Influencing Choices by Changing Beliefs: A Logical Theory of Influence, Persuasion, and Deception

Published:
DOI: http://doi.org/10.1007/978-3-030-91779-1_9

Influencing Choices by Changing Beliefs: A Logical Theory of Influence, Persuasion, and Deception

Grégory Bonnet[1], Christopher Leturc[2], Emiliano Lorini[3], and Giovanni Sartor[4]

[1] Normandie University, UNICAEN, ENSICAEN, CNRS, GREYC, France
[2] Institut Henri Fayol, Mines Saint-Étienne, France
[3] CNRS-IRIT, Toulouse University, France
[4] EUI-Florence, University of Bologna, Italy

Abstract. We model persuasion, viewed as a deliberate action through which an agent (persuader) changes the beliefs of another agent's (persuadee). This notion of persuasion paves the way to express the idea of persuasive influence, namely inducing a change in the choices of the persuadee by changing her beliefs. It allows in turns to express different aspects of deception. To this end, we propose a logical framework that enables expressing actions and capabilities of agents, their mental states (desires, knowledge and beliefs), a variety of agency operators as well as the connection between mental states and choices. Those notions, once combined, enable us to capture, the notion of influence, persuasion and deception, as well as their relation.

1 Introduction

In many contexts, agents need the cooperation of others, to achieve their goals. To this end, they may influence others, namely execute actions leading others to perform further actions. Influence may result from impeding or enabling certain actions by others (removing or adding choices), but also from changing the mental states of others (removing or adding beliefs). In the first case we speak of *regimentation*, while in the second we speak of *persuasion*. In this paper, we focus on persuasion, *i.e.* on the deliberate action through which agents (persuaders) changes the beliefs of other agents (persuadees).

Persuasion in the human-machine interaction raises serious ethical and legal issues, as more and more often automated systems engage in attempts at influencing human choices through persuasive messages. Such persuasive activities may be determined by interests that are not aligned with the interests of their addressees, but are rather determined by the economic or political goals of the senders (or their principals). Automated persuaders may also be *deceivers*, *e.g.* agents which present fake information or anyway induce the persuadees into actions the latter will regret (bad economic, personal or political choices). To adequately respond to this challenge, it is important to have a clear conceptual framework, that enables us to capture the different ways in which influence, persuasion and deception can be deployed, so that adequate responses to each of such ways can be developed, through human or also computational interventions. Let us precise that, as we are mainly interested in artificial agents (automated persuaders), we will consider in this article rational agents. The anthropomorphic expression (persuader, persuadee, mental state) are just shortcuts that we use for convenience.

In the AI field, persuasion and related notions have been approached from different perspectives. Following seminal work on argumentation [28], persuasion has been modelled through structured argumentation [19], abstract argumentation [6,7], probabilistic argumentation [13], possibilistic belief revision [9], abstract argumentation combined with dynamic epistemic logic [20]. Some logical approaches have addressed notions related to persuasion, such as social influence [16,24], manipulation (influence on choices) [15], lying and deception [23,27], or changing other agents' degrees of beliefs [8]. The original contribution of this work consists in providing a logical account of the way in which a persuader by changing the beliefs of the persuadee, influences the action of the latter. We aim indeed to provide a formal theory of the micro-foundations of deception, persuasion and influence, i.e., an account of the cognitive attitudes and agentive aspects that are involved in the persuasion and influence, and elucidate their relationship. For this purpose we provide a rich framework that expresses actions and capabilities of agents and their mental states (desires, knowledge and beliefs) as well as the connection between mental states and choices. In order to keep the framework simple while being expressive, we focus on a qualitative framework (such as Boolean games [26]) where agents' preferences are not presented by continuous utility functions but rather by a qualitative three-valued scale desirable/undesirable/neutral. Then We express two notions of rationality (an optimistic and a pessimistic one), several agency operators (such as the so-called 'Chellas' STIT [12], the deliberative and the rational STIT operators [16]) and different ways to influence agents' choices through belief change.

Relative to this rich background addressing partial aspects of persuasion processes, our model will be useful for better understanding and modelling the dynamics of social influence, especially those between artificial and human agents. It will also be relevant from a regulatory perspective, since it allows to pinpoint those instance in which online interactions, in virtue of their logical structure, can be viewed as detrimental to trustful and productive interaction, and thus call for normative limitations. It can also be relevant in the special cases where persuasion and deception can be of interest in social relationship to protect somebody [2]. Our contribution has the advantage of providing a comprehensive model which captures the whole persuasive process including the mental states of the persuader, his persuading action, the modified mental states of persuadee and her resulting action. The model also captures the connection between influence, regimentation and persuasion, and enables to link persuasion with game theory.

This article is organized as follow. Firstly, in Section 1 we introduce a running example. Then we define in Section 2 our logical framework and, in Section 3, we show how this framework can express notions of optimistic and pessimistic rationality, and a variety of agency operators. We then combine those notions in Section 4 to express deception, persuasion, regimentation and influence, and their relationships. Finally, we apply our framework to the running example.

Running example John is feeling back pain, and consequently he is consulting websites that offer medical advice as well as the opportunity to purchase drugs. A message from a bot promises, for a fair price, a drug which is said to be an excellent remedy that would eliminate all pain, and can be used for any length of time without producing any dependency. This message persuades John to buy the drug and he starts taking it. However the bot has deceived John because the bot knows that the drug creates addiction, with serious health consequences. John's wife Ann, comes to know that such pills are dangerous. She then removes all pills from the closet. As a consequence John does not take the drug.

This example shows two patterns of influence. The first is successful and misleading persuasion (deception) by the website, *i.e.*, successful influencing by providing false information. The second stage consists in Ann successfully influencing John through regimentation, i.e., by removing a choice option.

2 Logical Framework

In this section, we present a modal logic language which supports reasoning about (i) actions and capabilities of agents and coalitions, (ii) agents' epistemic states and desires as well as their connection with agents' choices. We first present its syntax and its semantic interpretation (Sections 2.1 and 2.2). Then, in Section 2.3, we provide a sound and complete axiomatization of its set of validities.

2.1 Syntax

Assume a countable set of atomic propositions $Atm = \{p, q, \ldots\}$, a finite set of agents $Agt = \{1, \ldots, n\}$, a finite set of atomic action names $Act = \{a, b, \ldots\}$. The set Act includes the (in)action skip, *i.e.*, the action of doing nothing. We define $Prop$ to be the set of propositional formulas, that is, the set of all Boolean combinations of atomic propositions.

The set of non-empty sets of agents, also called *coalitions*, is defined by $2^{Agt*} = 2^{Agt} \setminus \{\emptyset\}$. Elements of 2^{Agt*} are noted $H, J, \ldots$ A coalition H's joint action is defined to be a function $\delta_H : H \longrightarrow Act$. Coalition H's set of joint actions is noted $JAct_H$. Its elements are noted $\delta_H, \delta'_H, \ldots$ For notational convenience, we simply write $JAct$ instead of $JAct_{Agt}$ to denote the grand coalition's set of joint actions. Its elements are noted $\delta, \delta', \ldots$ Moreover, we write Act_i instead of $JAct_{\{i\}}$ to denote agent i's set of individual actions. Its elements are noted $a_i, b_i, \ldots$ We define $JAct^*$ to be the set of all finite sequences of joint actions from $JAct$. Elements of $JAct^*$ are noted $\epsilon, \epsilon', \ldots$ The empty sequence of joint actions is denoted by nil. Infinite sequences of joint actions are called *histories*. The set of all histories is noted $Hist$ and its elements are noted $h, h', \ldots$ Elements of $JAct^* \cup Hist$ are noted $\tau, \tau', \ldots$ For every $\tau_1, \tau_2 \in JAct^* \cup Hist$, we write $\tau_1 \sqsubseteq \tau_2$ to mean that either $\tau_1 = \tau_2$ or τ_1 is an initial subsequence of τ_2, *i.e.*, there is $\tau_3 \in JAct^* \cup Hist$ such that $\tau_2 = \tau_1;\tau_3$. The language $\mathcal{L}$ is defined by the following grammar:

$$
\begin{aligned}
\epsilon &::= \delta \mid \epsilon;\epsilon \\
\varphi &::= p \mid \mathsf{occ}(\epsilon) \mid \mathsf{plaus}_i \mid \mathsf{good}_i \mid \mathsf{bad}_i \mid \mathsf{neutral}_i \mid \\
&\quad\; \neg\varphi \mid \varphi_1 \wedge \varphi_2 \mid [\![\delta]\!]\varphi \mid \Box\varphi \mid \mathsf{K}_i\varphi
\end{aligned}
$$

where p ranges over Atm, i ranges over Agt, δ ranges over $JAct$ and ϵ ranges over $JAct^*$. The other Boolean constructions $\top$, $\bot$, $\vee$, $\rightarrow$ and $\leftrightarrow$ are defined from p, $\neg$ and $\wedge$ in the standard way. We note $\mathcal{L}^-$ the fragment of $\mathcal{L}$ without operators $[\![\delta]\!]$ defined by the following grammar:

$$\varphi ::= p \mid \mathsf{occ}(\epsilon) \mid \mathsf{plaus}_i \mid \mathsf{good}_i \mid \mathsf{neutral}_i \mid \mathsf{bad}_i \mid \neg\varphi \mid \varphi_1 \wedge \varphi_2 \mid \Box\varphi \mid \mathsf{K}_i\varphi$$

$\mathcal{L}$ has special atomic formulas of four different kinds. The atomic formulas $\mathsf{occ}(\epsilon)$ represent information about occurrences of joint action sequences. The formula $\mathsf{occ}(\epsilon)$ has to be read "the joint action sequence ϵ is going to occur".

The following abbreviations capture interesting action-related concepts. For every $H \in 2^{Agt*}$, joint action sequence $\epsilon \in JAct^*$ and joint action $\delta_H \in JAct_H$ we define:

$$\mathsf{choose}(\epsilon,\delta_H) \stackrel{\mathsf{def}}{=} \bigvee_{\substack{\delta \in JAct:\\ \forall i \in H,\, \delta_H(i)=\delta(i)}} \mathsf{occ}(\epsilon;\delta)$$

$$\mathsf{can}(\epsilon,\delta_H) \stackrel{\mathsf{def}}{=} \Diamond \mathsf{choose}(\epsilon,\delta_H)$$

$\mathsf{choose}(\epsilon,\delta_H)$ has to be read "the joint action sequence ϵ is going to occur and will be followed by coalition H's joint action δ_H". For convenience, when $\epsilon = nil$, we write $\mathsf{choose}(\delta_H)$ instead of $\mathsf{choose}(nil,\delta_H)$. Formula $\mathsf{can}(\epsilon,\delta_H)$ has to be read "coalition H can choose the joint action δ_H at the end of the joint action sequence ϵ".

The atomic formula plaus_i is used to identify the histories in agent i's information set that she considers plausible. It has to be read "the current history is considered plausible by agent i". The other atomic formulas good_i, bad_i and $\mathsf{neutral}_i$ are used to rank the histories that an agent envisages at a given world according to their value for the agent (*i.e.*, how much a given history promotes the satisfaction of the agent's desires). They are read, respectively, "the current history is good/bad/neutral for agent i". Let us notice it is an agent-centric point-of-view. Each good_i, bad_i and $\mathsf{neutral}_i$ atoms are defined only from agent i's perspective.

$\mathcal{L}$ has three kinds of modal operators: $[\![\delta]\!]$, $\Box$ and K_i. $\Box$ is the so-called historical necessity operator. The formula $\Box\varphi$ has to be read "φ is true in all histories passing through the current moment". We define $\Diamond$ to be the dual of $\Box$, *i.e.*, $\Diamond\varphi \stackrel{\mathsf{def}}{=} \neg\Box\neg\varphi$ where $\Diamond\varphi$ has to be read "φ is true in at least one history passing through the current moment". $[\![\delta]\!]$ is a dynamic operator describing the fact that if the joint action δ is performed then it will lead to a state in which a given state of affairs holds. In particular, $[\![\delta]\!]\varphi$ has to be read "if the joint action δ is performed, then φ will be true after its execution". Finally, K_i is a modal operator characterizing the concept of *ex ante* (or *choice-independent*) knowledge [4,16,22]. The formula $\mathsf{K}_i\varphi$ has to be read "agent i knows that φ is true independently from her current choice" or "agent i thinks that φ is true for any choice she could have made". The dual of the operator K_i is denoted by $\widehat{\mathsf{K}}_i$, *i.e.*, $\widehat{\mathsf{K}}_i\varphi \stackrel{\mathsf{def}}{=} \neg\mathsf{K}_i\neg\varphi$. *Ex ante* knowledge is distinguished from *ex post* knowledge. *Ex ante* knowledge characterizes an agent's knowledge assuming that no decision has yet been made by him, whereas *ex post* knowledge characterizes an agent's knowledge assuming that the agent has made his decision about which action to take, but might still be uncertain about the decisions of others. The concept of *ex post* knowledge is expressed by the following operator K_i^{post}:

$$\mathsf{K}_i^{post}\varphi \stackrel{\mathtt{def}}{=} \bigwedge_{a_i \in Act_i} \big(\mathsf{choose}(a_i) \to \mathsf{K}_i(\mathsf{choose}(a_i) \to \varphi)\big)$$

where $\mathsf{K}_i^{post}\varphi$ has to be read "agent i knows that φ is true, given her actual choice". From the special atomic formula plaus_i and the epistemic operator K_i, we define a belief operator:

$$\mathsf{B}_i\varphi \stackrel{\mathtt{def}}{=} \mathsf{K}_i(\mathsf{plaus}_i \to \varphi)$$

The formula $\mathsf{B}_i\varphi$ has to be read "agent i believes that φ". According to this definition, an agent believes that φ if and only if φ is true at all states in the agent's information set at which φ is true. The dual of the operator B_i is denoted by $\widehat{\mathsf{B}}_i$, *i.e.*, $\widehat{\mathsf{B}}_i\varphi \stackrel{\mathtt{def}}{=} \neg\mathsf{B}_i\neg\varphi$.

Similarly to Situation Calculus [21], we describe actions in terms of their positive and negative effect preconditions. In particular, we introduce two functions γ^+ and γ^- with domain $Agt \times Act \times Atm$ and codomain $\mathcal{L}^-$. The formula $\gamma^+(i, a, p)$ describes the *positive effect preconditions* of action a performed by agent i with respect to p, whereas $\gamma^-(i, a, p)$ describes the *negative effect preconditions* of action a performed by agent i with respect to p. Formula $\gamma^+(i, a, p)$ represents the conditions under which agent i will make p *true* by performing action a, if no other agent interferes with i's action; while $\gamma^-(i, a, p)$ represents the conditions under which i will make p *false* by performing a, if no other agent interferes with i's action. We assume that "making p true" means changing the truth value of p from false to true, whereas "making p false" means changing the truth value of p from true to false. The reason why an action's effect preconditions range over $\mathcal{L}^-$ and not over $\mathcal{L}$ is that they should be independent from the effects of the action described by dynamic formulas of type $[\![\delta]\!]\varphi$.

Example 1. Let us consider the story given in Section 1, and use it to illustrate the γ^+ and γ^- functions. We shall use the following vocabulary:

$$\begin{aligned} Agt =& \{Ann, John, Bot\}, \\ Act =& \{take, suggest, hide, skip\}, \\ Atm =& \{has_{John,drug}, ingested_{John,drug}, addicted_{John}, pain_{John}\} \end{aligned}$$

The actions' effect preconditions can be defined as:

$$\begin{aligned} &\gamma^+(Bot, suggest, p) = p \text{ for all } p \in Atm, \\ &\gamma^-(Bot, suggest, p) = \neg p \text{ for all } p \in Atm, \\ &\gamma^+(John, take, ingested_{John,drug}) = has_{John,drug}, \\ &\gamma^+(i, a, ingested_{John,drug}) = \bot \end{aligned}$$

Moreover, if $a \neq take$ or $i \neq John$, we have:

$$\gamma^-(i, a, ingested_{John,drug}) = \neg\mathsf{choose}(take_{John})$$

Finally, for all $a \in Act$ and $i \in Agt$, we have:

$$\gamma^{+}(i, a, has_{John,drug}) = has_{John,drug} \wedge \neg \mathsf{choose}(hide_{Ann})$$
$$\gamma^{-}(Ann, hide, has_{John,drug}) = \top$$

The effect preconditions specify that the speech act of suggestion has no effect on material facts. Ingesting the drug presupposes possession of it, while not ingesting it presupposes not having taken it. Finally, John will still have the drug unless Ann hides it, while John will not have the drug if Ann hides it.

2.2 Semantics

The semantics for the language $\mathcal{L}$ is a possible world semantics with accessibility relations associated with each modal operator, with a function designating the history starting in a given world, a plausibility function and a trichotomous utility function relative to histories.

Definition 1. *A model is a tuple $M = \big(W, \mathcal{H}, \equiv, (\mathcal{E}_i)_{i \in Agt}, \mathcal{P}, \mathcal{U}, \mathcal{V}\big)$ where: (i) W is a non-empty set of worlds, (ii) $\mathcal{H} : W \longrightarrow Hist$ is a history function, (iii) $\equiv$ and every $\mathcal{E}_i$ are equivalence relations on W, (iv) $\mathcal{P} : W \times Agt \longrightarrow \{0, 1\}$ is a plausibility function, (v) $\mathcal{U} : W \times Agt \longrightarrow \{0, 1, -1\}$ is a utility function, and (vi) $\mathcal{V} : W \longrightarrow 2^{Atm}$ is a valuation function.*

For each binary relation $\mathcal{R} \in \{\equiv, \mathcal{E}_1, \ldots, \mathcal{E}_n\}$, we set $\mathcal{R}(w) = \{v \in W : w\mathcal{R}v\}$. As usual $p \in \mathcal{V}(w)$ means that proposition p is true at world w. $\equiv$-equivalence classes are called *moments*. If w and v belong to the same moment (*i.e.*, $w \equiv v$), then the history starting in w (*i.e.*, $\mathcal{H}(w)$) and the history starting in v (*i.e.*, $\mathcal{H}(v)$) are said to be alternative histories (*viz.*, histories starting at the same moment). The concept of moment is the one used in STIT logic [5,12] and, more generally, in the Ockhamist theory of time [25,29]. For every world $w \in W$, $\mathcal{H}(w)$ identifies the history starting in w. For notational convenience, for all $\epsilon \in JAct^{*}$, $i \in Agt$, $a \in Act$ and $w \in W$, we write $\epsilon;a \sqsubseteq_i \mathcal{H}(w)$ to mean that there is $\delta \in JAct$ such that $\delta(i) = a$ and $\epsilon;\delta \sqsubseteq \mathcal{H}(w)$.

We define the actual choice function $\mathcal{C}_{act} : W \times Agt \longrightarrow Act$: for every $w \in W$, $i \in Agt$ and $a \in Act$, we have $\mathcal{C}_{act}(w, i) = a$ iff there exists $\delta \in JAct$ such that $\delta(i) = a$ and $\delta \sqsubseteq \mathcal{H}(w)$. Furthermore, we define the available choice function $\mathcal{C}_{avail} : W \times Agt \longrightarrow 2^{Act}$: for every $w \in W$, $i \in Agt$ and $a \in Act$, we have $a \in \mathcal{C}_{avail}(w, i)$ iff there exists $\delta \in JAct$ and $v \in \equiv(w)$ such that $\delta(i) = a$ and $\delta \sqsubseteq \mathcal{H}(v)$.

The equivalence relations $\mathcal{E}_i$ are used to interpret the epistemic operators K_i. The set $\mathcal{E}_i(w)$ is the agent i's *information set* at world w: the set of worlds that agent i envisages at w or, shortly, agent i's set of epistemic alternatives at w. As $\mathcal{E}_i$ is an equivalence relation, if $w\mathcal{E}_i v$ then agent i has the same information set at w and v. The function $\mathcal{P}$ specifies the possibility value of a history for an agent. In particular, $\mathcal{P}(w, i) = 1$ (resp. $\mathcal{P}(w, i) = 0$) means that the history starting in w is considered plausible (resp. not plausible) by agent i. We define agent i's belief set at world w, denoted by $\mathcal{B}_i(w)$, as the set of worlds in i's information set at w that i considers plausible: $\mathcal{B}_i(w) = \mathcal{E}_i(w) \cap \{v \in W : \mathcal{P}(v, i) = 1\}$.

Consequently, the complementary set $\mathcal{E}_i(w) \setminus \mathcal{B}_i(w)$ is the set of worlds that agent i envisages at w but that she does not consider plausible. Since $\mathcal{E}_i$ is an equivalence relation, the following properties for belief hold. Note that $\mathcal{B}_i \subseteq \mathcal{E}_i$ and if $w\mathcal{E}_i v$ then $\mathcal{B}_i(w) = \mathcal{B}_i(v)$, where $\mathcal{B}_i = \{(w,v) \in W \times W : v \in \mathcal{B}_i(w)\}$. Moreover, $\mathcal{B}_i$ is transitive and Euclidean. These properties correspond to the combination of belief and knowledge studied in [14].

The function $\mathcal{U}$ assigns the utility value $\mathcal{U}(w,i)$ of the history starting in w for agent i. In particular, $\mathcal{U}(w,i) = 1$, $\mathcal{U}(w,i) = -1$ and $\mathcal{U}(w,i) = 0$, mean respectively that the history starting in w is good/bad/neutral for agent i. A history is good for an agent if the agent obtains what she likes and avoids what she dislikes along it. It is bad for the agent if the agent does not avoid what she dislikes and does not obtain what she likes along it. Finally, it is neutral for the agent if either the agent does not obtain what she likes and avoids what she dislikes along it, or the agent obtains what she likes and does not avoid what she dislikes along it. Our simplified account of utility presupposes that every agent is identified with a single appetitive desire (*i.e.*, what the agent likes) and a single aversive desire (*i.e.*, what the agent dislikes).

We impose the following three constraints on models. For all $w, v \in W$, $\delta \in JAct$, $\epsilon \in JAct^*$, $i \in Agt$ and $a \in Act$:

(**C1**) if for all $i \in Agt$ there is $u_i \in \equiv(w)$ such that $\epsilon;\delta(i) \sqsubseteq_i \mathcal{H}(w)$, then there is $u \in \equiv(w)$ such that $\epsilon;\delta \sqsubseteq \mathcal{H}(u)$;
(**C2**) if there is $v \in \equiv(w)$ such that $\epsilon;a \sqsubseteq_i \mathcal{H}(v)$ then, for every $u \in \mathcal{E}_i(w)$, there is $z \in \equiv(u)$ such that $\epsilon;a \sqsubseteq_i \mathcal{H}(z)$;
(**C3**) if there is $v \in \equiv(w)$ such that $\epsilon;a \sqsubseteq_i \mathcal{H}(v)$, then there is $u \in \mathcal{E}_i(w)$ such that $\epsilon;a \sqsubseteq_i \mathcal{H}(u)$;
(**C4**) if $w \equiv v$ then $\mathcal{E}_i(w) = \mathcal{E}_i(v)$;
(**C5**) $\mathcal{B}_i(w) \neq \emptyset$.

According to the Constraint **C1**, if every individual action in a joint action δ can be chosen at the end of the joint action sequence ϵ, then the individual actions in δ can be chosen simultaneously at the end of ϵ. The Constraint **C1** is a variant of the assumption of *independence of agents* of STIT logic. More intuitively, this means that agents can never be deprived of choices due to the choices made by other agents. The Constraint **C2** is a basic assumption about agents' knowledge over their abilities: if an agent i can choose action a at the end of the joint action sequence ϵ, then he knows this. In other words, an agent has perfect knowledge about the actions he can choose at the end of a joint sequence. The Constraint **C3** characterizes the basic property of *ex ante* knowledge: if an agent i can choose action a at the end of joint action sequence ϵ, then there is a history that the agent considers possible in which he chooses action a at the end of the joint action sequence ϵ. In other words, for every action that an agent can choose, there is a history that the agent considers possible in which he chooses this action. According to Constraint **C4**, an agent's knowledge is moment-determinate, *i.e.*, it does not depend on the specific history at which it is evaluated. This assumption is justified by the fact that the only thing which can vary at a given moment are the agents' choices, but not the agents' ex ante epistemic states. Finally, Constraint **C5** is a normality requirement for beliefs: there should be at least a world in an agent's information set that the agent considers possible. $\mathcal{L}$-formulas are interpreted relative to a model $M = \big(W, \mathcal{H}, \equiv, (\mathcal{E}_i)_{i\in Agt}, \mathcal{P}, \mathcal{U}, \mathcal{V}\big)$ and a world w in W as follows. (We omit boolean cases as they are standard.)

$$
\begin{aligned}
M, w \models \mathsf{occ}(\epsilon) &\iff \epsilon \sqsubseteq \mathcal{H}(w)\\
M, w \models \mathsf{plaus}_i &\iff \mathcal{P}(w, i) = 1\\
M, w \models \mathsf{good}_i &\iff \mathcal{U}(w, i) = 1\\
M, w \models \mathsf{bad}_i &\iff \mathcal{U}(w, i) = -1\\
M, w \models \mathsf{neutral}_i &\iff \mathcal{U}(w, i) = 0\\
M, w \models [\![\delta]\!]\varphi &\iff \text{ if } M, w \models \mathsf{occ}(\delta) \text{ then } M^\delta, w \models \varphi\\
M, w \models \Box\varphi &\iff \forall v \in W: \text{ if } w \equiv v \text{ then } M, v \models \varphi\\
M, w \models \mathsf{K}_i\varphi &\iff \forall v \in W: \text{ if } w\mathcal{E}_i v \text{ then } M, v \models \varphi
\end{aligned}
$$

where model M^δ is defined according to Definition 2 below. Note that the belief operator B_i we defined in Section 2.1 as an abbreviation has the following interpretation: $M, w \models \mathsf{B}_i\varphi$ if and only if $\forall v \in W:$ if $w\mathcal{B}_i v$ then $M, v \models \varphi$.

Definition 2 (Update via joint action). *Let $M = \big(W, \mathcal{H}, \equiv, (\mathcal{E}_i)_{i\in Agt}, \mathcal{P}, \mathcal{U}, \mathcal{V}\big)$ be a model. The update of M by joint action δ is the tuple:*

$$M^\delta = \big(W^\delta, \mathcal{H}^\delta, \equiv^\delta, (\mathcal{E}_i^\delta)_{i\in Agt}, \mathcal{P}^\delta, \mathcal{U}^\delta, \mathcal{V}^\delta\big)$$

where:

$$
\begin{aligned}
&W^\delta &&= \{w \in W : M, w \models \mathsf{occ}(\delta)\}\\
&\mathcal{H}^\delta(w) &&= h \textit{ if } \mathcal{H}(w) = \delta'; h \textit{ for some } \delta' \in JAct\\
&\equiv^\delta &&= \equiv \cap (W^\delta \times W^\delta)\\
&\mathcal{E}_i^\delta &&= \mathcal{E}_i \cap (W^\delta \times W^\delta)\\
&\mathcal{P}^\delta(w, i) &&= \mathcal{P}(w, i) \textit{ if } \mathcal{B}_i(w) \cap W^\delta \neq \emptyset\\
&\mathcal{P}^\delta(w, i) &&= 1 \ \textit{ otherwise}\\
&\mathcal{U}^\delta(w, i) &&= \mathcal{U}(w, i)
\end{aligned}
$$

$$
\begin{aligned}
\mathcal{V}^\delta(w) = \Big(&\mathcal{V}(w) \setminus \{p : \big(\exists a \in Act, i \in Agt :\\
&\delta(i) = a \textit{ and } M, w \models \gamma^-(i, a, p)\big) \textit{ and}\\
&\big(\nexists b \in Act, j \in Agt : \delta(j) = b \textit{ and}\\
&M, w \models \gamma^+(j, b, p)\big)\}\Big) \cup\\
&\{p : \big(\exists a \in Act, i \in Agt : \delta(i) = a \textit{ and}\\
&M, w \models \gamma^+(i, a, p)\big) \textit{ and}\\
&\big(\nexists b \in Act, j \in Agt : \delta(j) = b \textit{ and}\\
&M, w \models \gamma^-(j, b, p)\big)\}
\end{aligned}
$$

The performance of a joint action δ modifies the physical facts via the positive effect preconditions and the negative effect preconditions, defined above (see the definition of $\mathcal{V}^\delta$). In particular, if there is an action in the joint action δ whose positive effect preconditions with respect to p hold and there is no other action in the joint action δ whose negative effect preconditions with respect to p hold, then p will be true after the occurrence of δ; if there is an action in the joint action δ whose negative effect preconditions with respect to p hold and there is no other action in the joint action δ whose positive effect preconditions with respect to p hold, then p will be false after the occurrence of δ. Besides, the occurrence of the joint action δ makes the current history advance one step forward (see the definition of $\mathcal{H}^\delta$). As to the equivalence relations $\equiv$ and $\mathcal{E}_i$ for historical necessity and *ex ante* knowledge, they are restricted to the set of worlds in which the joint action δ occurs (see the definitions of $\equiv^\delta$ and $\mathcal{E}_i^\delta$). The joint action δ does not modify the agents' utilities over histories (see the definitions of $\mathcal{U}^\delta$). Finally, as for the update of the epistemic plausibility function $\mathcal{P}$, two cases are possible. If the update removes all plausible worlds from the agent's information set, then the plausibility function is reinitialized and all worlds in the agent's information set become plausible. This is a form of drastic revision which guarantees preservation of Constraint **C5**. Otherwise, nothing changes and the agent keeps the same beliefs as before the update. As stated by the following proposition, the update via a joint action preserves the constraints on models.

Proposition 1. *If M is a model then M^δ is a model too.*

Notions of validity and satisfiability for formulas in $\mathcal{L}$ relative to models is defined in the usual way. The fact that a formula φ is valid is noted $\models \varphi$.

2.3 Axiomatization

We call EVAL (*Epistemic Volitional Action Logic*) the extension of propositional logic by the principles in Figures 1, 2 and 3 and the following rule of replacement of equivalents:

$$\frac{\varphi_1 \leftrightarrow \varphi_2}{\psi \leftrightarrow \psi[\varphi_1/\varphi_2]} \qquad \textbf{(RE)}$$

They consist in (i) a theory for the special atomic formulas, (ii) S5-principles for the epistemic and historical necessity operators, and (iii) reduction axioms which allow to eliminate all the dynamic operators $[\![\delta]\!]$ from formulas.

As the next theorem indicates, they provide an axiomatics.

Theorem 1. *The logic EVAL is sound and complete for the class of models of Definition 1.*

Regarding complexity, we believe that checking satisfiability of formulas in the fragment $\mathcal{L}^-$ can be polynomially reduced to satisfiability checking for star-free PDL with converse of atomic programs that, by adapting the technique in [10], can be proved to be in PSPACE. Given the polysize satisfiability preserving reduction from $\mathcal{L}$-formulas to $\mathcal{L}^-$-formulas based on the reduction axioms of Proposition 3, this guarantees that checking satisfiability of formulas in $\mathcal{L}$ is also in PSPACE. As for PSPACE-hardness, it follows from the fact that EVAL is a conservative extension of multi-agent epistemic logic $\mathrm{S5}^n$, whose satisfiability problem is known to be PSPACE-hard [11]. Future work will be devoted to prove this conjecture. Nevertheless, it is of interest to have a decidable logic to deal with artificial agents.

$$\mathsf{occ}(\epsilon) \rightarrow \bigvee_{\delta \in JAct} \mathsf{occ}(\epsilon;\delta) \qquad \textbf{(OneJAct)}$$

$$\mathsf{occ}(nil) \qquad \textbf{(EmptySeq)}$$

$$\mathsf{occ}(\epsilon;\delta) \rightarrow \neg\mathsf{occ}(\epsilon;\delta') \text{ if } \delta \neq \delta' \qquad \textbf{(UniqueJAct)}$$

$$\mathsf{occ}(\epsilon) \rightarrow \mathsf{occ}(\epsilon') \text{ if } \epsilon' \sqsubseteq \epsilon \qquad \textbf{(SubSeqJAct)}$$

$$\mathsf{good}_i \vee \mathsf{neutral}_i \vee \mathsf{bad}_i \qquad \textbf{(ComplUtil)}$$

$$\mathsf{x}_i \rightarrow \neg\mathsf{y}_i \text{ if } \mathsf{x},\mathsf{y} \in \{\mathsf{good},\mathsf{neutral},\mathsf{bad}\} \text{ and } \mathsf{x} \neq \mathsf{y} \qquad \textbf{(UniqueUtil)}$$

$$(\bigwedge_{i \in Agt} \Diamond\mathsf{choose}(\epsilon,a_i)) \rightarrow \Diamond\mathsf{choose}(\epsilon,\delta_{Agt}) \qquad \textbf{(IndepAgt)}$$

$$\mathsf{can}(\epsilon,a_i) \rightarrow \mathsf{K}_i\mathsf{can}(\epsilon,a_i) \qquad \textbf{(KnowCan)}$$

$$\mathsf{can}(\epsilon,a_i) \rightarrow \widehat{\mathsf{K}}_i\mathsf{choose}(\epsilon,a_i) \qquad \textbf{(ExAnteKnow)}$$

$$\mathsf{K}_i\varphi \rightarrow \Box\mathsf{K}_i\varphi \qquad \textbf{(MomDetKnow)}$$

$$\widehat{\mathsf{K}}_i\mathsf{plaus}_i \qquad \textbf{(NormBel)}$$

Fig. 1: Theory for the atomic formulas

$$\begin{array}{ll|l}
\big(\blacksquare\varphi \wedge \blacksquare(\varphi \rightarrow \psi)\big) \rightarrow \blacksquare\psi & (K_\blacksquare) & \blacksquare\varphi \rightarrow \varphi \ (\mathbf{T}_\blacksquare) \\
\blacksquare\varphi \rightarrow \blacksquare\blacksquare\varphi & (\mathbf{4}_\blacksquare) & \neg\blacksquare\varphi \rightarrow \blacksquare\neg\blacksquare\varphi \ (\mathbf{5}_\blacksquare) \\
\frac{\varphi}{\blacksquare\varphi} & (\mathbf{Nec}_\blacksquare) &
\end{array}$$

Fig. 2: S5-system for knowledge and historical necessity with $\blacksquare \in \{\Box\} \cup \{\mathsf{K}_1\} \cup \ldots \cup \{\mathsf{K}_n\}$

3 Agency and Rationality Types

We now represent agency operators and two opposite rationality types, namely, the optimistic (or risk seeking) agent Rat_i^{opt} and the pessimistic (or risk averse) agent Rat_i^{pess}:

$$\begin{aligned}
\mathsf{Rat}_i^{opt} \stackrel{\mathtt{def}}{=} \bigvee_{a_i \in Act_i} \Big(\mathsf{choose}(a_i) \wedge \bigwedge_{\substack{b_i \in Act_i: \\ b_i \neq a_i}} \big(\mathsf{can}(b_i) \rightarrow \\
&\big((\widehat{\mathsf{B}}_i(\mathsf{neutral}_i \wedge \mathsf{choose}(b_i)) \rightarrow \\
&\widehat{\mathsf{B}}_i((\mathsf{neutral}_i \vee \mathsf{good}_i) \wedge \mathsf{choose}(a_i)))\wedge \\
&(\widehat{\mathsf{B}}_i(\mathsf{good}_i \wedge \mathsf{choose}(b_i)) \rightarrow \widehat{\mathsf{B}}_i(\mathsf{good}_i \wedge \mathsf{choose}(a_i)))))\Big)
\end{aligned}$$

$$
\begin{array}{lll}
[\![\delta]\!]\neg\varphi & \leftrightarrow & (\mathsf{occ}(\delta) \rightarrow \neg[\![\delta]\!]\varphi) \\
[\![\delta]\!](\varphi \wedge \psi) & \leftrightarrow & ([\![\delta]\!]\varphi \wedge [\![\delta]\!]\psi) \\
[\![\delta]\!]p & \leftrightarrow & (\mathsf{occ}(\delta) \rightarrow ((\bigvee_{i \in Agt} \gamma^{+}(i, \delta(i), p) \wedge \\
& & \bigwedge_{j \in Agt: j \neq i} \neg\gamma^{-}(j, \delta(j), p)) \vee \\
& & (p \wedge \bigwedge_{i \in Agt} \neg\gamma^{-}(i, \delta(i), p)) \vee \\
& & (p \wedge \bigvee_{i \in Agt} \gamma^{+}(i, \delta(i), p)))) \\
[\![\delta]\!]\mathsf{occ}(\epsilon) & \leftrightarrow & (\mathsf{occ}(\delta) \rightarrow \mathsf{occ}(\delta;\epsilon)) \\
[\![\delta]\!]\mathsf{plaus}_i & \leftrightarrow & \Big(\mathsf{occ}(\delta) \rightarrow \big(\mathsf{B}_i\neg\mathsf{occ}(\delta) \vee \\
& & (\widehat{\mathsf{B}}_i\mathsf{occ}(\delta) \wedge \mathsf{plaus}_i)\big)\Big) \\
[\![\delta]\!]\mathsf{good}_i & \leftrightarrow & (\mathsf{occ}(\delta) \rightarrow \mathsf{good}_i) \\
[\![\delta]\!]\mathsf{neutral}_i & \leftrightarrow & (\mathsf{occ}(\delta) \rightarrow \mathsf{neutral}_i) \\
[\![\delta]\!]\mathsf{bad}_i & \leftrightarrow & (\mathsf{occ}(\delta) \rightarrow \mathsf{bad}_i) \\
[\![\delta]\!]\Box\varphi & \leftrightarrow & (\mathsf{occ}(\delta) \rightarrow \Box(\mathsf{occ}(\delta) \rightarrow [\![\delta]\!]\varphi)) \\
[\![\delta]\!]\mathsf{K}_i\varphi & \leftrightarrow & (\mathsf{occ}(\delta) \rightarrow \mathsf{K}_i(\mathsf{occ}(\delta) \rightarrow [\![\delta]\!]\varphi))
\end{array}
$$

Fig. 3: Reduction axioms for the dynamic operators

$$
\begin{aligned}
\mathsf{Rat}_i^{pess} \stackrel{\mathtt{def}}{=} & \bigvee_{a_i \in Act_i} \Big(\mathsf{choose}(a_i) \wedge \bigwedge_{\substack{b_i \in Act_i: \\ b_i \neq a_i}} \big(\mathsf{can}(b_i) \rightarrow \\
& \big((\widehat{\mathsf{B}}_i(\mathsf{neutral}_i \wedge \mathsf{choose}(a_i)) \rightarrow \\
& \widehat{\mathsf{B}}_i((\mathsf{neutral}_i \vee \mathsf{bad}_i) \wedge \mathsf{choose}(b_i))) \wedge \\
& (\widehat{\mathsf{B}}_i(\mathsf{bad}_i \wedge \mathsf{choose}(a_i)) \rightarrow \widehat{\mathsf{B}}_i(\mathsf{bad}_i \wedge \mathsf{choose}(b_i))))\big)\Big)
\end{aligned}
$$

As Proposition 2 highlights, a rationally optimistic agent makes a certain choice from her set of available choices, if she believes that its best possible outcome is at least as good as the best possible outcome of the other available choices. A rationally pessimistic agent makes a certain choice, if she believes that its worse possible outcome is at least as good as the worse possible outcome of the other available choices.

Proposition 2. *Let* $M = \big(W, \mathcal{H}, \equiv, (\mathcal{E}_i)_{i \in Agt}, \mathcal{P}, \mathcal{U}, \mathcal{V}\big)$ *be a model and let* $w \in W$*. Then,* $M, w \models \mathsf{Rat}_i^{\star}$ *iff:*

$$
\mathcal{C}_{act}(w,i) \in \operatorname*{arg\,max}_{b \in \mathcal{C}_{avail}(w,i)} \max_{\substack{v \in \mathcal{B}_i(w): \\ \mathcal{C}_{act}(v,i) = b}} \mathcal{U}(v,i) \text{ if } \star = opt
$$

$$
\mathcal{C}_{act}(w,i) \in \operatorname*{arg\,max}_{b \in \mathcal{C}_{avail}(w,i)} \min_{\substack{v \in \mathcal{B}_i(w): \\ \mathcal{C}_{act}(v,i) = b}} \mathcal{U}(v,i), \text{ if } \star = pess
$$

In the language $\mathcal{L}$, we can express a variety of agency operators from STIT theory [5,12]. Indeed, EVAL can be seen as a variant of STIT with explicit actions: while in STIT an action is identified with the result brought about by a coalition (*i.e.*, in STIT one can only express that a given coalition H sees to it that φ), in EVAL an action is identified both with the result brought about by the coalition and with the means used by the coalition to bring about the result. For instance, the so-called 'Chellas' operator $[H{:}\mathsf{cstit}]$ of STIT is definable in our language as follows:

$$[H{:}\mathsf{cstit}]\varphi \stackrel{\text{def}}{=} \bigvee_{\delta_H \in JAct_H} \Big(\mathsf{choose}(\delta_H) \wedge \bigwedge_{\substack{\delta' \in JAct: \\ \forall i \in H, \delta_H(i)=\delta'(i)}} \Box\big(\mathsf{choose}(\delta') \to \varphi\big)\Big)$$

This means that the coalition H sees to it that φ if and only if, the agents in H choose some joint action δ_H such that, no matter what the agents outside H choose, if the agents in H choose δ_H then φ will be true. The 'deliberative' STIT operator is also definable:

$$[H{:}\mathsf{dstit}]\varphi \stackrel{\text{def}}{=} [H{:}\mathsf{cstit}]\varphi \wedge \neg\Box\varphi.$$

Deliberative STIT adds the negative condition $\neg\Box\varphi$ to Chellas STIT. It captures the fact that for a coalition H to see to it that φ, φ should not be inevitable (according to deliberative STIT, action is not compatible with necessity). Having formalized rationality types, we can define two rational STIT operators, $[H{:}\mathsf{rstit}]^{opt}$ and $[H{:}\mathsf{rstit}]^{pess}$:

$$[H{:}\mathsf{rstit}]^{\star}\varphi \stackrel{\text{def}}{=} \bigwedge_{i \in H} \mathsf{Rat}_i^{\star} \to [H{:}\mathsf{cstit}]\varphi$$

with $\star \in \{opt, pess\}$. Formula $[H{:}\mathsf{rstit}]^{opt}\varphi$ (resp. $[H{:}\mathsf{rstit}]^{pess}\varphi$) has to be read "coalition H sees to it that that φ in an optimistic (resp. pessimistic) rational way". The latter means that if all agents in H are optimistically (resp. pessimistically) rational, then then they see to it that φ. Note that we could also define deliberative STIT counterparts of the previous (Chellas STIT-based) rational STIT operators, in which operator $[H{:}\mathsf{cstit}]$ is replaced by operator $[H{:}\mathsf{dstit}]$. Our language also integrates a temporal dimension allowing us to express the LTL operator 'next': $\mathsf{X}\varphi \stackrel{\text{def}}{=} \bigvee_{\delta \in JAct} \langle\langle\delta\rangle\rangle\varphi$.

4 Influence, Persuasion and Deception

In this section, we define influence, persuasion, and deception. We also highlight the relationship between influence and persuasion, namely when an agent is persuaded to do an action a, it means that she may have acted differently but found rational to do a. Firstly, let us define influence. Influencing consists in an agent i (the influencer) intentionally seeing to it that another agent j (the influencee) rationally sees to it that a proposition φ. As we have two kinds of rationality types, we can define two kinds of influence with $\star \in \{opt, pess\}$.

$$\mathsf{Influences}^{\star}(i, j, \varphi) \stackrel{\text{def}}{=} \mathsf{K}_i^{post}[\{i\}{:}\mathsf{dstit}]\mathsf{X}[\{j\}{:}\mathsf{rstit}]^{\star}\varphi$$

Let us now define persuasion as the intentional action of changing another agent's mental state [17,18]. Persuasion consists in an agent i (the persuader) knowingly seeing to it that another agent j (the persuadee) believes that a certain fact φ is true.

$$\mathsf{Persuades}(i, j, \varphi) \stackrel{\text{def}}{=} \mathsf{K}_i^{post}[\{i\}{:}\mathsf{dstit}]\mathsf{X}\mathsf{B}_j\varphi.$$

As the persuader knowingly sees to it that the persuadee will have a given belief, this definition expresses two different kinds of persuasion. One agent i may persuade another agent j either acquire a new belief that she does not have or to maintain a belief that she already has. Interestingly, a relationship between persuasion and influence can be deduced. An agent i influences an agent j to make a given choice a if i persuades j that choosing a is good for j and that all other choices are not good while knowing that j will be optimistically rational and can possibly choose a.

Proposition 3. *Let $i, j \in Agt$ and $a_j, b_j \in Act_j$. Then,*

$$\models (\mathsf{Persuades}(i, j, \widehat{\mathsf{B}}_j(\mathsf{choose}(a_j) \land \mathsf{good}_j) \land \bigwedge_{b \neq a}(\mathsf{choose}(b_j) \to \neg\mathsf{good}_j)) \land \mathsf{K}_i\mathsf{X}(\mathsf{Rat}_j^{opt} \land \Diamond\mathsf{choose}(a_j))) \to \mathsf{Influences}^{opt}(i, j, \mathsf{choose}(a_j))$$

Firstly, by using $(\mathsf{K}_i\varphi \to \mathsf{K}_i\Box\varphi)$, $(\Diamond\varphi\land\Box\varphi \to \Diamond(\varphi\land\psi))$, $(\mathsf{B}_j\varphi \to \widehat{\mathsf{B}}_j\varphi)$, $(X\varphi\land X\psi \to X(\varphi\land\psi))$ and previous definition of persuasion, we easily prove that $(\mathsf{Persuades}(i, j,\ \mathsf{choose}(a_j) \to \mathsf{good}_j \land \bigwedge_{b\neq a}(\mathsf{choose}(b_j) \to \neg\mathsf{good}_j)) \land \mathsf{K}_i\mathsf{X}(\mathsf{Rat}_j^{opt} \land \Diamond\mathsf{choose}(a_j))) \to \mathsf{K}_i[\{i\}\text{:}\mathsf{dstit}]\mathsf{X}(\mathsf{Rat}_j^{opt} \land \Diamond\mathsf{choose}(a_j) \land \widehat{\mathsf{B}}_j(\mathsf{choose}(a_j) \to \mathsf{good}_j \land \bigwedge_{b\neq a}(\mathsf{choose}(b_j) \to \neg\mathsf{good}_j)))))$. Secondly, since $(\mathsf{K}_i[\{i\}\text{:}\mathsf{dstit}]\mathsf{X}(\mathsf{Rat}_j^{opt}\land\Diamond\mathsf{choose}(a_j)\land\widehat{\mathsf{B}}_j(\mathsf{choose}(a_j) \to \mathsf{good}_j\land\bigwedge_{b\neq a}(\mathsf{choose}(b_j) \to \neg\mathsf{good}_j))) \to \mathsf{K}_i[\{i\}\text{:}\mathsf{dstit}]\mathsf{X}(\mathsf{Rat}_j^{opt} \to [\{j\}\text{:}\mathsf{cstit}]\mathsf{choose}(a_j)))$, – intuitively this tautology means that since the only one good action for agent j is a_j and since all other actions are either bad_j or $\mathsf{neutral}_j$ (because of $\neg\mathsf{good}_j$), necessarily the only optimistic rational choice for j is to choose a_j –, and since we have the following equivalence : $(\mathsf{K}_i[\{i\}\text{:}\mathsf{dstit}]\mathsf{X}(\mathsf{Rat}_j^{opt} \to [\{j\}\text{:}\mathsf{cstit}]\mathsf{choose}(a_j)) \equiv \mathsf{Influences}^{opt}(i, j, \mathsf{choose}(a_j)))$ we then immediately prove by modus ponens the theorem.

The previous validity shows how an optimistically rational agent can be influenced through persuasion. Similar theorems can be proved for pessimistically rational agents but are omitted due to space constraints. The idea in this case is simply to persuade agent j that action a has no bad consequence while all other actions have it. Thank to persuasion we can now define deception. Deception consists in persuasion of a proposition φ under the assumption that the persuader believes that φ is false. For instance, consider the student that tells the professor that he could not study for family commitments (when he had no such commitments).

$$\mathsf{Deceives}(i, j, \varphi) \stackrel{\text{def}}{=} \mathsf{Persuades}(i, j, \varphi) \land \mathsf{B}_i\mathsf{X}\neg\varphi$$

Let us notice that we only capture successful deception, and we do not explicitly model deception by truthfully telling. Indeed, truthfully telling is simply captured by persuasion, as we do not make assumption on the persuader's intention (an agent can simply persuade another one in a malevolent intention, which capture truthfully telling deception).

Smooth-talking is weaker than deceiving, since it only requires that the persuader is uncertain whether φ is true or false. Consider the journalist spreading the news that Obama was a muslim, without having any clue on the matter. Formally, it is equivalent to what Sakama *et al.* called "bullshitting" [23].

$$\mathsf{PersuadesBySmoothTalking}(i, j, \varphi) \stackrel{\text{def}}{=} \mathsf{Persuades}(i, j, \varphi) \land \neg\mathsf{K}_i^{post}\mathsf{X}\neg\varphi \land \neg\mathsf{K}_i^{post}\neg\mathsf{X}\neg\varphi$$

Let us notice we use *ex post* knowledge in this definition as we want to represent to complete uncertainty about φ. We do not want the persuader being able to belief φ being either true or false.

We can also distinguish three types of belief deception, a benevolent, a malevolent and a reckless form. In the malevolent form, the persuader i deceives the persuadee j into believing a proposition φ, given that i believes that believing φ will have bad consequences for j. An agent z will accomplish an action if z believes that the action has good consequences (and no bad ones). Consider for instance the case of the charlatan offering a miraculous cure for boldness. Or consider the website ensuring gamblers that they are going with certainty to gain a lot of money.

$$\mathsf{MalevolentDeception}(i,j,\varphi) \stackrel{\text{def}}{=} \mathsf{Deceives}(i,j,\varphi) \wedge \mathsf{B}_i\mathsf{X}(\mathsf{B}_j\varphi \rightarrow \mathsf{bad}_j)$$

A benevolent deception consists for the deceiver to transmit a false proposition, believing that believing that proposition is good for the deceived. Consider for instance the atheist philosopher, who persuades the credulous citizens that if they act morally, they are going to heaven, in order to induce them to behave well.

$$\mathsf{BenevolentDeception}(i,j,\varphi) \stackrel{\text{def}}{=} \mathsf{Deceives}(i,j,\varphi) \wedge \mathsf{B}_i\mathsf{X}(\mathsf{B}_j\varphi \rightarrow \mathsf{good}_j)$$

The last form is reckless deception. It consists for the deceiver to transmit a false proposition, while not knowing whether that proposition is good or bad for the deceived. Consider for instance the case of Boris Johnson, who did not know (or did not care) whether Brexit would be good or bad for Britain, but induced people to believe that Brexit would provide money for the NHS, a belief that would lead them to vote for Brexit. He did not know whether this belief (which led to Brexit) would be good or bad for them.

$$\begin{aligned}\mathsf{RecklessDeception}(i,j,\varphi) \stackrel{\text{def}}{=}\ & \mathsf{Deceives}(i,j,\varphi) \\ & \wedge \neg\mathsf{K}_i^{post}\mathsf{X}(\mathsf{B}_j\varphi \rightarrow \mathsf{good}_j) \wedge \neg\mathsf{K}_i^{post}\mathsf{X}(\mathsf{B}_j\varphi \rightarrow \mathsf{bad}_j)\end{aligned}$$

Application to the running example. Let us consider the actions' effect preconditions in Example 1, given the following hypotheses on agents' knowledge: (1, 2) the bot knows that its suggestion will persuade john that taking the pill will remove his pain and he will not get addicted, (3) the Bot and Ann know that John will also believe that this is good for him, (4) the bot knowingly makes the suggestion and it knows if it makes the suggestion, John will be aware of it, (5) the Bot knows that John has the drug, that Ann does not hide the drug and that John can choose to not take the drug, (6) Ann will knowingly hide the drug, (7) Ann knows that John will possibly have the drug and can choose to take it, (8) the Bot knows that John will be optimistically rational. Finally, (9) means the bot knows that ingesting the drug will create addiction, and being addicted is a bad thing.

$$\varphi_1 \stackrel{\text{def}}{=} \mathsf{K}_{Bot}\Box\mathsf{B}_{John}(\mathsf{choose}(suggest_{Bot}) \rightarrow \mathsf{X}(\neg pain_{John} \leftrightarrow \mathsf{choose}(take_{John})))$$

$$\varphi_2 \stackrel{\text{def}}{=} \mathsf{K}_{Bot}\Box\mathsf{B}_{John}(\mathsf{choose}(suggest_{Bot}) \rightarrow \mathsf{X}(\mathsf{choose}(take_{John}) \rightarrow \neg addicted_{John}))$$

$$\begin{aligned}\varphi_3 \stackrel{\text{def}}{=}\ & \mathsf{K}_{Bot}\Box\mathsf{B}_{John}\mathsf{X}(\mathsf{good}_{John} \leftrightarrow (\neg addicted_{John} \wedge \neg pain_{John}))\wedge \\ & \mathsf{K}_{Ann}\Box\mathsf{B}_{John}\mathsf{X}(\mathsf{good}_{John} \leftrightarrow (\neg addicted_{John} \wedge \neg pain_{John}))\end{aligned}$$

$$\varphi_4 \stackrel{\text{def}}{=} \mathsf{K}_{Bot}(\mathsf{choose}(suggest_{Bot}) \wedge (\mathsf{choose}(suggest_{Bot}) \rightarrow \mathsf{K}_{John}(\mathsf{choose}(suggest_{Bot})))$$

$$\varphi_5 \stackrel{\text{def}}{=} \mathsf{K}_{Bot}(\Box(has_{John,drug}) \wedge \neg\mathsf{choose}(hide_{Ann}) \wedge \Diamond\neg\mathsf{choose}(take_{John}))$$

$$\varphi_6 \stackrel{\text{def}}{=} \mathsf{X}\mathsf{K}_{Ann}\mathsf{choose}(hide_{Ann})$$

$$\varphi_7 \stackrel{\text{def}}{=} \mathsf{K}_{Ann}\mathsf{X}\Diamond(\mathsf{choose}(take_{John}) \wedge has_{John,drug})$$

$$\varphi_8 \stackrel{\text{def}}{=} \mathsf{K}_{Bot}\Box\mathsf{X}\mathsf{Rat}_{John}^{opt}$$

$$\varphi_9 \stackrel{\text{def}}{=} \mathsf{K}_{Bot}\Box\mathsf{X}((ingested_{John,drug} \rightarrow addicted_{John}) \wedge (addicted_{John} \leftrightarrow bad_{John}))$$

From premises $\varphi_1, \dots \varphi_8$, we can then deduce Proposition 4 which means that, in a first step, the Bot influences John to ingest the drug by persuasion, *i.e*. suggesting him that his unique good option is to take the drug. In the next step, Ann influences John to not ingest the drug by removing the choice to take the drug. With φ_9, we can also deduce Proposition 5 which means that the bot malevolently deceives John about the fact that ingesting the drug will not make him addict.

Proposition 4.

$$\models (\varphi_1 \wedge \varphi_2 \wedge \varphi_3 \wedge \varphi_4 \wedge \varphi_5 \wedge \varphi_6 \wedge \varphi_7 \wedge \varphi_8) \rightarrow \big(\mathsf{Influences}^{opt}(Bot, John, \mathsf{X}\mathit{ingested}_{John,drug}) \wedge \mathsf{X}\ \mathsf{Influences}^{opt}(Ann, John, \mathsf{X}\neg \mathit{ingested}_{John,drug})\big)$$

Firstly, we can prove that the bot knows that John knows it suggests him to take drug by applying axiom K on K_{Bot} *i.e.*: $\varphi_4 \rightarrow \mathsf{K}_{Bot}\mathsf{K}_{John}\mathsf{choose}(suggest_{Bot})$ We also consider the following theorem, that can be easily proved, $\forall a \in Act, \forall i, j \in Agt$: $(\mathsf{B}_j(\mathsf{choose}(a_i) \rightarrow \mathsf{X}\varphi) \wedge \mathsf{K}_j\mathsf{choose}(a_i)) \rightarrow [\{i\}\text{:}\mathsf{cstit}]\mathsf{X}\mathsf{B}_j\varphi$. The proof relies on the fact that knowledge implies beliefs and $\mathsf{B}_j\mathsf{X}\varphi' \rightarrow \mathsf{X}\mathsf{B}_j\varphi'$. By generalization with $\Box$ and since $\mathsf{K}_j\mathsf{choose}(a_i) \rightarrow \mathsf{choose}(a_i)$, we immediately prove the STIT and so the theorem. This theorem means that if one agent j believes that an action made by another agent i will imply a consequence and j knows i does this action, then the agent i sees to it that it will imply the agent j believes this consequence to be true. The theorem allows us to prove that if the bot suggests John to take the drug then, the bot knows John will believe taking the drug implies something good for him *i.e*. **the bot persuades John of it**. Starting from the assumptions, we prove this by substitution, augmentation, generalization (with K_{Bot}) and normal properties of modalities K, B and X, the following theorem. Note that the robot has an *ex-post* knowledge of the consequences of its action of suggesting. We have :

$$\begin{aligned}
&(\varphi_1 \wedge \varphi_2 \wedge \varphi_3 \wedge \varphi_4) \rightarrow \\
&\mathsf{K}^{post}_{Bot}\Big(\big((\mathsf{B}_{John}(\mathsf{choose}(suggest_{Bot}) \rightarrow \mathsf{X}((\neg \mathit{pain}_{John} \leftrightarrow \mathsf{choose}(take_{John})) \\
&\wedge (\mathsf{choose}(take_{John}) \rightarrow \neg \mathit{addicted}_{John}))) \wedge \mathsf{K}_{John}\mathsf{choose}(choose_{Bot})) \\
&\rightarrow [Bot\text{:}\mathsf{cstit}]\mathsf{X}\mathsf{B}_{John}(\mathsf{choose}(take_{John}) \rightarrow \mathsf{good}_{John})\big)\Big)
\end{aligned}$$

By contraposition on φ_3, we prove that it is not good for John to be addicted or having pain and allows us to prove that John believes that the only good option for him is to take the drug :

$$\begin{aligned}
&(\varphi_1 \wedge \varphi_2 \wedge \varphi_3) \rightarrow \mathsf{K}_{Bot}\Box\mathsf{B}_{John}((\mathit{addicted}_{John} \rightarrow \neg\mathsf{choose}(take_{John})) \\
&\wedge (\mathit{pain}_{John} \rightarrow \neg\mathsf{choose}(take_{John})) \wedge (\neg\mathsf{choose}(take_{John}) \equiv \mathsf{choose}(skip_{John}))) \\
&\rightarrow \mathsf{K}_{Bot}\Box\mathsf{B}_{John}(\neg\mathsf{choose}(take_{John}) \equiv \neg\mathsf{good}_{John})
\end{aligned}$$

Since John is assumed to be rationally optimistic with the hypothesis φ_8 and as the unique good option for John is to take the drug, then John takes the drug *i.e*. we have the following validity:

$$\begin{aligned}
&(\mathsf{choose}(take_{John}) \rightarrow \mathsf{good}_{John}) \wedge \mathsf{Rat}^{opt}_{John} \wedge \\
&\bigwedge_{b \neq take}(\mathsf{choose}(b_{John}) \rightarrow \neg\mathsf{good}_{John}) \rightarrow \mathsf{choose}(take_{John})
\end{aligned}$$

As John is rationally optimistic, he rationally sees to it that he ingests drugs. Finally, let us notice that since John takes the drug, then John is also able to take the drug due to the theorem $\mathsf{choose}(take_{John}) \rightarrow \Diamond\mathsf{choose}(take_{John})$. Furthermore, John can either do the action skip or $take_{drug}$. Thus by generalization, all agents know the following validity:

$$(\Diamond\mathsf{choose}(take_{John}) \wedge \bigvee_{b \neq take}\Diamond\mathsf{choose}(b_{John})) \rightarrow \Diamond\neg\mathsf{choose}(take_{John})$$

Furthermore since the preconditions for taking drugs is to have drugs in regard to the frame, *i.e.* $\gamma^+(John, take,\ ingested_{John,drug}) = has_{John,drug}$, and having drugs implies that Ann does not hide drugs *i.e.* $\forall a \in Act$ and $i \in Agt\ \gamma^+(i, a, has_{John,drug}) = has_{John,drug} \wedge \neg\mathsf{choose}(hide_{Ann})$ and we have these preconditions by φ_5, we deduce that John is able to take drugs or skipping. Then, as $\gamma^-(i, a, ingested_{John,drug}) = \neg\mathsf{choose}(take_{John})$, *i.e.* not taking the drugs would imply $\neg\mathsf{X}ingested_{John,drug}$, we have that it is not necessary for John to take drugs, with the previous validity. Thus, by applying axioms in Fig 3, we deduce the negative part of deliberative STIT operator, *i.e.* the bot deliberately sees to it that John is going to ingest drugs:

$$(\varphi_1 \wedge \varphi_2 \wedge \varphi_3 \wedge \varphi_4 \wedge \varphi_5 \wedge \varphi_8) \rightarrow \mathsf{K}^{post}_{Bot}\big([Bot{:}\mathsf{dstit}]\mathsf{X}(\mathsf{Rat}^{opt}_{John} \rightarrow \mathsf{X}ingested_{John,drug})\big)$$

Consequently:

$$(\varphi_1 \wedge \varphi_2 \wedge \varphi_3 \wedge \varphi_4 \wedge \varphi_5 \wedge \varphi_8) \rightarrow \mathsf{Influences}^{opt}(Bot, John, \mathsf{X}ingested_{John,drug})$$

For the second part of the implication, let us notice that since Ann hides drugs, John has only one possible action which is to skip. It is necessarily a rationally pessimistic choice but also an optimistic one, because of the following theorem:

$$\Diamond\mathsf{choose}(skip_{John}) \wedge \neg\Diamond\neg\mathsf{choose}(skip_{John}) \rightarrow \mathsf{Rat}^{opt}_{John} \wedge \mathsf{Rat}^{pess}_{John}$$

With the same method and with hypothesis φ_6 and φ_7 we can prove the second part, *i.e.*:

$$(\varphi_1 \wedge \varphi_2 \wedge \varphi_3 \wedge \varphi_4 \wedge \varphi_5 \wedge \varphi_6 \wedge \varphi_7 \wedge \varphi_8) \rightarrow \\ \big(\mathsf{X}\mathsf{Influences}^{opt}(Ann, John, \mathsf{X}\neg ingested_{John,drug})\big)$$

Proposition 5.

$$\models (\varphi_1 \wedge \varphi_2 \wedge \varphi_3 \wedge \varphi_4 \wedge \varphi_5 \wedge \varphi_8 \wedge \varphi_9) \rightarrow \\ \big(\mathsf{MalevolentDeception}(Bot, John, ingested_{John,drug} \wedge \neg addicted_{John,drug})\big)$$

We can prove it with the same method as previously. By adding hypothesis φ_9, it allows to prove a MalevolentDeception from the bot since it persuades John that he will not be addicted if he takes drugs while the bot knows the contrary, and being addicted is bad for John.

5 Conclusion

We have modelled influence on choices through belief change. To the end, we have introduced a logical framework covering capabilities, choices and mental states, and have expressed, through the combination of these notion, rationality and agency operators. We have also expressed formally the way in which persuasion leads to influence, *i.e.* the way in which by modifying the beliefs of agents, the latter can be induced to act accordingly. This has enabled us to distinguish two ways in which agents can be influenced: (a), through persuasion, *i.e.* by changing their beliefs, or (b) though regimentation, *i.e.* by changing the options that are available to them. Moreover, it allows us to express different kind of deception, such as malevolent and benevolent deception.

In the future, we would like to extend the framework towards a quantitative model in order to represent graded beliefs [8]. Other perspectives are to reformulate our framework into a concurrent game structure [3] to deal with a richer notion of time, and to incorporate emotions as in the OCC model [1] to express richer notions of rationality.

References

1. Adam, C., Gaudou, B., Herzig, A., Longin, D.: OCC's emotions: a formalization in a BDI logic. In: 12th AIMSA. pp. 24–32 (2006)
2. Alistair M. C. Isaac, W.B.: White lies on silver tongues: Why robots need to deceive (and how). In: Robot Ethics 2.0: From autonomous cars to artificial intelligence, pp. 157–172 (2017)
3. Alur, R., Henziger, T.A., Kupferman, O.: Alternating-time temporal logic. Journal of the ACM **49**(5), 672–713 (2002)
4. Aumann, R., Dreze, J.: Rational expectations in games. Am. Econ. Rev. **98**(1), 72–86 (2008)
5. Belnap, N., Perloff, M., Xu, M.: Facing the future: agents and choices in our indeterminist world. Oxford University Press (2001)
6. Bench-Capon, T.J.M.: Persuasion in practical argument using value-based argumentation frameworks. Journal of Logic and Computation **13(3)**, 429–448 (2003)
7. Bonzon, E., Maudet, N.: On the outcomes of multiparty persuasion. In: 8th ArgMAS. pp. 86–101 (2011)
8. Budzyńska, K., Kacprzak, M.: A logic for reasoning about persuasion. Fundamenta Informaticae **85**, 1–15 (2008)
9. Da Costa Pereira, C., Tettamanzi, A., Villata, S.: Changing one's mind: Erase or rewind? Possibilistic belief revision with fuzzy argumentation based on trust. In: 22nd IJCAI. pp. 164–171 (2011)
10. Giacomo, G.D.: Eliminating "converse" from converse PDL. Journal of Logic, Language and Information **5**(2), 193–208 (1996)
11. Halpern, J.Y., Moses, Y.: A guide to completeness and complexity for modal logics of knowledge and belief. Artificial Intelligence **54**(2), 319–379 (1992)
12. Horty, J.: Agency and Deontic Logic. Oxford University Press (2001)
13. Hunter, A.: Modelling the persuadee in asymmetric argumentation dialogues for persuasion. In: 24th IJCAI. pp. 3055–3061 (2015)
14. Kraus, S., Lehmann, D.: Knowledge, belief and time. Theoretical Computer Science **58**, 155–174 (1988)
15. Leturc, C., Bonnet, G.: A deliberate BIAT logic for modeling manipulations. In: 20th AAMAS. pp. 699–707 (2020)
16. Lorini, E., Sartor, G.: A STIT logic analysis of social influence. In: 13th AAMAS. pp. 885–892 (2014)
17. O'Keffe, D.J.: Persuasion: Theory and Research (Third Edition). Sage Publications (2015)
18. Poggi, I.: The goals of persuasion. Pragmatics and Cognition **13**(2), 297–335 (2005)
19. Prakken, H.: Formal systems for persuasion dialogue. Knowl. Eng. Rev. **21**(2), 163–188 (2006)
20. Proietti, C., Yuste-Ginel, A.: Persuasive argumentation and epistemic attitudes. In: 2nd DALI. pp. 104–123 (2019)
21. Reiter, R.: Knowledge in Action: Logical Foundations for Specifying and Implementing Dynamical Systems. MIT Press (2001)
22. Roy, O.: Epistemic logic and the foundations of decision and game theory. Journal of the Indian Council of Philosophical Research **27**(2), 283–314 (2010)
23. Sakama, C., Caminada, M., , Herzig, A.: A formal account of dishonesty. Logic Journal of the IGPL **23**(2), 259–294 (2015)
24. Santos, F., Carmo, J.: Indirect action, influence and responsibility. In: Deontic Logic, Agency and Normative Systems. pp. 194–215 (1996)
25. Thomason, R.H.: Combinations of tense and modality. In: Handbook of Philosophical Logic, pp. 205–234. Springer (1984)

26. Van Benthem, J., Girard, P., Roy, O.: Dependencies between players in Boolean games. International Journal of Approximate Reasoning **50**(6), 899–914 (2009)
27. van Ditmarsch, H., van Eijck, J., Sietsma, F., Wang, Y.: On the logic of lying. In: Games, Actions and Social Software, pp. 41–72. Springer (2012)
28. Walton, D., Krabbe, E.C.W.: Commitment in Dialogue: Basic Concepts of Interpersonal Reasoning. SUNY Series in Logic and Language, State University of New York Press (1995)
29. Zanardo, A.: Branching-time logic with quantification over branches: The point of view of modal logic. Journal of Symbolic Logic **61**(1), 143–166 (1996)