coherent with participants expecting their interaction partner to act differently (e.g. to choose (B,A)). In other words, participants seem to strongly identify with each player role (i.e.,

they do not use Rawls' original position to make their decision): when they act as player 1, they do not consider what they would do as player 2, and vice versa.

However we observe a change in behavior when we consider the team treatment. A Wilcoxon signed rank test indeed indicates that (B,B) is selected significantly more often in the team treatment (57%) than in the university treatment (34%, $p < 0.001$). Considering the independence of strategies played in the role of player 1 and player 2 further emphasizes this result. We observe no correlation in the case of the university treatment (Pearson's chi-squared test, χ^2=0.384, $p = 0.535$) but a significant correlation in the team treatment (Pearson's chi-squared test, χ^2=5.694, $p = 0.017$). For further details, see Table 15 in the Appendix. In other words, these results suggest that increasing social ties leads people to take Rawls' original position into account.

Similar results can be obtained for the entrance game. In this case, six distinct strategies need to be considered (see Table 14 for the payoff matrix).[15] The transformed entrance game has three pure-strategy Nash equilibria: (1) both players selecting ((In,A),A); (2) both players selecting (Out,B); (3) one player selecting ((In,A),B) while the other chooses (Out,A). As in the baseline game, the latter solution is not consistent with making a decision in Rawls' original position. In this case, of the six strategies available, only ((In,A),A) is evolutionary stable.

Similarly to the baseline game, we observe that, in the entrance game, social ties still lead people to act as if they were in the original position (see Table 3). A Wilcoxon signed rank test again reveals that ((In,B),B) is selected significantly more often in the team treatment (35%) than in the university treatment (13%, $p < 0.001$). More precisely, in the case of the university treatment, we observe no correlation between players' choices in both roles (Pearson's chi-squared test, χ^2=0.950, $p = 0.622$). However in the team treatment, a significant correlation is observed (Pearson's chi-squared test, χ^2=8.897, $p = 0.012$). For further details, see Table 16 in the Appendix.

[15]For simplicity, we omit counterfactual strategies (i.e., ((Out,A),·) and ((Out,B),·)) that are irrelevant to this analysis.

Strategies	treatment		
	team	**club**	**university**
((In,A),A)	12%	7%	10%
((In,A),B)	12%	16%	12%
((In,B),A)	4%	9%	7%
((In,B),B)	**35**%	22%	13%
(Out,A)	7%	17%	19%
(Out,B)	30%	**29**%	**39**%

Table 3: Meta-strategies in the original position of the entrance game across treatments (69 observations).

Specifically note that in the team treatment, the fairest outcome ((In,B),B) becomes the modal choice.

Finally we consider the implications of these results with respect to our subjective measures of connectedness in the particular case of interactions between teammates. In the context of the baseline game, being closely tied with other team members according to the group connectedness measure makes participants select (B,B) significantly more often (73% in group H^G; 46% in group L^G; Mann-Whitney test: $p = 0.024$). On the other hand, this effect does not replicate through the alternative self connectedness measure (63% in group H^S; 52% in group L^S; Mann-Whitney test: $p = 0.337$). Furthermore, looking at behavior in the entrance game reveals similar results: being closely tied with other team members according to the group connectedness measure makes participants select ((In,B),B) significantly more often (57% in group H^G; 18% in group L^G; Mann-Whitney test: $p < 0.001$). Unlike in the baseline game, using the self connectedness measure replicates this effect (47% in group H^S; 19% in group L^S; Mann-Whitney test: $p = 0.016$). These results, which are illustrated in greater details through Figures 9 and 10 from the Appendix, indicate that both self and group connectedness lead to choices that are more in tune with choices that should be taken in Rawls' original position.

6 Discussion

The experimental study presented in this paper provides evidence that an increase of (objective) social ties through exogenously assigned groups involving real social interactions can help individuals solve asymmetric coordination problems with a unique clearly identifiable best outcome for the group.[16] This evidence is stronger for group members perceiving higher (subjective) social ties among them.

In this section, we will discuss whether and in which measure relevant theories in the literature can explain the effect of social ties observed in our study.

Theories of social preferences cannot fully explain the effect of social ties that we observe. For example an increase in inequity aversion (Bolton and Ockenfels, 2000; Fehr and Schmidt, 1999) would predict the choice of Out for player 1 in the entrance game. Our results show the opposite tendency. Alternatively, assuming that social ties correlate with stronger reciprocal fairness (Charness and Rabin, 2002) would predict high degrees of miscoordination in the baseline game.[17] Again we observe the opposite result. Intuitively, such theories of social preferences are not best candidates to reasonably explain coordination in the above games because they all rely on the assumption that people make decisions with the individual intention to promote fairness (e.g., acting as a benefactor). This kind of *benefactor behavior* (Bacharach, 1999) is indeed not sufficient because it still requires each player to think about each other's possible actions in order to coordinate (in the baseline game, one may think that B is the fairest option if the other chooses B, and that A is the

[16] It is worth noting that our results cannot be generalized to other interactive situations that do not satisfy this constraint. For example, it is not difficult to define an asymmetric coordination game where a particular outcome maximizes the total payoff of the pair whereas another maximizes equality in payoffs. In this case, what is the most group beneficial outcome is not clear anymore, and therefore, social ties may not be sufficient to increase coordination in this context. We however postpone further investigation of such scenarios to future work.

[17] According to the model in Charness and Rabin (2002), both outcomes (A,A) and (B,B) are always Nash equilibria of the baseline game, no matter how strongly motivated players are to maximize the welfare of the group.

fairest option if the other chooses A). For more details, see Attanasi et al. (2014).

Coordination on the (B,B) outcome in the baseline game, and on the ((In,B),B) outcome in the entrance game can however be explained by theories of team reasoning (Bacharach, 1999; Sugden, 2000, 2003). When an individual engages in team reasoning, he identifies himself as a member of a group and conceives that group as a unit of agency acting in pursuit of some collective objective (Sugden, 2000). In other words, such an individual will act for the interest of his group by identifying and implementing a strategy profile that maximizes the collective payoff of the group. This type of thinking is clearly distinct from that assumed in the theories of social preferences previously discussed. Under Bacharach's concept of *unreliable team interaction* (Bacharach, 1999), which relies on such team reasoning, a given player identifies with a team with a certain probability p and chooses the action which maximizes the team benefit. With probability $1 - p$ the player is self-interested and maximizes his own benefit. In the context of our experiment, given a sufficiently high probability of team reasoning, the players should coordinate on the (B,B) outcome in the baseline game, and on the ((In,B),B) outcome in the entrance game. Following this theory, our results imply that players in the team treatment are more likely to use team reasoning than in the university treatment (especially when they believe in strong group connectedness within their team). However, note that the theory only considers binary types of reasoning: either one follows team reasoning, or not. Yet, given the multiple levels of self evoked by social identity theory, we might consider a more gradual notion of group identification. As a result, existing theories of team reasoning fail to fully capture the possibility of different levels of social connectedness (see Attanasi et al. (2014) for more details regarding team reasoning and its main limitation in the context of social ties).

Another theory, which can explain the observed behavior in our experiment, is the theory of *empathetic preferences* (see Binmore, 1994, 1998, 2005). Binmore argues that an individual may be equipped with some empathetic preferences, which consist in combining his actual own preferences with his

preferences when *imagining* himself to be in the other person's position. For example, in the context of the baseline game, an empathizing player 1 would compare his own preferences while being himself (i.e., player 1) with his preferences while being player 2.[18] If making a decision based on such an interpersonal comparison of preferences, player 1 is said to *empathize* with player 2. The idea is thus linked to Rawls' concept of original position (see Section 5.3). According to the analysis of the meta-strategies discussed in Section 5.3, we can thus say that players empathize more with each other in the presence of social ties. However, Binmore's theory can also not fully capture the concept of gradual social ties as it does not quantify the degree of empathizing behavior, that is, how choices are altered for intermediate levels of empathy.

As an alternative, the model of *homo moralis* (Alger and Weibull, 2013, p. 2276) can also be used to interpret our results. Homo moralis faces a trade-off between maximizing his own material payoff, and doing 'the right thing,' that is, choose a strategy that, if used by all individuals, would lead to the highest possible payoff." Given our experimental findings, the degree of morality seems to be stronger in the presence of stronger social ties. It should however be noted that this model assumes a symmetric game structure and thus strictly has to be related to the findings discussed in Section 5.3 (i.e., assuming participants make their decision behind the veil of ignorance).

Finally, in contrast with the above theories that can be described as some kind of *hard-wired* psychological mechanisms, the concept of social norms (Bicchieri, 2006) can also be considered to justify the effects of social ties observed in our experiment. Indeed, such a normative approach follows the idea that a person's utility can be directly influenced by the conformity of his own choices with internalized rules that have been socially defined (i.e., deviating from such norms is costly to the individual). In particular, equilibria of coordination games can be seen as 'norms' insofar as they are unintended collective

[18]Binmore points out that, when projecting himself to be in player 2's position, player 1 must not consider his own preferences as player 1, he must instead imagine himself while having player 2's preferences: since player 2 prefers outcome (B,B) to outcome (A,A), player 1 should share this preference when putting himself in player 2's position, even though he prefers (A,A) to (B,B) as player 1.

outcomes of individual choices (Bicchieri, 2006, p. 51)

It has recently been argued that prosocial behavior could be driven by the desire to adhere to social norms (e.g., Krupka and Weber, 2013): sociality is driven not by preferences over payoffs of others, but rather by preferences for following well-established social rules. These norms specify the most socially appropriate action for an agent in a given strategic setting. Hence, different norms can be active in different contexts, and, within the same context, different subjects can be sensitive to different norms. Following this approach, Kimbrough and Vostroknutov (2015) have elicited individual norm-sensitivity and shown how it relates to play in different social dilemma games. Furthermore, Goette et al. (2012) have shown that membership to real groups as well as minimal groups can lead to different behaviors in terms of norm enforcement: in-group norm violators are more leniently punished than out-group defectors.

Therefore, discussing which norm appears to be active in our experimental game is worth doing. In the context of our experiment, members of a particular volleyball team may learn, throughout repeated training sessions, to enforce the norm of "*maximizing team benefit*" (after all, team performance, not individual performance, is what matters most in such team sports). This rule may then be enforced by subjects when asked to play the above coordination games with one of their teammates. This kind of normative interpretation is clearly plausible to justify increased coordination in the team treatment of our experiment. However, how it can account for the increased coordination between players with intermediate levels of ties is less obvious (Figure 3(a) illustrates that, in the entrance game, players 1 select (In,B) more often in the club treatment than in the university treatment). As for the previously discussed theories, the main challenge of such a normative approach is to fully capture the gradual nature of social ties (e.g., can different levels of ties trigger different social norms?). We postpone such a relevant analysis to future work.

References

Akerlof, G. A. and Kranton, R. E. (2000). Economics and identity. *Quarterly Journal of Economics*, 115:715–753.

Alger, I. and Weibull, J. W. (2013). Homo moralis: Preference evolution under incomplete information and assortative matching. *Econometrica*, 81:2269–2302.

Attanasi, G., Hopfensitz, A., Lorini, E., and Moisan, F. (2014). The effects of social ties on coordination: conceptual foundations for an empirical analysis. *Phenomenology and the Cognitive Sciences*, 14:43–73.

Bacharach, M. (1999). Interactive team reasoning: a contribution to the theory of cooperation. *Research in Economics*, 23:117–147.

Balkenborg, D. (1994). *An experiment on forward versus backward induction*. Rheinische Friedrich-Wilhelms-Universität Bonn.

Bernhard, H., Fischbacher, U., and Fehr, E. (2006). Parochial altruism in humans. *Nature*, 442:912–915.

Bicchieri, C. (2006). *The grammar of society: The nature and dynamics of social norms*. Cambridge University Press.

Binmore, K. (1994). *Game theory and the social contract - Volume 1: Playing fair*. Cambridge, MA: MIT Press.

Binmore, K. (1998). *Game theory and the social contract - Volume 2: Just playing*. Cambridge, MA: MIT press.

Binmore, K. (2005). *Natural justice*. Oxford University Press, USA.

Bolton, G. E. and Ockenfels, A. (2000). A theory of equity, reciprocity and competition. *American Economic Review*, 100:166–193.

Brandts, J. and Holt, C. (1989). *Forward induction: Experimental evidence from two-stage games with complete information*. Departament d'Economia i d'Història Econòmica, Universitat Autònoma de Barcelona.

Brandts, J. and Holt, C. (1995). Limitations of dominance and forward induction: Experimental evidence. *Economics Letters*, 49:391–395.

Brewer, M. B. (1979). In-group bias in the minimal intergroup situation: A cognitive-motivational analysis. *Psychological Bulletin*, 86:307–324.

Brewer, M. B. (1999). The psychology of prejudice: Ingroup love and outgroup hate? *Journal of social issues*, 55:429–444.

Buchan, N., Johnson, E., and Croson, R. (2006). Let's get personal: An international examination of the influence of communication, culture and social distance on other regarding preferences. *Journal of Economic Behavior & Organization*, 60:373–398.

Cachon, G. and Camerer, C. (1996). Loss-avoidance and forward induction in experimental coordination games. *Quarterly Journal of Economics*, 111:165–194.

Charness, G., Rigotti, L., and Rustichini, A. (2007). Individual behavior and group membership. *American Economic Review*, 97:1340–1352.

Charness, G. B. and Rabin, M. (2002). Understanding social preferences with simple tests. *Quarterly Journal of Economics*, 117:817–869.

Chavanne, D., McCabe, K., and Paganelli, M. P. (2011). Whose money is it anyway? ingroups and distributive behavior. *Journal of Economic Behavior & Organization*, 77:31–39.

Chen, R. and Chen, Y. (2011). The potential of social identity for equilibrium selection. *American Economic Review*, 101:2562–2589.

Chen, Y. and Li, S. (2009). Group identity and social preferences. *American Economic Review*, 99:431–457.

Cooper, R., De Jong, D., Forsythe, R., and Ross, T. (1992). Forward induction in coordination games. *Economics Letters*, 40:167–172.

Cooper, R., De Jong, D., Forsythe, R., and Ross, T. (1993). Forward induction in the battle-of-the-sexes games. *American Economic Review*, 83:1303–1316.

Crawford, V., Gneezy, U., and Rottenstreich, Y. (2008). The power of focal points is limited: Even minute payoff asymmetry may yield large coordination failures. *American Econonic Review*, 98:1443–1458.

Fehr, E. and Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics*, 114:817–868.

Fershtman, C. and Gneezy, U. (2001). Discrimination in a segmented society: An experimental approach. *Quarterly Journal of Economics*, 116:351–377.

Gaechter, S., Starmer, C., and Tufano, F. (2012). The power of social relations for coordination: The magic of 'oneness'. *Working Paper*.

Goeree, J. K., McConnell, M. A., Mitchell, T., Tromp, T., and Yariv, L. (2010). The 1/d law of giving. *American Economic Journal: Microeconomics*, 2:183–203.

Goette, L., Huffman, D., and Meier, S. (2012). The impact of social ties on group interactions: Evidence from minimal groups and randomly assigned real groups. *American Economic Journal: Microeconomics*, 4:101–115.

Hogg, M. A. (2002). Social identity. In Leary, M. and Tangney, J., editors, *Handbook of self and identity*, pages 462–479. The Guilford Press.

Kimbrough, E. O. and Vostroknutov, A. (2015). Norms make preferences social. *Journal of the European Economic Association, forthcoming*.

Krupka, E. L. and Weber, R. A. (2013). Identifying social norms using coordination games: Why does dictator game sharing vary? *Journal of the European Economic Association*, 11:495–524.

Leider, S., Möbius, M. M., Rosenblat, T., and Do, Q.-A. (2009). Directed altruism and enforced reciprocity in social networks. *The Quarterly Journal of Economics*, 124:1815–1851.

Newman, M. E. (2003). The structure and function of complex networks. *SIAM Review*, 45:167–256.

Rawls, J. (1971). *A theory of justice*. Harvard University Press, Cambridge.

Shahriar, Q. (2009). Forward induction works! an experimental study to test the robustness and the power. *Working Paper*.

Sugden, R. (2000). Team preferences. *Economics and Philosophy*, 16:175–204.

Sugden, R. (2003). The logic of team reasoning. *Philosophical Explorations*, 6:165–181.

Tajfel, H., Billig, M., Bundy, R., and Flament, C. (1971). Social categorization and intergroup behaviour. *European Journal of Social Psychology*, 1:149–178.

Tajfel, H. and Turner, J. (1979). An integrative theory of intergroup conflict. In Austin, W. G. and Worchel, S., editors, *The Social Psychology of Intergroup Relations*, pages 33–47, Monterey, CA: Brooks/Cole.

Turner, J., Hogg, M. A., Oakes, P. J., Reicher, S. D., and Wetherell, M. S. (1987). *Rediscovering the Social Group: A Self-Categorization Theory*. Oxford: Blackwell.

Van Huyck, J. B., Battalio, R. C., and Beil, R. O. (1993). Asset markets as an equilibrium selection mechanism: Coordination failure, game form auctions, and tacit communication. *Games and Economic Behavior*, 5:485–504.

Watts, D. J. and Strogatz, S. H. (1998). *Nature*, 393:440–442.

A Measurement of subjective connectedness

Figures 7 and 8 illustrate the kind of questions that were used in our experiment. Note that the individuals' faces have been voluntarily blurred to ensure anonymity.

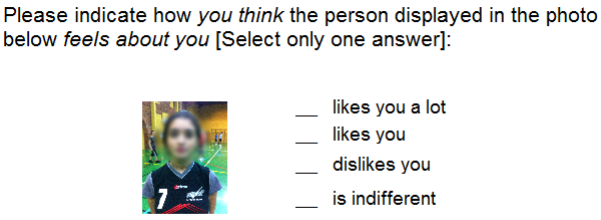


Figure 7: Measuring an individual's self connectedness with another team member.

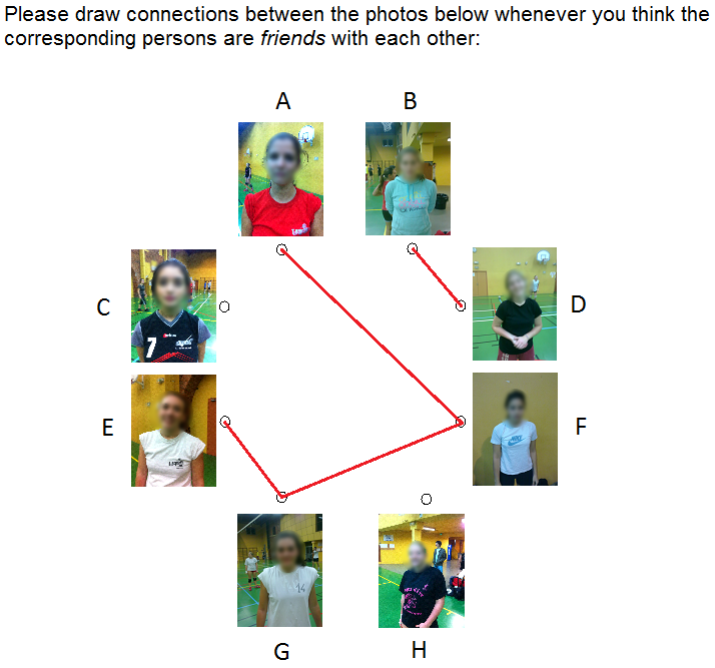


Figure 8: Measuring individual C's belief of the group connectedness.

B Experimental Instructions

B.1 Preliminary Instructions

We are going to present two games that you will have to play with some unknown participants. One of these games will then be drawn in order to determine your actual earnings.

Each game considers two players. You will be asked to take a decision as *player 1* **and** as *player 2*. At the end of the experiment, we will randomly assign one of these two roles to you.

Your actual earning will then depend on your decision in the role that will be assigned to you as well as your partner's decision in the selected game. Therefore, each of your decisions is important. So please take every question seriously by carefully answering them.

Moreover your participation to this experiment relies on the fact that you answered every single question.

If anything is unclear or if you have any question, please do not hesitate to raise your hand so that we can bring you the clarification that you need.

B.2 Instructions of the baseline game

During this experiment, you will interact with some randomly selected player in a game that is defined as follows.

In the first stage, some initial amount are given to both you and your opponent:

- **20 euros** for *player 1*
- **10 euros** for *player 2*

No decision needs to be taken by any player during this stage.

In the second stage, every player will then have to choose simultaneously between two distinct moves **A** and **B**.

In the second stage:

- If every player chooses to play **A**, 5 euros will be withdrawn from *player 2*'s initial amount and 15 euros will be added to *player 1*'s initial amount. Thus *player 1* will get 35 euros while *player 2* will get 5 euros.

- If every player chooses to play **B**, 5 euros will be withdrawn from *player 1*'s initial amount and 25 euros will be added to *player 2*'s initial amount. Thus *player 1* will get 15 euros while *player 2* will get 35 euros.

- If the players' choices are different from each other, then both players' amount will be reset to zero (each will thus get 0 euro).

The following table summarizes the various choices and payoffs from the second stage:

	A	**B**
A	(1): 35 (2): 5	(1): 0 (2): 0
B	(1): 0 (2): 0	(1): 15 (2): 35

This simultaneous decision ends both the second stage and the game. All along the game, both players will remain anonymous to one another. You will receive the corresponding amount if this game is eventually being selected.

These instructions concern the three situations described below.

B.3 Questions for the baseline game

In the context of the previous game, you will play with **X** [19] *(select one answer per question).*

- Please indicate your choice if you are acting as **player 1**:

  In the second stage, you play:

[19] Depending on the matching process, **X** may stand for "a university student", "a club member", or "a teammate". Each subject answered the following two questions (as player 1 and as player 2) for all three values of **X** (See Section 4 for details about the matching process).

○ A ○ B

- Please indicate your choice if you are acting as **player 2**:

  In the second stage, you play:

○ A ○ B

Note that the three previous pair of questions (with, as opponent: a university student, a club member, or a teammate) are independent from one another. Please make sure to answer each of them.

B.4 Instructions of the entrance game

During this experiment, you will interact with some randomly selected player in a game that is defined as follows.

In the first stage, some initial amount are given to both you and your opponent:

- **20 euros** for *player 1*
- **10 euros** for *player 2*

Then, the two following options become available to *player 1*:

- The **"Out"** option implies that every player keeps their initial amount and the game ends.
- The alternative option (**"In"**) implies entering a second stage where each player will have to take another decision. In the latter case, both players will then have to choose simultaneously between two distinct moves **A** and **B**.

In the second stage:

- If every player chooses to play **A**, 5 euros will be withdrawn from *player 2*'s initial amount and 15 euros will be added to *player 1*'s initial amount. Thus player 1 will get 35 euros while *player 2* will get 5 euros.

- If every player chooses to play **B**, 5 euros will be withdrawn from *player 1*'s initial amount and 25 euros will be added to *player 2*'s initial amount. Thus *player 1* will get 15 euros while *player 2* will get 35 euros.

- If the players' choices are different from each other, then both players' amount will be reset to zero (each will thus get 0 euro).

The following table summarizes the various choices and payoffs from the second stage:

	A	**B**
A	(1): 35 (2): 5	(1): 0 (2): 0
B	(1): 0 (2): 0	(1): 15 (2): 35

This simultaneous decision ends both the second stage and the game. All along the game, both players will remain anonymous to one another. You will receive the corresponding amount if this game is eventually being selected.

These instructions concern the three situations described below.

B.5 Questions for the entrance game

In the context of the previous game, you will play with **X** [20] *(select one answer per question).*

- Please indicate your choice while you are acting as **player 1**:

  In the first stage, you play:

[20] Depending on the matching process, **X** may stand for "a university student", "a club member", or "a teammate". Each subject answered the following two questions (as player 1 and as player 2) for all three values of **X** (See Section 4 for details about the matching process).

○ In ○ Out

In the second stage (assume that you played **"In"** first), you play:

○ A ○ B

- Please indicate your choice while you are acting as **player 2**:

  In the second stage (assume that your opponent played **"In"** first), you play:

  ○ A ○ B

Note that the three previous pair of questions (with, as opponent: a university student, a club member, or a teammate) are independent from one another. Please make sure to answer each of them.

C Additional tables

Throughout this section, all figures include Wilcoxon signed rank test results concerning mean rank differences in behavior across treatments (i.e., team, club, university). Moreover, some tables also include Mann-Whitney test results that allow comparing behavior across the various groups from Table 1. Note that in all statistical tests, only p values lower than 0.2 are displayed in the tables. p values larger than 0.2 are classified as not significant (n.s.).

C.1 Baseline game

players	**Matching types**			**Wilcoxon signed rank test** (p values)		
	team	**club**	**university**	**team** vs **university**	**team** vs **club**	**club** vs **university**
1	68%	58%	49%	0.002	0.090	0.109
2	77%	77%	74%	n.s.	n.s.	n.s.

Table 4: Choosing B in the baseline game for each player in each type of matching (69 observations).

Groups	**Player**	**Matching types**			**Wilcoxon signed rank test** (p values)		
		team	**club**	**university**	**team** vs **university**	**team** vs **club**	**club** vs **university**
H^S	1	76%	68%	58%	0.008	n.s.	0.102
(38 obs.)	2	79%	74%	74%	n.s.	n.s.	n.s.
L^S	1	58%	45%	39%	0.058	n.s.	n.s.
(31 obs.)	2	74%	80%	74%	n.s.	n.s.	n.s.
H^S **vs** L^S	1	0.109	0.053	0.115			
(p values)	2	n.s.	n.s.	n.s.			

Table 5: Choosing B in the baseline game based on groups H^S ($k_i^S \geq K^S$) and L^S ($k_i^S < K^S$).

Groups	Player	Matching types			Wilcoxon signed rank test (p values)		
		team	**club**	**university**	**team** vs **university**	**team** vs **club**	**club** vs **university**
H^G	1	83%	73%	63%	0.014	0.179	0.179
(30 obs.)	2	83%	83%	73%	0.179	n.s.	0.083
L^G	1	56%	46%	38%	0.035	n.s.	n.s.
(39 obs.)	2	72%	72%	74%	n.s.	n.s.	n.s.
H^G **vs** L^G	1	0.018	0.024	0.042			
(p values)	2	n.s.	n.s.	n.s.			

Table 6: Choosing B in the baseline game based on groups H^G ($k_i^G \geq K^G$) and L^G ($k_i^G < K^G$).

C.2 Entrance game

Tables 7, 9 and 11 depict player 1's choice in the entrance game. Note that these tables ignore counterfactual strategies (i.e., strategies (Out,A) and (Out,B)). Moreover, these tables include Wilcoxon signed rank tests that compare how often did subjects perform a given strategy (e.g., (In,A)) with how often did they choose any other strategy (e.g., (In,B) or Out) in any two treatments. Tables 8, 10 and 12 depict player 2's choice in the entrance game, together with Wilcoxon signed rank tests.

Choices	Matching types			Wilcoxon signed rank test (p values)		
	team	**club**	**university**	**team** vs **university**	**team** vs **club**	**club** vs **university**
(In,A)	23%	23%	21%	n.s.	n.s.	n.s.
(In,B)	39%	30%	20%	0.003	0.157	0.035
Out	38%	47%	59%	0.004	0.083	0.059

Table 7: Player 1's behavior in the entrance game (69 observations).

Matching types			Wilcoxon signed rank test (p values)		
team	**club**	**university**	**team** vs **university**	**team** vs **club**	**club** vs **university**
77%	67%	64%	0.049	0.108	n.s.

Table 8: Choosing B for player 2 in the entrance game (69 observations).

Groups	**Choices**	**Matching types**			**Wilcoxon signed rank test** (p values)		
		team	**club**	**university**	**team** vs **university**	**team** vs **club**	**club** vs **university**
H^S (38 obs.)	(In,A)	18%	8%	13%	n.s.	0.102	n.s.
	(In,B)	50%	39%	24%	0.007	n.s.	0.034
	Out	32%	53%	63%	0.001	0.011	0.157
L^S (31 obs.)	(In,A)	29%	42%	32%	n.s.	n.s.	n.s.
	(In,B)	26%	19%	16%	0.180	n.s.	n.s.
	Out	45%	39%	52%	n.s.	0.157	n.s.
H^S **vs** L^S (p values)	(In,A)	n.s.	0.001	n.s.			
	(In,B)	0.042	0.073	n.s.			
	Out	0.193	n.s.	n.s.			

Table 9: Player 1's behavior in the entrance game based on groups H^S ($k_i^S \geq K^S$) and L^S ($k_i^S < K^S$).

Groups	**Matching types**			**Wilcoxon signed rank test** (p values)		
	team	**club**	**university**	**team** vs **university**	**team** vs **club**	**club** vs **university**
H^S (38 obs.)	76%	71%	68%	n.s.	n.s.	n.s.
L^S (31 obs.)	77%	61%	58%	0.083	0.095	n.s.
H^S **vs** L^S (p values)	n.s.	n.s.	n.s.			

Table 10: Choosing B for player 2 in the entrance game based on groups H^S ($k_i^S \geq K^S$) and L^S ($k_i^S < K^S$).

Groups	**Choices**	**Matching types**			**Wilcoxon signed rank test** (p values)		
		team	**club**	**university**	**team** vs **university**	**team** vs **club**	**club** vs **university**
H^G (38 obs.)	(In,A)	13%	23%	13%	n.s.	n.s.	n.s.
	(In,B)	57%	30%	20%	0.002	0.021	0.180
	Out	30%	47%	67%	0.002	0.058	0.058
L^G (31 obs.)	(In,A)	31%	23%	28%	n.s.	n.s.	n.s.
	(In,B)	26%	31%	20%	n.s.	n.s.	0.102
	Out	43%	46%	52%	n.s.	n.s.	n.s.
H^G **vs** L^G (p values)	(In,A)	0.091	n.s.	0.140			
	(In,B)	0.009	n.s.	n.s.			
	Out	n.s.	n.s.	n.s.			

Table 11: Player 1's behavior in the entrance game based on groups H^G ($k_i^G \geq K^G$) and L^G ($k_i^G < K^G$).

Groups	**Matching types**			**Wilcoxon signed rank test** (p values)		
	team	**club**	**university**	**team** vs **university**	**team** vs **club**	**club** vs **university**
H^G (30 obs.)	90%	70%	63%	0.011	0.057	n.s.
L^G (39 obs.)	67%	64%	64%	n.s.	n.s.	n.s.
H^G **vs** L^G (p values)	0.024	n.s.	n.s.			

Table 12: Choosing B for player 2 in the entrance subgame based on groups H^G ($k_i^G \geq K^G$) and L^G ($k_i^G < K^G$).

D Behind the veil of ignorance

Tables 13 and 14 represent respectively the payoff matrices of the baseline game and the entrance game when played behind the veil of ignorance.

Player X	**Player Y**			
	(A,A)	(A,B)	(B,A)	(B,B)
(A,A)	20	2.5	17.5	0
(A,B)	17.5	0	35	17.5
(B,A)	2.5	10	0	7.5
(B,B)	0	7.5	17.5	25

Table 13: Average payoffs for row Player X in the transformed baseline game.

Player X	**Player Y**					
	((In,A),A)	((In,A),B)	((In,B),A)	((In,B),B)	(Out,A)	(Out,B)
((In,A),A)	20	2.55	17.5	0	22.5	5
((In,A),B)	17.5	0	35	17.5	22.5	5
((In,B),A)	2.5	10	0	7.5	5	12.5
((In,B),B)	0	7.5	17.5	25	5	12.5
(Out,A)	12.5	12.5	10	10	15	15
(Out,B)	10	10	27.5	27.5	15	15

Table 14: Average payoffs for row Player X in the transformed entrance game.

Tables 15 and 16 depict tests of independence of behavioral variables in the baseline game (playing A/B as player 1 vs playing A/B as player 2) and the entrance game (playing (In,A)/(In,B)/Out as player 1 vs playing A/B as player 2) respectively.

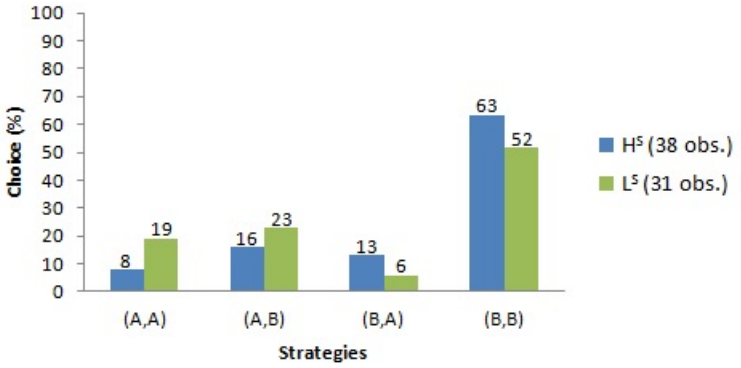
Decision as player 1 (A/B) vs decision as player 2 (A/B) (Pearson's chi-squared test)	**Matching types**		
	team	**club**	**university**
χ^2	5.694	1.729	0.384
p value	0.017	0.189	0.535

Table 15: Independence of decisions as both players in the baseline game (69 observations).

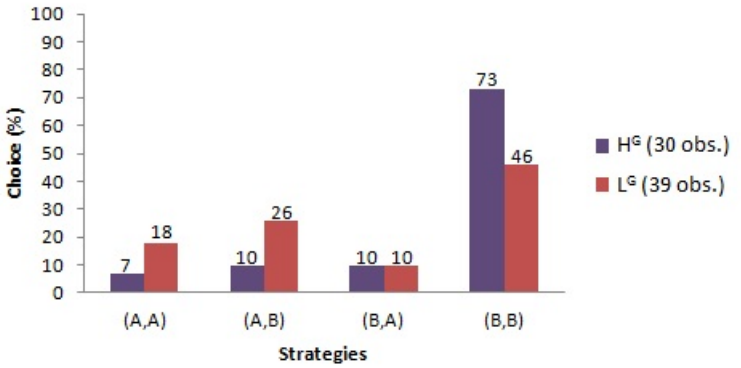
Decision as player 1 ((In,A)/(In,B)/Out) vs decision as player 2 (A/B) (Pearson's chi-squared test)	**Matching types**		
	team	**club**	**university**
χ^2	8.897	0.495	0.950
p value	0.012	0.781	0.622

Table 16: Independence of decisions as both players in the entrance game (69 observations).

Figures 9 and 10 illustrate the observed behavior in the baseline game and the entrance game respectively, under the assumption that those games are played behind the veil of ignorance.

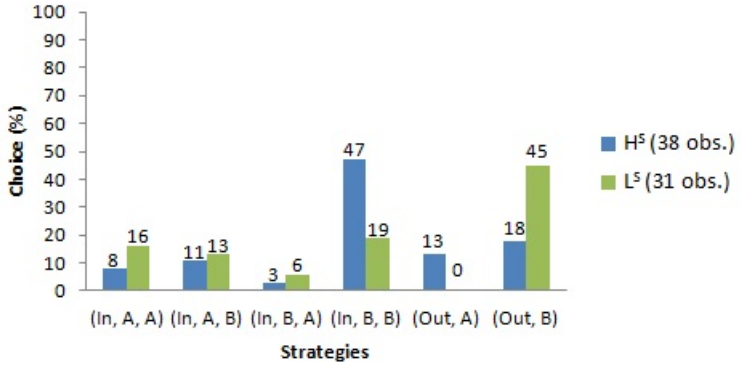


(a) Classification based on self connectedness

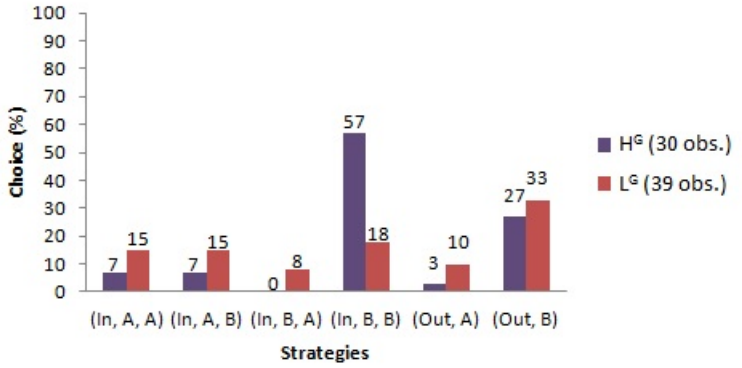


(b) Classification based on group connectedness

Figure 9: Behavior in the original position of the baseline game (Team treatment).



(a) Classification based on self connectedness



(b) Classification based on group connectedness

Figure 10: Behavior in the original position of the entrance game (Team treatment).