A Deliberate BIAT Logic for Modeling Manipulations

Christopher Leturc, Grégory Bonnet

A Deliberate BIAT Logic for Modeling Manipulations

Christopher Leturc
Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC
Caen, France
christopher.leturc@unicaen.fr

Grégory Bonnet
Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC
Caen, France
gregory.bonnet@unicaen.fr

ABSTRACT

In many applications, selfish, dishonest or malicious agents may find an interest in manipulating others. While many works deal with designing robust systems, few works deal with logical reasoning about manipulation. Based on social science literature, we propose a new logical framework to express and reason about manipulation, defined as a deliberate effect to instrumentalize a victim while making sure to conceal that instrumentalization. Since manipulation relies on deliberate effects of a manipulator, we propose a new BIAT operator to catch deliberate effects. We first prove that this logical framework is sound and complete. Then we formally define manipulation and we show our logical framework also expresses related notions such as *coercion*, *persuasion*, or *deception*.

KEYWORDS

Logics for agents and multi-agent systems; Reasoning in agent-based systems

ACM Reference Format:
Christopher Leturc and Grégory Bonnet. 2020. A Deliberate BIAT Logic for Modeling Manipulations. In *Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), Auckland, New Zealand, May 9–13, 2020,* IFAAMAS, 9 pages.

In many applications, selfish, dishonest or malicious agents may find an interest in manipulating others. In computer science and social science, manipulation is viewed as controlling or influencing somebody or something, often in a dishonest way so that they do not realize it. For example, reputation systems evaluate the trust that one agent should place in another depending on other agent's testimonies. In those systems, agents may have interest in lying so as to mislead others, and push them to interact with some specific agents [18, 19, 30]. This is manipulation in the sense that, to be effective, the liar must ensure that the other agents are unaware he intended to mislead them.

In the field of artificial intelligence, many works dealt with manipulation in social choice theory [12, 26, 29, 33], game theory [11, 46], recommendation systems [24, 28]. However, very few works have focused on modeling manipulation using a formal logic to describe and reason about it. In modal logic literature, some related works have examined modeling social influence [22] and deception [32, 45]. Interestingly, manipulation can be seen as the combination of both approaches, namely modeling the deliberate effect of influencing another agent while concealing this influence. By *deliberate effect*, we mean an effect that an agent has fully decided and anticipated [8]. Furthermore, in this article, rather than talking about influence, we prefer talking about *instrumentalization*. According to the Oxford dictionnary, it means "To make or render (something) instrumental to accomplishing a purpose or result; to use as a means to an end". Thus instrumentalization can be seen as a special case of influence restricted to other agents' actions.

Thus, we propose in this article a new modal logic that allows to reason about manipulation. The contribution of this work is twofold. Firstly, we propose a new deliberate BIAT modality which does not already exist in the literature and combine it with a classical STIT modality to catch all consequences of actions and side-effects of actions. Secondly, we use this framework to provide a formal definition of manipulation, based on social science literature.

The remainder of this article is structured as follows. In Section 1, we propose a general definition of manipulation and make a review of related works in modal logic literature. In Section 2, in view of the previous definition of manipulation, we survey available logical tools that are able to express it. In Section 3, we propose a logical framework and show that it is sound and complete. In Section 4, we formally define manipulation, and show our formal framework also models *coercion*, *persuasion* and *deception*. Finally, in Section 5 we instantiate an example.

1 FORMALIZING MANIPULATION

In this section, we first present several works about manipulation from the point of view of social science and use them to give a general synthetic definition of manipulation. We then present some related works about modeling influence and dishonesty in multi-agent systems.

1.1 Manipulation in social science

In the field of politics, marketing and psychiatry, manipulation is sometimes defined as the act of altering the judgment of individuals, depriving them of part of their judgment and deliberate choices [9, 21, 39]. However, according to most psychologists, this definition brings rational persuasion, deception or even coercion into the field of manipulation while "most people would distinguish manipulation from persuasion, on one hand, and from coercion, on the other" [31]. There seems to be a consensus on that "manipulation is not exactly coercion, not precisely persuasion, and not entirely similar to deception" [15].

One of the major characteristic of manipulation is that it is an invisible exercise of power and so it is necessarily hidden from the target. As Goodin [14] stated: "One person manipulates another when he deceptively influences him, causing the other to act contrary to his putative will. Manipulation is something which actually happens invisibly. [...] By the time we start talking about manipulation at all, the act has been already exposed. But when we do speak of manipulation, either in the past tense ('I was manipulated') or in

Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), B. An, N. Yorke-Smith, A. El Fallah Seghrouchni, G. Sukthankar (eds.), May 9–13, 2020, Auckland, New Zealand.

the second third person ('You/they are being manipulated') we think we really are telling him something he does not already know". Similarly, Handelman [15] highlights that "the inevitable conclusion is that during a manipulation interaction the target cannot identify that he operates under a manipulative influence". Consequently manipulation is an instrumentalization and it is hidden from the target. Hence, by considering a synthesis of the definitions given in [2, 10, 15, 40] and adapting it to multi-agent systems, we retain the following definition of manipulation.

Definition 1.1. An agent (called a manipulator) is manipulating another agent (called a victim) if, and only if, the manipulator deliberately instrumentalizes the victim while making sure to conceal that instrumentalization to the victim.

1.2 Related works in artificial intelligence

In artificial intelligence, works on manipulation only focus either on exhibiting manipulations to show a system's weakness, or on designing robust systems such as in reputation systems [42], in social choice theory [12] or in game theory [46]. However these approaches do not express manipulation as social science does but only focus on decision-making processes. Modal logics allow to explicitly describe notions of intention, belief and knowledge that are fundamental to manipulations. Several logical approaches have already studied similar notions such as social influence [7, 22, 34], lying and deception [32, 45].

Concerning the social influence, Lorini and Sartor [22] define with a STIT logic the fact that an agent i makes sure that another agent j realizes in the future the action expected by i. Bottazzi and Troquard [7] define the influence with a BIAT logic as the effect of an agent i to bring another agent j to do something. Let us notice that STIT and BIAT formalisms appear to be appropriate to model instrumentalization and we detail them in Section 2.2.

Concerning deception, lying is the intention of an agent i to inform an agent j about something in order to make j believing it while the agent i believes in its opposite [32, 45]. Van Ditmarsch et al. [45] use dynamic doxastic logics with a modality to describe the action of private announcements which is used to describe lying. Sakama et al. [32] use a modal logic and introduce a modality of communication between two agents, a modality of belief as well as a modality of intention in order to express lying, bullshitting or deception by withholding information.

After having reviewed works on manipulation in point of view of social science and then by having reviewed works on formal logics that deal with related notions of deception and social influence, there is, to the best of our knowledge, no work in formal logics that deals with modeling manipulation.

2 INGREDIENTS TO MODEL MANIPULATION

In this section, since manipulation is a deliberate effet with concealment, we survey related works in literature about how to represent concealment and deliberate effects.

2.1 Manipulation as a lack of being aware?

Logicians [17, 35, 43] expressed the meaning of being aware about something in dynamic logical frameworks. According to them, an agent i is aware of a formula ϕ if ϕ is in the agent's awareness correspondence function [35]. This function is a set-valued function that associates for each possible world and each agent, the set of formula that the agent is aware of. We might think that we need to consider concealment of manipulation as a lack of being aware rather than a lack of knowledge. However in this work, we consider concealment of manipulation as a lack of knowledge and we let the awareness representation as a future perspective.

2.2 Modeling deliberate effects

Since manipulation is a deliberate effect, we need an action logic to represent both deliberate effects and consequences of actions. To formalize the notion of actions in logic, many formalisms exist. For instance, dynamic logics [16] and temporal logics [3] consider several action modalities where each action modality is associated with a program and its outputs. Giordano et al. [13] consider distinct modalities for each possible action and add to their formalism a consequence operator in order to catch causality and ramifications. Dynamic epistemic logics express the logical consequences generated by public or private announcements of agents [44]. Many other formalisms exist such as *fluent calculus*. For a detailed survey, the interested reader may refer to Segerberg et al. [36]. However in our case, we do not want to consider explicit actions as distinct modalities because manipulation can take many forms (e.g. lies, rumor propagation, emotional blackmail) and does not depend on particular actions but rather on its results. Thus, two approaches seem relevant: the STIT [4, 5, 22] and the BIAT formalism [27, 34, 41], which both consider, in an abstract way, the fact of ensuring that something is done.

Both STIT and BIAT formalisms consider actions as the fruit of their consequences. This level of abstraction is well-adapted to define manipulation. The BIAT approaches consider a modality E_i which means that the agent i *brings it about*. BIAT is *side-effects free*, i.e. indirect consequences of actions are not considered as intended effects. While STIT approaches represent a modality $[STIT]_i$ which describes the fact that the agent i *sees to it that* and catches *all consequences of actions*. Although these two approaches are often confused, the main difference between these two formalisms lies in the semantics of these modalities. STIT approaches consider a S5 system whereas BIAT is a non normal system based on neighborhood functions. Furthermore, standard STIT logics use a notion of temporality while BIAT logics do not.

In the literature, STIT approaches already defined a deliberate effect. Lorini and Sartor [22] consider that something is done deliberately by one agent i if i sees to it that something is done while it is not necessarily the case. Let us imagine a situation in which one agent i caused a car crash deliberately to take advantage of car insurance. But after this car accident, a person was dead on the road. By following the formal definition of Lorini and Sartor's deliberate STIT, we would deduce that the agent i deliberate sees to it that "the car is crashed" but also, all indirect consequences as "a person is dead". Consequently by following STIT reasoning and since it was not necessarily the case (if the agent did not choose to cause this accident) that the person is dead, we would also deduce that i deliberate sees to it that "the person is dead". However we claim the opposite. Even if the agent i deliberately caused this car accident, he did not deliberately kill the victim. Furthermore a deliberate

effect must be known by the agent. Indeed when we deliberate do or do not something, then we know what we are doing. Because they use standard STIT approach, Lorini and Sartor do not and cannot consider positive and negative introspection on knowledge. To the best of our knowledge, there is no other deliberate effect operator with the following properties:

- negative and positive introspection;
- side-effect free.

Consequently, since BIAT is side-effect free and makes it easy to express positive and negative introspection, we define in the sequel a new deliberate BIAT operator that takes into account these points.

3 A MODAL LOGIC FOR MANIPULATION

We propose in this section a modal logic that considers several modalities: deliberate effects, all consequences of actions, belief and knowledge. As explained previously, STIT semantic catches all (and indirect) consequences of actions while BIAT semantics catches the deliberate effects. Thus, we distinguish a modality of deliberate effects (expressed by a BIAT-like modality E_i^d) from a notion to capture consequences of actions performed (expressed by a STIT-like modality E_i). Thanks to these modalities we are able to express instrumentalization and concealment.

3.1 Language

Let $\mathcal{P} = \{a, b, c, ...\}$ be a set of propositional letters, and $\mathcal{N}$ be a finite set of agents with $i, j \in \mathcal{N}$ two agents, and $p \in \mathcal{P}$ be a propositional variable. We define $\mathcal{L}_{KBE}$ the language with the following BNF grammar rule:

$$\phi ::= p \mid \neg\phi \mid \phi \vee \phi \mid \phi \wedge \phi \mid \phi \Rightarrow \phi \mid K_i\phi \mid B_i\phi \mid E_i\phi \mid E_i^d\phi$$

The formula $E_i\phi$ means that the actions of i lead to ϕ. So E_i represents the effects of actions which may have been deliberated or not, such as side-effects. The formula $E_i^d\phi$ means that ϕ is a deliberate effect[1] by agent i. This modality is semantically represented with a neighborhood function. Each deliberate effect is represented as a set of possible worlds. Hence, the set of all sets of worlds in the neighborhood function represents all deliberate effects. Finally the formulas $K_i\phi$ and $B_i\phi$ mean respectively that the agent knows that ϕ, and the agent believes that ϕ.

3.2 Associated semantics

We consider the following logical frame:

$$C = (\mathcal{W}, \{\mathcal{B}_i\}_{i\in\mathcal{N}}, \{\mathcal{K}_i\}_{i\in\mathcal{N}}, \{\mathcal{E}_i\}_{i\in\mathcal{N}}, \{\mathcal{E}_i^d\}_{i\in\mathcal{N}})$$

where $\mathcal{W}$ is a nonempty set of possible worlds, $\{\mathcal{B}_i\}_{i\in\mathcal{N}}$, $\{\mathcal{K}_i\}_{i\in\mathcal{N}}$, $\{\mathcal{E}_i\}_{i\in\mathcal{N}}$ are sets of binary relationships, and $\{\mathcal{E}_i^d\}_{i\in\mathcal{N}}$ is a set of neighborhood functions i.e. $\forall i \in \mathcal{N}, \mathcal{E}_i^d : \mathcal{W} \to 2^{2^{\mathcal{W}}}$.

We define a model as $\mathcal{M} = (\mathcal{W}, \{\mathcal{B}_i\}_{i\in\mathcal{N}}, \{\mathcal{K}_i\}_{i\in\mathcal{N}}, \{\mathcal{E}_i\}_{i\in\mathcal{N}}, \{\mathcal{E}_i^d\}_{i\in\mathcal{N}}, V)$ with $V : \mathcal{P} \to 2^{\mathcal{W}}$ an interpretation function. For all $w \in \mathcal{W}$, and $\phi, \psi \in \mathcal{L}_{KBE}$, and $p \in \mathcal{P}$:

(1) $\mathcal{M}, w \models \top$
(2) $\mathcal{M}, w \not\models \bot$
(3) $\mathcal{M}, w \models p$ iff $w \in V(p)$
(4) $\mathcal{M}, w \models \neg\phi$ iff $\mathcal{M}, w \not\models \phi$
(5) $\mathcal{M}, w \models \phi \vee \psi$ iff $\mathcal{M}, w \models \phi$ or $\mathcal{M}, w \models \psi$
(6) $\mathcal{M}, w \models \phi \wedge \psi$ iff $\mathcal{M}, w \models \phi$ and $\mathcal{M}, w \models \psi$
(7) $\mathcal{M}, w \models \phi \Rightarrow \psi$ iff $\mathcal{M}, w \models \neg\phi$ or $\mathcal{M}, w \models \psi$
(8) $\mathcal{M}, w \models B_i\phi$ iff $\forall v \in \mathcal{W}, w\mathcal{B}_i v : \mathcal{M}, v \models \phi$
(9) $\mathcal{M}, w \models K_i\phi$ iff $\forall v \in \mathcal{W}, w\mathcal{K}_i v : \mathcal{M}, v \models \phi$
(10) $\mathcal{M}, w \models E_i\phi$ iff $\forall v \in \mathcal{W}, w\mathcal{E}_i v : \mathcal{M}, v \models \phi$
(11) $\mathcal{M}, w \models E_i^d\phi$ iff $|\phi| \in \mathcal{E}_i^d(w)$,
with $|\phi| := \{v \in \mathcal{W} : \mathcal{M}, v \models \phi\}$

Let us remind that ϕ is valid in $\mathcal{M}$ (written $\mathcal{M} \models \phi$) if, and only if, for all worlds $w \in \mathcal{W}$, ϕ is satisfiable in w i.e. $\mathcal{M}, w \models \phi$ is true. A formula ϕ is valid in a frame C (written $\models_C \phi$ or $C \models \phi$) if, and only if, for all models $\mathcal{M}$ built on C, $\mathcal{M} \models \phi$. In this case ϕ is a tautology of C, written $\models_C \phi$.

For the modalities of knowledge and belief, we conventionally constrain our frame C so that, for any agent $i \in \mathcal{N}$, $\mathcal{K}_i$ is *reflexive*, *transitive* and *confluent*[2] and $\mathcal{B}_i$ is *serial*, *transitive* and *Euclidean*. The constraints and relations between these two modalities have already been well studied [38]. Thus, we first consider that an agent i believes what it knows, namely:

$$\forall w \in \mathcal{W} : \mathcal{B}_i(w) \subseteq \mathcal{K}_i(w) \quad (KB1)$$

If an agent believes something then it knows it believes it:

$$\forall w, u, v \in \mathcal{W} : w\mathcal{K}_i u \wedge u\mathcal{B}_i v \Rightarrow w\mathcal{B}_i v \quad (KB2)$$

In the same way, an agent knows what it does not believe:

$$\forall w, u, v \in \mathcal{W} : w\mathcal{K}_i u \wedge w\mathcal{B}_i v \Rightarrow u\mathcal{B}_i v \quad (KB3)$$

The constraint E1 expresses the fact that once the actions leading to ϕ are done by the agent i, then ϕ is true (axiom T):

$$\forall w \in \mathcal{W} : w\mathcal{E}_i w \quad (E1)$$

From this constraint, we immediately deduce that $\mathcal{E}_i$ is also *serial* and thus this system satisfies the property D. $\mathcal{E}_i$ is also *transitive* because when the actions of agent i lead to ϕ, these actions also lead to the fact that these actions are done properly. Moreover, if an agent i does not perform actions that lead to some consequences ϕ, then agent i indirectly performs actions that lead to not realize the actions that lead to ϕ. Thus the relation $\mathcal{E}_i$ is *Euclidean*.

The constraints for the deliberate effect modality is defined as follows: the main difference with the operator E_i is that an agent i cannot deliberately brings about a tautology (called nNEd).

$$\forall w \in \mathcal{W} : W \notin \mathcal{E}_i^d(w) \quad (nNEd)$$

In addition, when an agent i deliberately brings about a state of the world, then the agent i performs actions that lead to this state of the world. There is therefore a link between the deliberate effect modality and the consequences of actions, represented by the constraint (called EdE):

$$\forall w \in \mathcal{W} : S \in \mathcal{E}_i^d(w) \Rightarrow \mathcal{E}_i(w) \subseteq S \quad (EdE)$$

[1] In this article, deliberate effects may also be associated with the expressions "bring about something" or "see to it that".

[2] A binary relation $\mathcal{R}$ on $\mathcal{W}$ is *confluent* if, and only if, the following property is satisfied $\forall w, u, v \in \mathcal{W}, w\mathcal{R}u \wedge w\mathcal{R}v \to \exists z \in \mathcal{W} : u\mathcal{R}z \wedge v\mathcal{R}z$. Here we do not consider a S5 system with negative introspection but a S4.2 system. For details, the interesting reader may refer to [38] who gave arguments to support S4.2 rather than S5 for modeling knowledge.

When an agent i deliberately brings about ϕ while deliberately brings about ψ, then agent i deliberately brings about $\phi \wedge \psi$:

$$\forall w \in \mathcal{W} : S \in \mathcal{E}_i^d(w) \wedge T \in \mathcal{E}_i^d(w) \Longrightarrow S \cap T \in \mathcal{E}_i^d(w) \quad (CE)$$

However we do not consider the reciprocal. A deliberate effect when it concerns a whole is not equivalent to the sum of its parts. For instance, when you deliberately eat a cake with hazelnuts, you do not deliberately eat the cake's dough and you do not deliberately eat the cake's hazelnuts independently. Furthermore the reciprocal $\forall w \in \mathcal{W} : S \cap T \in \mathcal{E}_i^d(w) \Longrightarrow S \in \mathcal{E}_i^d(w) \wedge T \in \mathcal{E}_i^d(w)$ which is associated with $E_i^d(\phi \wedge \psi) \Rightarrow E_i^d\phi \wedge E_i^d\psi$ cannot be considered for a technical reason. An immediate result which comes from topology [25] says that it is also equivalent to $\forall w \in \mathcal{W} : S \in \mathcal{E}_i^d(w) \wedge S \subseteq T \Longrightarrow T \in \mathcal{E}_i^d(w)$ and it is inconsistent with the constraint $\forall w \in \mathcal{W} : W \notin \mathcal{E}_i^d(w)$ $(nNED)$.

Finally, the deliberate effect modality satisfies positive (EdKP) and negative (EdKN) introspection in relation to the modality of knowledge.

$$\forall w \in \mathcal{W} : \mathcal{E}_i^d(w) \subseteq \bigcap_{v \in \mathcal{W} : w\mathcal{K}_i v} \mathcal{E}_i^d(v) \quad (EdKP)$$

$$\forall w \in \mathcal{W} : \forall S \notin \mathcal{E}_i^d(w) \Longrightarrow S \notin \bigcup_{v \in \mathcal{W} : w\mathcal{K}_i v} \mathcal{E}_i^d(v) \quad (EdKN)$$

Respectively, these constraints mean that an agent who deliberately brings about a certain consequence, knows that he deliberately brings about it. If an agent i does not deliberately brings about a certain consequence, then agent i knows that it deliberately does not bring about this consequence.

3.3 Associated axiomatic system

Given the constraints on our framework, the associated axiomatic system is given in Figure 1: $\vdash \phi$ means that ϕ is a theorem. Moreover, for all modalities $\Box \in \{K_i, B_i, E_i, E_i^d\}$, we have the modus ponens (MP), the substitution (SUB) and the rule of inference (RE) i.e. from $\vdash \phi \Leftrightarrow \psi$, infer $\vdash \Box\phi \Leftrightarrow \Box\psi$. However, the rule of necessitation (NEC) is only verified for normal modalities i.e. for all $\Box \in \{K_i, B_i, E_i\}$, from $\vdash \phi$, infer $\vdash \Box\phi$. Finally, we have duality $(DUAL)$ i.e. for all $(\Box, \Diamond) \in \{(B_i, \langle B_i\rangle), (K_i, \langle K_i\rangle), (E_i, \langle E_i\rangle), (E_i^d, \langle E_i^d\rangle)\}$, $\vdash \Box\phi \Leftrightarrow \neg\Diamond\neg\phi$.

3.4 Soundness

It is well known that the semantics of a normal modality of a system S5 that preserves validity is an equivalence relation [6]. Since the relation $\mathcal{E}_i$ is an equivalence relation, the rules of S5 preserve the validity. Then a relation $\mathcal{K}_i$ which is *reflexive*, *transitive* and *confluent* is sound with a S4.2 system. Concerning the inference rules between the modality K_i and B_i, Stalnaker [38] showed they are valid in our logical frame. Moreover, it is well known that a serial, transitive and Euclidean relation preserves the validity of a KD45 system for the modality B_i. Thus, in this section, we only focus on the non-normal properties associated with the neighborhood semantics $\mathcal{E}_i^d$. The following properties are in [25]:

- (1) $\mathcal{C} \models \neg E_i^d \top$ iff $\forall w \in \mathcal{W} : \mathcal{W} \notin \mathcal{E}_i^d(w)$
- (2) $\mathcal{C} \models E_i^d p \wedge E_i^d q \Rightarrow E_i^d(p \wedge q)$ iff $\forall w \in \mathcal{W} : S \in \mathcal{E}_i^d(w) \wedge T \in \mathcal{E}_i^d(w) \Longrightarrow S \cap T \in \mathcal{E}_i^d(w)$

(PC)	All tautologies of Propositional Calculus
$(\mathbf{S4}_{K_i})$	All S4-axioms for K_i
$(\mathbf{4.2}_{K_i})$	$\vdash \langle K_i\rangle K_i\phi \Rightarrow K_i\langle K_i\rangle\phi$
$(\mathbf{KD45}_{B_i})$	All KD45-axioms for B_i
$(\mathbf{S5}_{E_i})$	All S5-axioms for E_i
$(\mathbf{K_iB_i})$	$\vdash K_i\phi \Rightarrow B_i\phi$
$(\mathbf{4}_{K_i,B_i})$	$\vdash B_i\phi \Rightarrow K_iB_i\phi$
$(\mathbf{5}_{K_i,B_i})$	$\vdash \neg B_i\phi \Rightarrow K_i\neg B_i\phi$
$(\mathbf{E_i^dE_i})$	$\vdash E_i^d\phi \Rightarrow E_i\phi$
$(\mathbf{C}_{E_i^d})$	$\vdash E_i^d\phi \wedge E_i^d\psi \Rightarrow E_i^d(\phi \wedge \psi)$
$(\neg\mathbf{N}_{E_i^d})$	$\vdash \neg E_i^d \top$
$(\mathbf{4}_{K_i,E_i^d})$	$\vdash E_i^d\phi \Rightarrow K_iE_i^d\phi$
$(\mathbf{5}_{K_i,E_i^d})$	$\vdash \neg E_i^d\phi \Rightarrow K_i\neg E_i^d\phi$

Figure 1: Axiomatic system KBE

Other properties are standard to prove by using contraposition and building a right countermodel. Consequently it is straightforward to prove that our KBE system is sound.

Theorem 3.1. *The KBE system is sound.*

Proof. Substitution, modus ponens, and necessitation preserve the validity for any normal modality [6] and for E_i^d: (PC), (SUB), (MP), (RE), $(DUAL)$ also preserve validity [25]. So KBE is sound. □

3.5 KBE Completeness

In order to prove that our system is complete, we apply a Henkin-like proof method by building a canonical model which relies on *Maximal Consistent Sets* (MCS) and a notion of *minimal canonical model* for neighborhood semantics [25].

Theorem 3.2. *The KBE system is complete.*

Proof. Due to space restrictions and because it is standard, we only present a sketch of the proof. The proof is based on the Henkin-like proof method and the sets of MCS. We consider a minimal canonical model $\mathcal{M}^c$ which is a model s.t. for all $i \in \mathcal{N}$, and $w \in \mathcal{W}^c$ with $\mathcal{W}^c$ a set of MCS, for each $(\mathcal{R}^c, \Box) \in \{(\mathcal{K}_i^c, K_i), (\mathcal{B}_i^c, B_i), (\mathcal{E}_i^c, E_i)\}$, $w\mathcal{R}^c v$ iff $\Box\phi \in w \Longrightarrow \phi \in v$, and $\mathcal{E}_i^{d^c}(w) := \{||\phi|| : E_i\phi \in w\}$ with $||\phi|| := \{w \in \mathcal{W}^c : \phi \in w\}$ and $V^c(p) = ||p||$ with p a propositional letter. We firstly prove by induction that it satisfies the truth lemma [25] and secondly we prove that all frame properties hold. Particularly $\mathcal{K}_i$, $\mathcal{B}_i$ and $\mathcal{E}_i$ are standard [6]. However, E_i^d may raise an issue due to the property $(EdKN)$. Let us notice that, when $X \notin \mathcal{E}_i^{d^c}(w)$ and $w\mathcal{K}_i^c v$ with $w, v \in \mathcal{W}^c$, we must consider two cases. If X is in the form $X = ||\phi||$, then $\neg E_i^d\phi \in w$. Furthermore as $\vdash \neg E_i^d\phi \Rightarrow K_i\neg E_i^d\phi$, we have $\neg E_i^d\phi \Rightarrow K_i\neg E_i^d\phi \in w$ (cf. truth lemma) and by a MCS property on implication, we have immediately $K_i\neg E_i^d\phi \in w$. Finally, with the definition of the canonical model and since $w\mathcal{K}_i v$, we have $X \notin \mathcal{E}_i^{d^c}(v)$. However if ϕ is not in the form $X = ||\phi||$, then by definition of the *minimal canonical model*, we deduce that for all $u \in \mathcal{W}^c$ such that there is no $\phi \in \mathcal{L}_{KBE}, u = ||\phi||$ then $X \notin \mathcal{E}_i^{d^c}(u)$. So $X \notin \mathcal{E}_i^{d^c}(v)$.

In conclusion, we prove that for all valid formulas ϕ in our frame C, ϕ is valid for all models $\mathcal{M}$ on C. Thus, ϕ is valid in all canonical model $\mathcal{M}^c$ and so $\vdash \phi$. Consequently, our KBE system is complete. □

Moreover it is easy to show that the KBE system has the *deduction theorem*, is also *strongly complete* and *strongly sound* [6, 25].

3.6 Some frame properties

Let us remark that the property D holds for E_i^d. Other interesting theorems can also be deduced. In particular, when an agent ensures that another agent believes something, then it ensures that the other agent does not believe that a third-party agent can know the opposite.

Theorem 3.3.

(1) $\vdash \neg E_i^d \bot \quad (D_{E_i^d})$

(2) *concealing contrary beliefs:* $\vdash E_i^d B_j \phi \Rightarrow E_i \neg B_j K_k \neg \phi$

(3) *concealing knowledge:* $\vdash E_i^d \neg B_j \phi \Rightarrow E_i \neg B_j K_k \phi$

Proof. All these theorems can be obtained by our Hilbert proof system. We only give sketches about the rules to apply.

(1) Just notice that $\vdash \neg E_i \bot \Rightarrow \neg E_i^d \bot$.

(2) The first step is to show that $\vdash (B_j\phi \Rightarrow \neg B_j K_k \neg \phi)$. This theorem directly comes from reductio ad absurdum on $B_j\phi \wedge B_j K_k \neg \phi$ and by using (T_{K_k}). The second step is to apply necessitation $\vdash E_i(B_j\phi \Rightarrow \neg B_j K_k \neg \phi)$ and with $(K_{E_i}) + (MP)$, immediately we have $\vdash E_i B_j \phi \Rightarrow E_i \neg B_j K_k \neg \phi$. Finally with $(E_i^d E_i) + (MP)$, we deduce $\vdash E_i^d B_j \phi \Rightarrow E_i \neg B_j K_k \neg \phi$.

(3) Firstly with (NEC_{B_j}) on (T_{K_k}), we have $\vdash B_j K_k \phi \Rightarrow B_j \phi$. Secondly by applying (NEC_{E_i}) on the contraposition of this theorem, we obtain $\vdash E_i \neg B_j \phi \Rightarrow E_i \neg B_j K_k \phi$. Finally we prove the theorem $\vdash E_i^d \neg B_j \phi \Rightarrow E_i \neg B_j K_k \phi$ with $(E_i^d E_i) + (MP)$. □

These theorems tell us that when an agent i brings it about new beliefs in another agent j they also maintained consistency (i.e. by preventing j to know that a third party agent may know the opposite as it is the case for i). Let us notice that $\vdash E_i^d \neg K_j \phi \Rightarrow E_i \neg K_j K_k \phi$ is also a theorem by following the same method as in (3). Moreover, as the contraposition of $\vdash K_k \phi \Rightarrow B_k \phi$ is $\vdash \neg B_k \phi \Rightarrow \neg K_k \phi$ and $\vdash E_i^d \phi \Rightarrow E_i \phi$ is $\vdash \neg E_i \phi \Rightarrow \neg E_i^d \phi$, we deduce two immediate corollaries to these theorems:

(1) $\vdash E_i^d B_j \phi \Rightarrow \neg E_i^d K_j K_k \neg \phi$

(2) $\vdash E_i^d \neg B_j \phi \Rightarrow \neg E_i^d K_j K_k \phi$

We also prove a *qui facit per alium facit per se* principle i.e. "he who acts through another does the act himself".

Theorem 3.4. *(Qui facit per alium facit per se)*

$$\vdash (E_i^d E_j \phi \vee E_i^d E_j^d \phi) \Rightarrow E_i \phi$$

Proof. For the left part of the disjunction with $(E_i^d E_i)$ and (NEC_{E_i}) on (T_{E_j}), we immediatly have $\vdash E_i^d E_j \phi \Rightarrow E_i \phi$. For the right part of the disjunction with $(E_i^d E_i)$, we have by substitution $\vdash E_i^d E_j \phi \Rightarrow E_i E_j \phi$. Then with (NEC_{E_i}) on (T_{E_j}), $\vdash E_i^d E_j^d \phi \Rightarrow E_i \phi$. Consequently by disjunction elimination we deduce the theorem $\vdash (E_i^d E_j \phi \vee E_i^d E_j^d \phi) \Rightarrow E_i \phi$. □

This theorem means that an agent influencing another agent to act illegally also acts itself. Thus in a legal context, a manipulative agent is also responsible for illegal acts perpetrated by a manipulated agent. Therefore the manipulator has some responsibility in these acts committed by this principle.

4 MODELING MANIPULATIONS

In this section we first define formally what a manipulation is. Secondly we show our logical framework can model coercion, persuasion and some form of deception, and thus is consistent with "manipulation is not exactly coercion, not precisely persuasion, and not entirely similar to deception" [15].

4.1 What manipulation is

In terms of manipulation, a manipulator always intended to influence the intentions: by pushing his victim to do something, or by preventing his victim from doing something. We call this influence an *instrumentalization*. Moreover, the manipulator always deliberately intended to conceal this instrumentalization. Thus, we can characterize manipulation depending on (1) what the manipulator wanted the victim to realize ; (2) whether the victim deliberately intended to realize the manipulator's will; (3) how the manipulator intended to conceal. Hence, we consider *constructive manipulations* when a manipulator brings his victim about doing something, and *destructive manipulations* when the manipulator aims at preventing an agent from doing something. Since manipulation is a deliberate effect of the manipulator, we also need to distinguish between bringing another agent about doing something in a deliberate way from doing it in an unintentional way. Thus, a *strong manipulation* is when the manipulator deliberately brings the manipulated agent about deliberately doing something, and a *soft manipulation* when the manipulator brings the manipulated agent about doing something. Finally we distinguish different forms of manipulation depending on whether the dissimulation is based on knowledge or beliefs: we call an *epistemic concealment* when the manipulator aims at preventing the victim to know his effects, and a *doxastic concealment* when the manipulator aims at preventing the victim to believe his effects.

The Table 1 and the Table 2 present different ways of expressing instrumentalization and concealment in the case of constructive manipulations and the case of destructive manipulations.

Instrumentalization	**Concealment**
Strong $(E_i^d E_j^d \phi)$	Epistemic $(E_i^d \neg K_j E_i^d E_j^d \phi)$
Soft $(E_i^d E_j \phi)$	Doxastic $(E_i^d \neg B_j E_i^d E_j \phi)$

Table 1: Constructive forms of manipulation

The Table 1 shows the different components of a constructive manipulation. For example, a *strong instrumentalization* is represented by the formula $E_i^d E_j^d \phi$. Literally, this formula describes that

the agent i employed a strategy leading to the agent j performing deliberately a set of actions which lead to the consequence ϕ. A *soft instrumentalization* can be represented by the formula $E_i^d E_j \phi$. Finally, in the case of constructive manipulations, an *epistemic concealment* can be represented by the formula $E_i^d \neg K_j E_i^d E_j \phi$ and a *doxastic concealment* by the formula $E_i^d \neg B_j E_i^d E_j \phi$.

Instrumentalization	Concealment
Strong ($E_i^d \neg E_j^d \phi$)	Epistemic ($E_i^d \neg K_j E_i^d \neg E_j^d \phi$)
Soft ($E_i^d \neg E_j \phi$)	Doxastic ($E_i^d \neg B_j E_i^d \neg E_j \phi$)

Table 2: Destructive forms of manipulation

The Table 2 describes the different components when a manipulation is destructive. For example, in this case of destructive manipulations, *soft instrumentalization* is represented by the formula $E_i^d \neg E_j \phi$, a *strong manipulation* is represented by the formula $E_i^d \neg E_j^d \phi$, then an *epistemic concealment* by the formula $E_i^d \neg K_j E_i^d \neg E_j^d \phi$ and finally, a *doxastic concealment* is represented by the formula $E_i^d \neg B_j E_i^d \neg E_j^d \phi$.

In the sequel, we combine these different forms of instrumentalization and concealment to define all the forms of manipulation that can be expressed in KBE. However since we use a non-normal modality for E_i^d which does not have the theorem $\Box(\phi \wedge \psi) \equiv \Box\phi \wedge \Box\psi$ with $\Box$ a normal modality, we have to consider all other possible formulas that this agent may deliberately brings about at the same time. Thus, we introduce a set of formulas Σ which is finite and *closed*[3]. Intuitively, this set represents the formulas on which agents can reason.

4.1.1 Soft constructive manipulations. A *soft constructive manipulation with epistemic concealment* – denoted $MCEK_{i,j}^{\Sigma}\phi$ below – is when a manipulator deliberately brings the victim about doing something (in a deliberate way or not) while making sure that the victim does not know the deliberate effects of the manipulator. We assume Σ represents all the formulas on which agents can reason. Thus, Σ is finite and closed. Furthermore, Σ contains $\{\top, \bot\}$ and $\phi \in \Sigma$. A *soft constructive manipulation with doxastic concealment* – denoted $MCEB_{i,j}^{\Sigma}\phi$ below – is similar but, in this case, the manipulator makes sure that the victim does not believe his deliberate effects. Formally, we define these manipulation forms such as:

$$MCEK_{i,j}^{\Sigma}\phi = \bigvee_{\psi \in \Sigma} E_i^d(E_j\phi \wedge \neg K_j E_i^d E_j \phi \wedge \psi)$$
$$MCEB_{i,j}^{\Sigma}\phi = \bigvee_{\psi \in \Sigma} E_i^d(E_j\phi \wedge \neg B_j E_i^d E_j \phi \wedge \psi)$$

Let us notice that, ψ represents all formulas from Σ which do not contradict $E_j\phi \wedge \neg K_j E_i^d E_j \phi$. Indeed, if ψ contradicts $E_j\phi \wedge \neg K_j E_i^d E_j \phi$, then we immediately deduce that $E_i^d \bot$. However it is necessarily false due to the theorem $\vdash \neg E_i^d \bot$. Moreover let us notice that Σ finds its analogy with the awareness correspondence function $\mathcal{A}_i : \mathcal{W} \to 2^{\mathcal{L}_{KBE}}$ as it is introduced in [35]. Here, we assume that all agents are aware of all formulas of Σ and therefore for all worlds $w \in \mathcal{W}$, for all $i \in \mathcal{N}$, we would have $\mathcal{A}_i(w) = \Sigma$. A remarkable point of this definition is that we can prove the existence of a manipulation only in relation to what we are aware of. Thus, even if Σ is a closed set of formulas on which agents can reason and we show that it is not the case that an agent i manipulates another agent j, we are never sure unless we assume that no agent considers formulas that are not in Σ.

For instance in advertising, the effects of an advertiser is to lead potential buyers buying a product (i.e. $E_i^d E_j \phi$). In general, this E_i^d is not hidden, and so it is not a manipulation but only influence. However, it becomes a manipulation when the advertiser uses a selling technique that he tries to hide from future buyers as subliminal images. Thus, in this case the advertiser seeks to conceal his real strategy (his E_i^d) to make the customer buying the product (i.e. $E_i^d \neg K_j E_i^d E_j \phi$).

4.1.2 Strong constructive manipulations. A *strong constructive manipulation with epistemic concealment* – denoted $MCE^d K_{i,j}^{\Sigma}$ below – is when the manipulator agent brings the other agent about doing something in a deliberate way while making sure that the victim does not know the deliberate effects of the manipulator. A *strong constructive manipulation with doxastic concealment* – denoted $MCE^d B_{i,j}^{\Sigma}$ below – is similar but, in this case, the manipulator makes sure that the victim does not believe his effects. Formally,

$$MCE^d K_{i,j}^{\Sigma}\phi = \bigvee_{\psi \in \Sigma} E_i^d(E_j^d\phi \wedge \neg K_j E_i^d E_j^d \phi \wedge \psi)$$
$$MCE^d B_{i,j}^{\Sigma}\phi = \bigvee_{\psi \in \Sigma} E_i^d(E_j^d\phi \wedge \neg B_j E_i^d E_j^d \phi \wedge \psi)$$

This manipulation may represent influences in a voting process for instance. The manipulator agent pushes the agents to vote for a given option ϕ and so influences the strategy of other agents to choose this option (i.e. $E_i^d E_j^d \phi$) while concealing his deliberate effects to influence their choice (i.e. $E_i^d \neg K_j E_i^d E_j^d \phi$).

4.1.3 Strong and soft destructive manipulations. As said in the introduction, another way to see manipulation is to consider that a manipulator may deliberately prevent the victim from doing something. We call this kind of manipulation a *destructive manipulation*. As previous, destructive manipulation can be declined in soft and strong destructive manipulations with either epistemic, or doxastic concealment. Formally,

$$MDEK_{i,j}^{\Sigma}\phi = \bigvee_{\psi \in \Sigma} E_i^d(\neg E_j\phi \wedge \neg K_j E_i^d \neg E_j \phi \wedge \psi)$$
$$MDE^d K_{i,j}^{\Sigma}\phi = \bigvee_{\psi \in \Sigma} E_i^d(\neg E_j^d\phi \wedge \neg K_j E_i^d \neg E_j^d \phi \wedge \psi)$$
$$MDEB_{i,j}^{\Sigma}\phi = \bigvee_{\psi \in \Sigma} E_i^d(\neg E_j\phi \wedge \neg B_j E_i^d \neg E_j \phi \wedge \psi)$$
$$MDE^d B_{i,j}^{\Sigma}\phi = \bigvee_{\psi \in \Sigma} E_i^d(\neg E_j^d\phi \wedge \neg B_j E_i^d \neg E_j^d \phi \wedge \psi)$$

An example of destructive manipulation is the case of eclipse attacks [37] in P2P networks. Such attacks consist in cutting off messages from and towards a node in order to exclude it from the

[3]We recall that a set of formulas Σ is said to be *closed* iff (1) if $\sigma \in \Sigma$ and θ is a subformula of σ, then $\theta \in \Sigma$ and (2) if $\sigma \in \Sigma$ and σ is not of the form $\neg\theta$, then $\neg\theta \in \Sigma$.

network. The hacker ensures, at the moment he acts, that the target node cannot communicate with other nodes in the network while not believing[4] it is currently under attack. Thus this attack can be described by a formula $E_i^d(\neg E_j\phi \wedge \neg B_j E_i^d \neg E_j\phi)$ with ϕ being any communication.

4.1.4 A general definition of manipulation. Finally, all these definitions can be merged in a general definition of manipulation:

$$M_{i,j}^{\Sigma}\phi = \bigvee_{\Box \in \{B,K\}} MCE\Box_{i,j}^{\Sigma}\phi \vee MCE^d\Box_{i,j}^{\Sigma}\phi \vee MDE\Box_{i,j}^{\Sigma}\phi \vee MDE^d\Box_{i,j}^{\Sigma}\phi$$

4.2 What manipulation is not

We can also express related notions like coercion, persuasion and deception which, as shown in Section 1, are different from manipulation. In the following, we consider a set of formulas Σ on which agents can reason and such that Σ is finite, closed and $\{\top, \bot\} \subseteq \Sigma$. Let $\phi \in \Sigma$ be a formula of Σ.

4.2.1 Coercion. The coercion is an influence of an agent over another agent by means of pressure without any dissimulation.

$$coe_{i,j}^{\Sigma}\phi = \bigvee_{\psi \in \Sigma} E_i^d(E_j^d\phi \wedge K_j E_i^d E_j^d \phi \wedge \psi)$$

A robber pointing a gun at somebody so as to get his wallet is not trying to manipulate the victim but he is influencing his behavior. The robber deliberately ensures that the victim knows that he is under pressure (by pointing the gun).

4.2.2 Persuasion. Persuasion consists in an agent making another one into believing something.

$$per_{i,j}^{\Sigma}\phi = \bigvee_{\psi \in \Sigma} E_i^d(B_j\phi \wedge \psi)$$

Interestingly, if ψ represents a form of dissimulation, then we talk about deception.

4.2.3 Deception. Deception consists in an agent making another believing something while hiding it some aspects linked to the newly believed statement. It may be half-truth, or deception by omission [32]. Due to space constraints, we only focus on *source concealment* (namely hiding the deliberate effects to make another agent into believing something) and *credible lies* (namely hiding we believe the opposite of the statement we want the other agent to believe).

Source concealment can represent agents that spread rumors. For instance, in the case of stock exchange market, it happens that some agents spread rumors in order to influence the others to buy or sell a product without they know that it is a part of their strategy [1]. Thus, it can be characterized by the fact that an agent makes sure to conceal his deliberate effects to make someone believes something.

$$con_{i,j}^{\Sigma}\phi = \bigvee_{\psi \in \Sigma} E_i^d(B_j\phi \wedge \neg K_j E_i^d B_j\phi \wedge \psi)$$

While Mahon [23] defines lying as "to make a believed-false statement (to another person) with the intention that that statement be believed to be true (by the other person)", a statement is a lie if there is also a deliberate effect to conceal the intended effects to lie. We call such a lie a *credible lie*:

$$cre_{i,j}^{\Sigma}\phi = B_i\neg\phi \wedge (\bigvee_{\psi \in \Sigma} E_i^d(B_j\phi \wedge \neg K_j B_i \neg\phi \wedge \psi)))$$

Let us notice that we do not need to introduce explicit communication modality as [32] did. Lying can be expressed as a deliberate effect of an agent i to make another agent j believing something while the agent i believes in the opposite. Here, the modality of communication could be reduced to a deliberate effect of making another agent believing something.

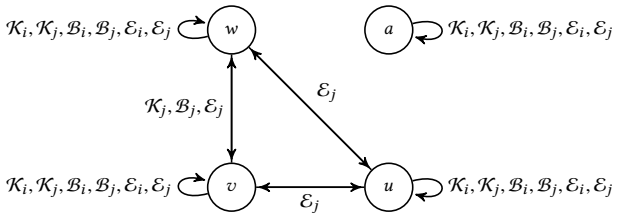

Figure 2: Mental states of the agents

5 APPLICATION OF KBE

The purpose of the example below is to instantiate the KBE system in a situation where it is possible for one agent to manipulate another. We consider an ecommerce website in which two agents perform a commercial transaction. Let i be the seller and j be the customer. The agent i says to the agent j: "You can trust me on the quality of the product. You will not find a better product anywhere else. You are free to check information by yourself !". Let us notice in the conversation when you use terms such as you "are free to" may be related to a technique of manipulation. For instance, these terms are the basis of a technique in the theory of free will compliance[5]. To represent this situation, we consider two propositional variables p and q:

p refers to "agent j trusts the agent i on product quality";
q refers to "agent j buys the product".

We consider several possible future scenarios[6] and represent them as a set of possible worlds $\mathcal{W} = \{a, w, v, u\}$ where:

w: "agent i builds trust to get agent j to buy the product";
v: "agent i does not deliberately influence j to buy the product but j buys the product and trusts i on product quality";
u: "agent j buys the product without trust in i on product quality and knows that the agent i intended to make him buy the product";
a: "agent j does not buy the product and does not trust i on product quality".

Let $\Sigma = Cl(\Gamma)$ be a set of formulas where $\Gamma = \{E_i^d(E_i^d p \Rightarrow E_i^d E_j q), E_i^d E_j q \wedge E_i^d \neg K_j E_i^d E_j q \Rightarrow E_i^d(E_j q \wedge \neg K_j E_i^d E_j q), E_i^d E_j^d q \wedge E_i^d K_j E_i^d E_j^d q \Rightarrow E_i^d(E_j^d q \wedge K_j E_i^d E_j^d q)\} \cup \{\top, \bot\}$. Σ is defined as the

[4]Here, the attack is viewed as a soft destructive manipulation with doxastic concealement. Obviously, it may also be defined with an epistemic concealment.

[5]It has been observed by sociopsychologists that the use of terms such as "you are free to" can strongly influence the choice of somebody to which desired by a manipulator [20].

[6]For the sake of readability, we do not consider all other possible scenarios such as the agent j does not buy the product but trusts i on product quality.

closure[7] of Γ, and is finite and closed. We assume that Σ represents all the formulas on which the agents can reason with. For example, the formula $E_i^d(E_i^d p \Rightarrow E_i^d E_j q)$ allows agents to reason about the situation described in the world w i.e. the agent i deliberately builds trust in order to get agent j to buy the product. Furthermore, as we consider the closure of Γ, we also express all subformulas and their single negation[8]. Then, the formula $E_i^d E_j q \wedge E_i^d \neg K_j E_i^d E_j q \Rightarrow E_i^d(E_j q \wedge \neg K_j E_i^d E_j q)$ allows to deduce a *soft constructive manipulation*. Finally, $E_i^d E_j^d q \wedge E_i^d K_j E_i^d E_j^d q \Rightarrow E_i^d(E_j^d q \wedge K_j E_i^d E_j^d q)$ allows agents to infer coercion.

The valuation function V of the model describing this situation is given by $V(p) = \{w, v\}$ and $V(q) = \{w, u, v\}$. The accessibility relations are given in Figure 2 and they are assumed to be:

(1) $\mathcal{K}_i(w) = \{w\}, \mathcal{K}_i(v) = \{v\}, \mathcal{K}_i(u) = \{u\}, \mathcal{K}_i(a) = \{a\}$
(2) $\mathcal{K}_j(w) = \{w, v\}, \mathcal{K}_j(v) = \{w, v\}, \mathcal{K}_j(u) = \{u\}, \mathcal{K}_j(a) = \{a\}$
(3) $\mathcal{B}_i(w) = \{w\}, \mathcal{B}_i(v) = \{v\}, \mathcal{B}_i(u) = \{u\}, \mathcal{B}_i(a) = \{a\}$
(4) $\mathcal{B}_j(w) = \{w, v\}, \mathcal{B}_j(v) = \{w, v\}, \mathcal{B}_j(u) = \{u\}, \mathcal{B}_j(a) = \{a\}$
(5) $\mathcal{E}_i(w) = \{w\}, \mathcal{E}_i(v) = \{v\}, \mathcal{E}_i(u) = \{u\}, \mathcal{E}_i(a) = \{a\}$
(6) $\mathcal{E}_j(w) = \{w, u, v\}, \mathcal{E}_j(v) = \{w, u, v\}, \mathcal{E}_j(u) = \{w, u, v\}, \mathcal{E}_j(a) = \{a\}$
(7) $\mathcal{E}_i^d(w) = \{\{w, v\}, \{w, u, v\}, \{w, u, a\}, \{w, v, a\}, \{w\}, \{w, a\}, \{w, u\}\}, \mathcal{E}_i^d(v) = \{\{v\}, \{w, v\}\}, \mathcal{E}_i^d(u) = \{\{u\}, \{w, u, v\}\}, \mathcal{E}_i^d(a) = \{\{w, a\}\}$
(8) $\mathcal{E}_j^d(w) = \{\{w, u, v\}\}, \mathcal{E}_j^d(v) = \{\{w, u, v\}\}, \mathcal{E}_j^d(u) = \{\{w, u, v\}\}, \mathcal{E}_j^d(a) = \{\{w, a\}\}$

(1) and (2) describe the fact that the agent i knows if the agent j trusts him and if the agent j buys the product. Moreover, (3) and (4) require that agents believe what they know and vice versa, i.e. $\mathcal{K}_i = \mathcal{B}_i$ and $\mathcal{K}_j = \mathcal{B}_j$.

(5) In the possible world w, the agent i ensures that p and q. (6) The agent j buys the product in $\{w, u, v\}$ but does not necessarily trust i on product quality.

(7) In w, the agent i deliberately ensures that the agent j trusts him and he deliberately ensures that if the agent j trusts him, then the agent i buys the product while making sure to hide his strategy to get him to buy the product. (8) Finally, the agent j in $\{w, u, v\}$ only intended to buy the product.

Let us notice that in w, this model expresses that the agent i has deliberately influenced the agent j to buy the product by building trust. Indeed, we have $|E_i^d p \Rightarrow E_i^d E_j q| = \{w, u, a\}$ and $\{w, u, a\} \in \mathcal{E}_i^d(w)$. So $\mathcal{M}, w \models E_i^d(E_i^d p \Rightarrow E_i^d E_j q)$. Thus, by applying the theorem $\models E_i^d \phi \Rightarrow \phi$, we deduce that in w, we have $\mathcal{M}, w \models E_i^d p \Rightarrow E_i^d E_j q$. But $|p| = \{w, v\}$ and $\{w, v\} \in \mathcal{E}_i^d(w)$, $\mathcal{M}, w \models E_i^d p$. Therefore, we have $\mathcal{M}, w \models E_i^d E_j q$.

In addition, we can notice that in w, the agent i also ensures to hide his strategy to get agent j to buy the product. Indeed, we have in v that, since $|E_j q| = \{w, u, v\}$ and $\{w, u, v\} \notin \mathcal{E}_i^d(v)$, we have $\mathcal{M}, v \models \neg E_i^d E_j q$. Thus, since the agent j cannot discern between the worlds w and v, we also deduce that $\mathcal{M}, w, v \models \neg K_j E_i^d E_j q$. Moreover, we can notice that $|\neg K_j E_i^d E_j q| = \{w, v, a\}$[9] and $\{w, v, a\} \in \mathcal{E}_i^d(w)$. Therefore, since $|\neg K_j E_i^d E_j q| \in \mathcal{E}_i^d(w)$, we have $\mathcal{M}, w \models E_i^d \neg K_j E_i^d E_j q$.

In conclusion, we have shown that $\mathcal{M}, w \models E_i^d E_j q \wedge E_i^d \neg K_j E_i^d E_j q$. Now, by the tautology $\models E_i^d \phi \wedge E_i^d \psi \Rightarrow E_i^d(\phi \wedge \psi)$, we deduce that $\mathcal{M}, w \models E_i^d(E_j q \wedge \neg K_j E_i^d E_j q)$ and it is equivalent to $\mathcal{M}, w \models E_i^d(E_j q \wedge \neg K_j E_i^d E_j q \wedge \top)$. So, we showed that, in this situation, there is a possible world in which the agent i is manipulating the agent j to make him buy the product by using a *soft constructive manipulation with epistemic concealment*. Moreover, if we decompose the deliberate effects of the agent i, $\mathcal{E}_i^d(w) = \{\{w, v\}, \{w, u, v\}, \{w, u, a\}, \{w, v, a\}, \{w\}, \{w, a\}, \{w, u\}\}$, we notice that the agent i intended to ensure p by considering the set $|p| = \{w, v\}$, and on the other hand to ensure q by considering the set $|q| = \{w, u, v\}$. This agent also has the strategy to get the other agent to buy the product with the set $|E_i^d p \Rightarrow E_i^d E_j q| = \{w, u, a\}$, and his strategy of dissimulation is represented by the set $|\neg K_j E_i^d E_j q| = \{w, v, a\}$. Finally, the sets $\{w\}$, $\{w, a\}$, $\{w, u\}$ are given by the imposed constraint (CE) on the frame to allow the tautology $\models E_i^d \phi \wedge E_i^d \psi \Rightarrow E_i^d(\phi \wedge \psi)$. These sets of possible worlds reflect the fact that when an agent sets up different planes, this agent also considers all combinations of all the different planes as a possible plane. For example, since $\{w, a\} = \{w, u, a\} \cap \{w, v, a\}$, $\{w, a\}$ is the combination of the respective planes $|E_i^d p \Rightarrow E_i^d E_j q|$ and $|\neg K_j E_i^d E_j q|$.

Finally let us notice that in the world u, the agent i coerced the agent j to push him to buy the product. Indeed, since we have $|E_j^d q| = \{w, u, v\}$ and $|K_j E_i^d E_j^d q| = \{u\}$[10] and that $|E_j^d q| \cap |K_j E_i^d E_j^d q| = \{u\} \in \mathcal{E}_i^d(u)$, we deduce that $\mathcal{M}, u \models E_i^d(E_j^d q \wedge K_j E_i^d E_j^d q)$ and so $\mathcal{M}, u \models E_i^d(E_j^d q \wedge K_j E_i^d E_j^d q \wedge \top)$. Consequently, we just have proved that $\mathcal{M}, u \models coe_{i,j}^{\Sigma} q$.

6 CONCLUSION AND FUTURE WORKS

In this article, we proposed a framework for expressing manipulation as the deliberate instrumentalization of a victim while making sure to conceal that instrumentalization. To this end, we proposed a new deliberate BIAT modality. Considering knowledge, belief, deliberate effects and consequences of actions as different modalities, we proved that our system was sound and complete. Furthermore it allowed us to deduce several theorems such as concealment of contrary beliefs and *qui facit per alium facit per se* principle. Finally we gave an explicit definition of what manipulation is, and we modeled coercion, persuasion and deception differently such as highlighted by the literature. In terms of perspectives, it should be interesting to extend the framework with dynamical aspects of awareness like it is described by Van Ditmarsch et al. [43] so as to define new manipulation forms with *awareness concealment*.

[7] Let Γ be a set of formulas. We recall that $Cl(\Gamma)$ is the *closure* of Γ iff $Cl(\Gamma)$ is the smallest closed set of formulas containing Γ.
[8] We recall that a set of formulas Σ is *closed under single negation* iff if $\sigma \in \Sigma$ and σ is not of the form $\neg\theta$, then $\neg\sigma \in \Sigma$.

[9] We explain why $a \in |\neg K_j E_i^d E_j q|$ and $u \notin |\neg K_j E_i^d E_j q|$. Firstly, notice that $|E_j q| \notin \mathcal{E}_i^d(a)$, and $\mathcal{M}, a \models \neg E_i^d E_j q$ and $\forall x \in \mathcal{W} : a \mathcal{K}_j x$, $\mathcal{M}, x \models \neg E_i^d E_j q$. Thus, $\mathcal{M}, a \models K_j \neg E_i^d E_j q$, and so $\mathcal{M}, a \models \neg K_j E_i^d E_j q$. Secondly, notice that $|E_i^d E_j q| = \{w, u\}$ and since $\forall x \in \mathcal{W}$, $u \mathcal{K}_j x$, $\mathcal{M}, x \models E_i^d E_j q$, we have $\mathcal{M}, u \models K_j E_i^d E_j q$ and so $u \notin |\neg K_j E_i^d E_j q|$.
[10] To make sure, just compute the set $|E_i^d E_j^d q| = \{w, u\}$ and so the only possible world x such that $\forall z \in \mathcal{W}$, $x \mathcal{K}_j z$, $\mathcal{M}, z \models E_i^d E_j^d q$ is the world $x = u$.

REFERENCES

[1] R. K. Aggarwal and G. Wu. 2006. Stock market manipulations. *The Journal of Business* 79, 4 (2006), 1915–1953.

[2] A. S. Akopova. 2013. Linguistic Manipulation: Definition and Types. *IJCRSEE* 1, 2 (2013), 78–82.

[3] R. Alur, T. A. Henzinger, and O. Kupferman. 2002. Alternating-time temporal logic. *Journal of the ACM (JACM)* 49, 5 (2002), 672–713.

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

[5] N. Belnap and M. Perloff. 1988. Seeing to it that: a canonical form for agentives. *Theoria* 54, 3 (1988), 175–199.

[6] P. Blackburn, M. De Rijke, and Y. Venema. 2002. *Modal Logic: Graph. Darst.* Vol. 53. Cambridge University Press.

[7] E. Bottazzi and N. Troquard. 2015. On Help and Interpersonal Control. In *The Cognitive Foundations of Group Attitudes and Social Interaction*. 1–23.

[8] Jan M Broersen. 2011. Making a start with the stit logic analysis of intentional action. *Journal of philosophical logic* 40, 4 (2011), 499–530.

[9] Harvey R St Clair. 1966. Manipulation. *Comprehensive psychiatry* 7, 4 (1966), 248–258.

[10] S. Cohen. 2017. Manipulation and Deception. *Australasian Journal of Philosophy* (2017), 1–15.

[11] D. Ettinger and P. Jehiel. 2010. A theory of deception. *Microeconomics* 2, 1 (2010), 1–20.

[12] A. Gibbard. 1973. Manipulation of voting schemes: a general result. *Econometrica* (1973), 587–601.

[13] L. Giordano, A. Martelli, and C. Schwind. 2000. Ramification and causality in a modal action logic. *Journal of logic and computation* 10, 5 (2000), 625–662.

[14] Robert E Goodin. 1980. Manipulatory politics. *The journal of Politics* (1980).

[15] Sapir Handelman. 2009. *Thought manipulation: the use and abuse of psychological trickery*.

[16] D. Harel, D. Kozen, and J. Tiuryn. 2001. Dynamic logic. In *Handbook of philosophical logic*. 99–217.

[17] B. Hill. 2010. Awareness dynamics. *Journal of Philosophical Logic* 39, 2 (2010), 113–137.

[18] Kevin Hoffman, David Zage, and Cristina Nita-Rotaru. 2009. A survey of attack and defense techniques for reputation systems. *Comput. Surveys* 42, 1 (2009), 1–17.

[19] A. Jøsang and J. Golbeck. 2009. Challenges for robust trust and reputation systems. In *Proceedings of the 5th International Workshop on Security and Trust Management*.

[20] Robert-Vincent Joule, Fabien Girandola, and Fran¸coise Bernard. 2007. How can people be induced to willingly change their behavior? The path from persuasive communication to binding communication. *Social and Personality Psychology Compass* 1, 1 (2007), 493–505.

[21] M. Kligman and C. M. Culver. 1992. An analysis of interpersonal manipulation. *The Journal of Medicine and Philosophy* 17, 2 (1992), 173–197.

[22] E. Lorini and G. Sartor. 2016. A STIT Logic for Reasoning About Social Influence. *Studia Logica* 104, 4 (2016), 773–812.

[23] James Edwin Mahon. 2008. Two definitions of lying. *International Journal of Applied Philosophy* 22, 2 (2008), 211–230.

[24] B. Mobasher, R. Burke, R. Bhaumik, and J. J. Sandvig. 2007. Attacks and remedies in collaborative recommendation. *IEEE Intelligent Systems* 22, 3 (2007).

[25] E. Pacuit. 2017. *Neighborhood semantics for modal logic*. Springer.

[26] D. C. Parkes and L. H. Ungar. 2000. Preventing strategic manipulation in iterative auctions: Proxy agents and price-adjustment. In *Association for the Advancement of Artificial Intelligence*. 82–89.

[27] I. Pörn. 2012. *Action theory and social science: Some formal models*. Springer.

[28] P. Resnick and R. Sami. 2008. Manipulation-resistant recommender systems through influence limits. *ACM SIGecom Exchanges* 7, 3 (2008), 10.

[29] M. S. Robinson. 1985. Collusion and the Choice of Auction. *The RAND Journal of Economics* (1985), 141–145.

[30] Y. Ruan and A. Durresi. 2016. A survey of trust management systems for online social communities – Trust modeling, trust inference and attacks. *Know ledge-Based Systems* 106 (2016), 150–163.

[31] J. Rudinow. 1978. Manipulation. *Ethics* 88, 4 (1978), 338–347.

[32] C. Sakama, M. Caminada, and A. Herzig. 2015. A formal account of dishonesty. *Logic Journal of the IGPL* 23, 2 (2015), 259–294.

[33] S. Sanghvi and D. Parkes. 2004. Hard-to-manipulate VCG-based auctions. *Harvard Univ., Cambridge, MA, USA, Tech. Rep* (2004).

[34] Filipe Santos and José Carmo. 1996. Indirect action, influence and responsibility. In *Deontic Logic, Agency and Normative Systems*. 194–215.

[35] B. Schipper. 2014. Awareness. *Handbook of Epistemic Logic* (2014), 77–146.

[36] K. Segerberg, J.-J. Meyer, and M. Kracht. 2009. The logic of action. (2009).

[37] A. Singh, T.-W. Ngan, P. Druschel, and D. Wallach. 2006. Eclipse attacks on overlay networks: Threats and defenses. In *IEEE 25th International Conference on Computer Communications*.

[38] R. Stalnaker. 2006. On logics of knowledge and belief. *Philosophical studies* 128, 1 (2006), 169–199.

[39] C. R. Sunstein. 2015. Fifty shades of manipulation. *Journal of Marketing Behavior* 213 (2015).

[40] P. Todd. 2013. Manipulation. *The international encyclopedia of ethics* (2013).

[41] N. Troquard. 2014. Reasoning about coalitional agency and ability in the logics of âĂIJbringing-it-aboutâĂİ. *International Conference on Autonomous Agents and Multiagent Systems* 28, 3 (2014), 381–407.

[42] Thibaut Vallée, Grégory Bonnet, and François Bourdon. 2014. Multi-armed bandit policies for reputation systems. In *International Conference on Practical Applications of Agents and Multi-Agent Systems*. Springer, 279–290.

[43] H. Van Ditmarsch, T. French, F. R Velázquez-Quesada, and Y. N. Wáng. 2018. Implicit, explicit and speculative knowledge. *Artificial Intelligence* 256 (2018), 35–67.

[44] Hans Van Ditmarsch, Wiebe van Der Hoek, and Barteld Kooi. 2007. *Dynamic epistemic logic*. Vol. 337. Springer Science & Business Media.

[45] H. Van Ditmarsch, J. Van Eijck, F. Sietsma, and Y. Wang. 2012. On the logic of lying. In *Games, Actions and Social Software*. 41–72.

[46] A. R. Wagner and R. C. Arkin. 2009. Robot deception: recognizing when a robot should deceive. In *CIRA*. IEEE, 46–54.