# Reasoning about manipulation in multi-agent systems

Christopher Leturc, Grégory Bonnet

# Reasoning about Manipulation in Multi-Agent Systems

Christopher Leturc[a] and Grégory Bonnet[b]

[a]Inria, Université Côte d'Azur, CNRS, I3S, France
[b]Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC, 14000 Caen

**ARTICLE HISTORY**
Compiled October 21, 2022

**ABSTRACT**
In many applications, selfish, dishonest or malicious agents may find an interest in manipulating others. While many works deal with designing robust systems or designing manipulative strategies, few works are interested in defining in a broad sense what is a manipulation and how we can reason with such a notion. In this article, based on a social science literature, we give at first a general definition of manipulation that can be applied in multi-agent systems. A manipulation is a deliberate effect of an agent (called a *manipulator*) to instrumentalize another agent (called a *victim*), while making sure to conceal that effect. Secondly, we show how this definition is related with different fields in computer science where the concept of manipulation is studied. Finally, we present a logical framework, called KBE, to express and reason about manipulations. Since manipulation relies on deliberate effects, KBE introduces a deliberate BIAT operator which abstracts deliberate consequences of actions. We prove that this logic is sound and complete and, we formally define manipulation. Furthermore, based on KBE, we express related notions such as coercion, persuasion, or deception and we show that these notions are different from manipulation.

**KEYWORDS**
Deception; Manipulation; Modal logics; Neighborhood semantics.

## 1. Introduction

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 (Hoffman, Zage, & Nita-Rotaru, 2009; Josang & Golbeck, 2009; Ruan & Durresi, 2016). However, agents may have interest in those systems in lying so as to mislead others, and push them to interact with some specific agents. Such behaviour is a 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 as in computational social choice (Gibbard, 1973; Parkes & Ungar, 2000; Robinson, 1985; Sanghvi & Parkes, 2004), game theory (Ettinger & Jehiel, 2010; Wagner & Arkin,

CONTACT Christopher Leturc Email: christopher.leturc@inria.fr

2009), or recommendation systems (Mobasher, Burke, Bhaumik, & Sandvig, 2007; Resnick & Sami, 2008). From a general perspective, there are many works on concepts closely related to manipulation (Masters, Smith, Sonenberg, & Kirley, 2021). For instance, in modal logic literature, some works have modeled social influence (Lorini & Sartor, 2016) and deception (Sakama, Caminada, & Herzig, 2015; Van Ditmarsch, Van Eijck, Sietsma, & Wang, 2012). Interestingly, social science deals with manipulation as a combination of many concepts, e.g. strategies, deception, obfuscation, intentionality (Akopova, 2013; Cohen, 2017; Handelman, 2009; Todd, 2013). A review of social science literature allows us to consider a general definition of manipulation, i.e. acting in such a way to produce a deliberate effect of influencing another agent while concealing this influence.

This article is an extended version of our previous work (Leturc & Bonnet, 2020). In this work, we defined a modal logic that allows one to reason about this concept of manipulation. To this end and to express deliberate effects, we have proposed a deliberate BIAT modality and combined it with a STIT-like modality to catch all consequences of actions and side-effects of actions. Concealment was expressed through epistemic and doxastic modalities. In the present article, we extend this previous work with two new contributions: (1) we provide a broad state-of-the-art about manipulation in social science, and about logical systems to express such notion; (2) we provide the complete proofs of all theorems given in Leturc and Bonnet (2020).

The remainder of this article is structured as follows. In Section 2, we present a review of the notion of manipulation drawn from social science literature, and we propose a general definition of manipulation. In Section 3, in view of the previous definition of manipulation, we survey available logical tools that are able to express it. In Section 4, we propose a logical framework and show that it is sound and complete. In Section 5, we formally define manipulation, and show our formal framework can also be used to model *coercion*, *persuasion* and *deception*. Finally, in Section 6 we instantiate an example.

## 2. Defining manipulation

The notion of manipulation is present in several fields of computer science such as in social choice theory (Gibbard, 1973), or in reputation systems (Vallée & Bonnet, 2015) where it is defined as a strategy allowing an agent to influence and control the individual (or collective) decision-making process of a group of agents using false information in order to push the latter to make a decision in favour of the manipulating agent. However, this is a restrictive notion of manipulation, as shown by works on social psychology, manipulative agent does not necessarily use the transmission of false information to manipulate. For example, the "foot in the door" consists in getting the victim to engage in a preparatory behavior to facilitate a decision in favor of the manipulator (Joule, Beauvois, & Deschamps, 2002). As another example, social proof consists in showing a victimized agent a set of behaviours that other agents exhibit in order to influence him or her to imitate those agents (Cialdini, 2012). All those manners – using false informations, behaviors or social proof – are different concrete strategies to manipulate. In order to model manipulation in a broader way, we need a sufficiently broad definition of manipulation and we propose to extract it through social science literature. For this reason, Section 2.1 presents a survey on manipulation from the point of view of these fields. This state-of-the-art then allows us to give a definition of the term in Section 2.2.

### *2.1. Manipulation in social science*

Social scientists often disagree on the definition of manipulation. For some of them, manipulation is considered to be the act of altering the judgment of individuals by depriving them of some of their deliberate choices (Kligman & Culver, 1992; Saint Clair, 1966; Sunstein, 2015). However as stated by (Handelman, 2009; Rudinow, 1978; Sorlin, 2017), such a definition would lead to consider rational persuasion, deception or coercion as manipulation whereas most people would distinguish them. Consequently, there are four ways to consider manipulation in social science literature: considering it as a "vague concept", as strategies to alter the others' judgment, as an exercise of power over others, and finally as the invisible exercise of power. We review each of these points-of-view and extract their main features to propose a definition of manipulation for modeling.

#### *2.1.1. Manipulation seen as a vague concept*

By reviewing several situations in which the term "manipulation" is frequently used but giving counterexamples to each of them, the philosopher Felicia Nimue Ackerman claims that it is impossible to characterize manipulation because it is a "combinatorially vague concept" (Ackerman, 1995). By "combinatorially vague concept", she means that there are a variety of conditions under which the term is frequently used but none of them are sufficient to discriminate with certainty between manipulative and non-manipulative situations. To support her claim, Ackerman made a broad review of those conditions, including for instance use of indirect, cunning and subtle means, falsification or omission of information, deception, pressure and unethical behaviors, presence of hidden motives, and so on. While Ackerman claims that we cannot know what combinations of these conditions actually constitute manipulation. We reject this point-of-view for at least two reasons: (1) considering manipulation as a vague concept makes impossible to give an explicit formal definition of manipulation and reasoning with it as proposed in this paper and, (2) it would mean that efforts to model manipulation are doomed to failure unless to consider machine learning techniques to classify the data space into categories such as "manipulation" or "not manipulation", which is currently not the aim of this paper.

#### *2.1.2. Manipulation seen as strategies to tamper the others' judgment*

For many social scientists, manipulation is characterized by the use of psychological methods that can alter the victim's judgement, making him vulnerable to the manipulator's influence and unable to perceive the manipulation. For this reason, manipulation can be defined as *strategies* to exploit the others' weaknesses: manipulation is any successful and intentional act of influencing an agent's beliefs or behavior, directly or indirectly, by causing changes in mental processes (Baron, 2003; Faden & Beauchamp, 1986; Kahneman, 2011; Mills, 1995; Rudinow, 1978; Wilkinson, 2013). Those changes can be:

(1) increasing or decreasing the available options to the victim;
(2) offering a reward or threatening punishment to the victim;
(3) directly influencing the victim's mental states.

This point-of-view focuses on the strategies behind a manipulation. Obviously, in the literature there exists many definitions of strategy and the reader may refer to Mintzberg's review (Mintzberg, 1987). A manipulation can be based on exploiting

| Basis of the strategy | Examples |
|---|---|
| Exploiting agents' weaknesses | Cognitive dissonance (Festinger, 1962), discursive deception (Van Dijk, 2006), free will compliance (Joule et al., 2002) |
| Deflecting norms | Reciprocity, social proof or social commitment (Cialdini, 2001) |
| Abusing trust | Aggressive marketing (Calo, 2013), authority abuse (Luhmann, 1979), self-disclosure (Aïmeur & Sahnoune, 2020) |
| Using rationality | Authority arguments Handelman (2009), nudges and discreet rewards (Wilkinson, 2013) |
| Relying on emotions | Coercion (Wood, 2014), emotional blackmail (Gunderson, 1984), seduction (Strauss, 2006) |
| Playing on knowledge / beliefs | Deception (Whiten & Byrne, 1988), hidding truth (Maillat & Oswald, 2009) or bullshitting (McGinn, 2014). |

**Table 1.** A taxonomy of manipulation strategies in social science

an agent weakness as a cognitive dissonance, deflecting norms as social commitment, abusing trust as betrayal, playing on the rational behavior of the victim, relying on emotions such as playing on fear with coercion, or a manipulation strategy can be based on knowledge or beliefs as lying, or *bullshitting*. Table 1 sums up and gives examples of those kind of manipulation strategies.

**Exploiting agent weaknesses**

The theory of free will compliance, introduced in social psychology by Joule et al. (2002), describes a set of behavioral manipulation techniques. Among them, baiting involves asking a victim for a simple service, such as asking the time. It turns out that after obtaining this first service, it is easier to ask for a more compelling service. Thus, a manipulator may have the deliberate intention to use one of these techniques to manipulate humans. Another kind of weakness are cognitive biases (Kahneman, 2011). An agent can exploit those cognitive bias in order to induce another agent to make commitments of which he is unable to see the importance.

**Deflecting norms**

A manipulator can deflect a social norm to his own benefit. In social psychology, presents a set of norms on which a manipulator can rely to push another agent to realize his desire. For example, the norm of *reciprocity* is an internalized norm that says to always reciprocate with agents who offer us a service. Thus, a manipulator may circumvent this norm by voluntarily offering a costless service or resource to a victim to push him to provide a costier service requested in return by the manipulator (Cialdini, 2001).

**Abusing trust**

According to Castelfranchi and Falcone (2010), trust is the backbone of society. It is also a tool of manipulation. Indeed, a manipulation strategy consists in building trust in order to better control the other agents of the system. In marketing, trust is a tool to get a customer to buy a product he does not need and without his awareness (Calo, 2013). For example, salespeople can announce a "scarcity" (Cialdini, 2001) to give a customer the illusion that a product is of quality because there is not much left in stock.

**Using rationality**

Rationality is a behaviour of agents who always want to maximize what they can hope to gain from a given situation. It can be used to influence decisions without the agent even being aware of this influence. Governments can use rationality as an argument to undermine political opponents. For instance, the 2003 Iraq war was presented as "rational warfare" to gain support (Handelman, 2009). Other examples of manipulations that can circumvent or subvert the victim's rational abilities are the use of nudges, or small discreet rewards to push an agent in the "good" direction (Wilkinson, 2013).

**Relying on emotions**

Manipulations between humans can have an emotional basis. A manipulative agent simulates certain emotional states, for example when a child cries to get a new toy from his parents. The child deliberately puts himself in this state to affect the parents' emotional state in order to get what he wants. In psychiatry, this emotionally based manipulation strategy is found in many examples such as when a person threatens to commit suicide. In this case, the threat aims at getting something from the others: a listening ear, a service, hoping to get back together with a person (Gunderson, 1984).

**Playing on knowledge and beliefs**

The linguist Eddo Rigotti presents a typology of the main processes that allow a manipulating agent to mislead another agent by playing on his beliefs and knowledge (Rigotti, 2005). For example, sophistry is a technique that consists in pushing another agent to deduce something false in order to exploit this error. Another example consists in hiding the truth of a proposition which could weaken or contradict the manipulator's discourse (Maillat & Oswald, 2009).

### *2.1.3. Manipulation seen as the exercise of power over others*

Beyond the strategies an agent can use to manipulate, for many researchers, manipulation is primarily characterized as a form of power exercised over others (Abell, 1977; Goodin, 1980; Kligman & Culver, 1992; Maoz, 1990; Todd, 2013).

**Manipulation is not simply a loss of autonomy.**

While manipulation is sometimes characterized as a loss of autonomy (Poulin, 2010; Raz, 1986) (e.g. when a parasite takes control of its host), the philospher Patrick Todd claims it is not enough to define manipulation because it is necessary to distinguish between manipulation as the manipulation of an object, and manipulation as a case in which agents act in a manipulative manner (Todd, 2013). However, loss of autonomy is necessary for the manipulation to take place.

**Manipulation is a deliberate intention.**

Many social scientists state that manipulation is primarily an intention to act on another agent with influence. This influence is said to be manipulative if it is deliberately and intentionally used by the manipulator while the influence is said to be not manipulative as long as it is sincere, i.e. in accordance with what the influencer takes to be true, relevant, and appropriate (Kligman & Culver, 1992; Noggle, 1996). Indeed, when an agent manipulates another, the act is knowingly deliberate. As social science rejects the concept of *unintentional manipulation*, using a mechanism of influence without knowing it, and therefore without having intended it, cannot be considered as a manipulation.

**Manipulation is the intention to change something.**

Kligman and Culver (1992) define manipulation as the manpilator's intention to change something in his environment. For example, a manipulator may deliberately withhold or selectively present certain information and omit others, exploiting the ignorance or beliefs of his victim in order to maintain control over his perceived options and direct him in the direction desired by the manipulator. For Maoz (1990), political manipulation is an attempt by one or more individuals to structure a collective decision as to maximize (resp. minimize) the chances of a favourable (resp. unfavourable) outcome. For Abell (1977), manipulation is a process where a manipulative agent $A$ control a manipulated agent B's preferences by reducing $B$'s understanding of the situation or $B$'s means of action. Interestingly, all those definitions are similar to those used in game theory and social choice theory. For example, manipulating a voting system consists in providing a false preference profile to ensure a result preferred to the one normally obtained with the true preference profile (Gärdenfors, 1976; Gibbard, 1973).

**Manipulation does not always go against the interests of the others.**

When one agent manipulates the other, he is doing it for his own interest, and this often goes against the interests of the others. That is why Goodin (1980) claims that manipulation is primarily a deceptive influence compelling one to act against one's will. In the same way, Barnhill (2014) states manipulation is also the intention of influencing certain traits or psychological dispositions in order to bring the victim into ideals of beliefs, desires or emotions in a way that is generally not in the victim's self-interest in the current context. However, all manipulations do not go against the self-interest of the manipulated agent. For example, the placebo effect that is sometimes used in medicine to make a patient believe that he will be cured with an ineffective drug is a manipulation in the patient's interest (Turner, Deyo, Loeser, Von Korff, & Fordyce, 1994). That is why some authors consider the existence of *benevolent manipulations* (Rosenberg & Pearlin, 1962) meaning somebody exhibits a manipulative behavior in order to push someone else to do something in the latter's interest. Obviously, it means that the manipulator has a representation (possibly impropered) of the "victim"'s interest.

**Manipulation is an instrumentalization of the others.**

Being benevolant or not, manipulation can be viewed in all cases as a psychosocial maneuver that uses aggression, coercion, intelligence, deception and trickery to influence someone in order to achieve a manipulator's desire (Bowers, 2003; Rigotti, 2005; Saint Clair, 1966). Rigotti (2005) said the victim of a manipulation pursues the aim of the manipulator in the illusion of pursuing his or her own goal. While Saint Clair (1966) states that there must be an incompatibility between the manipulator's desires and the manipulated's ones, Bowers (2003) claims that manipulation seeks to achieve a desired goal without regard to the interests or needs of the manipulated agent. In both cases, achieving the manipulator's desires is the fundamental point associated with manipulation. That is why social science considers the manipulation as a form of *intrumentalization*, namely using somebody as a mean to achieve a goal.

### *2.1.4. Manipulation seen as the invisible exercise of a power*

Interestingly, many researchers states manipulation is primarily an hidden influence (de Saussure & Schulz, 2005; Handelman, 2009; Van Dijk, 2006; Ware, 1981).

For instance in clinical settings, manipulation is then associated with a patient's efforts (e.g. somatic complaints, provocative actions or misleading messages, and self-destructive acts) to use covert means to gain control or support from significant persons (Gunderson, 1984). In linguistics, manipulation is a malevolant intention of the speaker with a character of hidden influence (Akopova, 2013; Rigotti, 2005): a manipulation is carried out when the addressee can no longer see the speaker's intentions behind what he or she is asserting. More generally, Handelman (2009) states the practical meaning of the manipulation is that the target is subject to a hidden influence and believes that his or her choices are made freely and independently. The manipulation is therefore intended to motivate the target to operate in a form that, under normal conditions, would likely cause him or her to resist or reject the interaction. That is why this influence must be above all indirect, invisible and secret in order to take place. While this secrecy can simply mean the victim cannot use his deliberative capabilities (Sunstein, 2015) to know or understand what is happening (Ware, 1981), awareness is also considered: manipulation can be viewed as a form of control in which the manipulated is not aware of what is happening or of the manipulator's strategy (de Saussure & Schulz, 2005; Van Dijk, 2006).

### *2.2. Towards a general definition of manipulation*

However, although we have highlighted significant differences among researchers, it appears that there is agreement among them on certain points. Hence, we use these commonalities to consider a general definition of manipulation. One of the main characteristics of manipulation is that it is, on the one hand, the exercise of power over others, and on the other hand, the exercise of a concealed power from the manipulated agents.

Firstly, manipulation is a power which is distinct from the more general concept of influence. While influence may be exerted in an unintentional way, manipulation is first and foremost a voluntary effect of a manipulator to use the victim to accomplish something (Akopova, 2013; Cohen, 2017; Handelman, 2009; Todd, 2013). Manipulation is necessarily intentional, and the use of the concept of "unintentional manipulation" is strongly rejected (Kligman & Culver, 1992). Indeed, we cannot call it manipulation if the so-called manipulator did not deliberate about manipulating a victim. Thus, an agent who unwittingly influences another agent without his knowledge cannot be manipulating. That is why we consider in the sequel manipulation as an *instrumentalization* of a victim. Unlike influence, which can relate to beliefs, knowledge or even intentions, instrumentalization is a deliberate influence on the effects produced by a victim whether those latters are deliberate or not.

Secondly, this instrumentalization is concealed from the victim. Manipulation cannot simply be reduced to coercion and persuasion (Handelman, 2009; Kligman & Culver, 1992; Sorlin, 2017) as it is something that happens completely invisibly, and by the time we start talking about manipulation, the act has already been committed. So, when we talk about manipulation, whether it is in the past tense or the second person, we are definitely stating something that the victim did not know. So the inevitable conclusion is that the target is necessarily unable to identify that he was subjected to a manipulative influence. The linguist Sandrine Sorlin claims that if one can say "he tried to manipulate me but failed", we cannot say "he manipulated me but failed" (Sorlin, 2017). He insists on the fact that success is embedded in manipulation. Manipulation implies the success of this enterprise.

Finally, manipulation is not similar to deception (Handelman, 2009; Kligman & Culver, 1992; Sorlin, 2017). Indeed, deception may be related to lying which can be one possible strategy to manipulate a victim. However not all manipulations are based on lies: e.g. telling the truth may be also a way to manipulate and induce somebody in the wrong way. That is why a manipulation should not be confused with its inner strategy. To summarize, manipulation has three fundamental characteristics:

(1) it is a deliberate effect of a manipulator (i.e. an applied strategy);
(2) it is an instrumentalization of a victim (i.e. an influence);
(3) it is always hidden from the victim (i.e. a concealment).

Consequently manipulation is an instrumentalization and it is concealed from the target. Hence, by considering a synthesis of the definitions given in (Akopova, 2013; Cohen, 2017; Handelman, 2009; Todd, 2013), we retain the following definition of manipulation.

**Definition 2.1.** A manipulation is a **deliberate effect** of an agent (called a *manipulator*) to **instrumentalize** another agent (called a *victim*), while making sure to **conceal that effect**.

It is important to notice that this definition makes sense in Artificial Intelligence literature, even in specific domains where manipulation has its own definition. In order to support our claim let us consider three domains: game theory (Ettinger & Jehiel, 2010), social choice theory (Gibbard, 1973) and reputation systems (Hoffman, Zage, & Nita-Rotaru, 2009) where manipulation has been widely studied. Obvious manipulation is not limited to those domains. The interested reader may refer to Masters et al. (2021) to have other examples.

(1) **Game theory** deals with the strategic interactions between agents and is based on the assumption they are rational, i.e. the agents always seek to maximize their personal or collective reward function according to the decisions they can make. In this setting, manipulation is just considered as a strategy: manipulators must decide if they have an interest to manipulate while victims must decide if they have an interest to not play an counter-manipulation strategy. A refinement of this notion has been proposed by Ettinger and Jehiel (2010): manipulation is expressed through a Bayesian game on which the victims are associated to *cognitive types* that describe how much they believe the other types of player have a manipulation strategy, and how much they are able to distinguish the strategies of the various types of their opponents. Here, manipulation is a deliberate strategy (a chosen action) with instrumentalization (the manipulator's utility depends on the victim's strategy) and concealment (as the less the victim is able to distinguish the manipulation strategy, the more interest there is to manipulate).
(2) **Social choice theory** deals with collective decision making based on the agents' preferences. Here, manipulation consists in as lying on our own preferences so as to obtain an outcome in our favor (Gibbard, 1973). Thus, it reduces manipulation as a particular strategy in which an agent is deliberately not sincere about its real preference profile. However, it also implies instrumentalization and concealment. Social choice theory deals with agents seeking for collective decisions. A manipulator lies (its deliberate strategy) and pushes other agents to make a given collective decision (it is an intrumentalization). Concealment results from

the fact that, to be effective, other agents must not be aware (or knowledgeable) of that strategy because, if it was the case, rational agents may adapt their behaviors to make the manipulation fail if it is in their interest (for instance by removing the manipulator from the decision process).

(3) **Reputation systems** are systems where agents interact, collect, share, and aggregate the results of their past interactions to decide where they should be agents they can trust for future interactions. In this setting, manipulation consists in either lying on his identity (to avoid to accumulate bad evaluations), producing wrong evaluation (to increase or reduce another agent's reputation), interacting in order to blur the other agents (so they produce wrong evaluations), or do not collecting or sharing some messages in order to isolate another agent (Hoffman, Zage, & Nita-Rotaru, 2009). As for the social choice theory, manipulation is here a deliberate strategy. Instrumentalization comes from the fact the manipulation wants to push the other agents to interact with him, or to forbid some agents to interact with the others. Finally, concealment comes from the fact that, if the other agents were aware or knew the manipulation, they should be able to isolate the manipulator.

While manipulation has been studied in multi-agent systems as a particular strategy or action e.g. in game theory or social choice theory as we showed, in this article we adopt a general position in regard to what is a manipulation in a multi-agent system by considering the definition 2.1. This definition makes clear the distinction with other related concepts such as e.g. coercion which is the instrumentalization of somebody by making him seeing a threat (thus without concealment) (McCloskey, 1980), persuasion which is the deliberate effect of changing beliefs (O'Keefe, 2015), lying which is the intention to induce a contrary beliefs from ours (Mahon, 2008), and so on. These concepts have been studied and disambiguated in the field of logic in multi-agent systems. Thus, in the next section, we provide a survey about related works in logic that are interested in defining such notions and how these formalisms can help us in defining manipulation.

## 3. Formalizing manipulation

From a general point-of-view, modal logics allow us 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 (Bottazzi & Troquard, 2015; Lorini & Sartor, 2016; Santos & Carmo, 1996), lying and deception (Sakama et al., 2015; Van Ditmarsch et al., 2012). In this section, since manipulation is a deliberate instrumentalization with concealment, we survey related works in literature about how to represent concealment (van der Hoek, Iliev, & Wooldridge, 2012) and deliberate effects (Broersen, 2008). First, we present works related on modeling of lies and dishonesty. In a second step, since manipulation is a deliberate effect to instrumentalize, we review logics that have been interested in representing influence and deliberate effects. Finally, we present works on modeling awareness.

### *3.1. Representing dishonesty, deception and lies*

Those three related concepts has been modeled by (Bonnet, Leturc, Lorini, & Sartor, 2021; Sakama et al., 2015; Van Ditmarsch et al., 2012). All those approaches are based

of the fact that an agent $i$ informs another agent $j$ about something so that $j$ believes it while the agent $i$ believes the opposite. For instance, Sakama et al. (2015) use modal logic and introduces a modality of communication between two agents, a modality of belief as well as a modality of intention to describe concepts such as lying, churning, concealment or deception. Bonnet et al. (2021) define persuasion as a deliberate action through which a persuader changes the beliefs of a persuadee, and define deception as persuading the persuadee of something under the assumption that the persuader believes that it is false. finally, Van Ditmarsch et al. (2012) uses dynamic doxastic logics with a modality to describe the action of private advertisements that is used to describe lying.

#### *3.1.1. A theory of dishonesty*

Sakama et al. (2015) define a logical theory of dishonesty in which they characterize notions of lying, bullshitting, concealment of information, deception, and half-truth. To formally describe these notions, they consider a formalism called BIC in which they considers for any agent $i, j \in \mathcal{N}$, modalities of beliefs $B_i$, intentions $I_i$ and communications, noted $C_{i,j}$, from an agent $i$ to another agent $j$. While Mahon (2008) first defines lying as making another person believe a false statement with the intent that the statement to be believed to be true by the other person, Sakama et al. (2015) take up this definition and define a simple lie with the predicate:

$$SimpleLie_{i,j}(\varphi) \triangleq C_{i,j}\varphi \wedge B_i\neg\varphi \wedge I_iB_j\varphi$$

However, Mahon (2008) also considers that lying to another person can be defined as the intention to make the intention to make the addressee believe that the statement is believed to be true by the speaker. Thus, Sakama et al. (2015) describe lying as:

$$Lie^*_{i,j}(\varphi) \triangleq SimpleLie_{i,j}(\varphi) \vee (C_{i,j}\varphi \wedge B_i\neg\varphi \wedge I_iB_jB_i\varphi)$$

To this notion of lying, Sakama et al. (2015) add the notion of lying by objective when the agent $i$ lying intends to make his addressee believe a proposition $\varphi$ by deduction. Here, the liar then lies about a statement that he does not believe $\sigma$, but believes that the statement will lead the other person to believe $\varphi$. This notion is then described by the predicate:

$$O - Lie^*_{i,j}(\sigma, \varphi) \triangleq I_iB_j\varphi \wedge \neg B_iB_j\varphi \wedge B_iB_j(\sigma \Rightarrow \varphi) \wedge B_i\neg\sigma \wedge C_{i,j}\sigma$$

Other notions are characterized as "bullshitting", which describes the action of communicating information that is not the case whether the agent communicating $i$ believes $\varphi$ and believes the opposite $\neg\varphi$, i.e. the predicate:

$$BS_{i,j}(\varphi) \triangleq C_{i,j}\varphi \wedge \neg B_i\varphi \wedge \neg B_i\neg\varphi$$

They also characterize withholding information as the fact that an agent $i$ intends not to reveal information $\varphi$ that $i$ believes to be true, that is:

$$WI_{i,j}(\varphi) \triangleq \neg C_{i,j}\varphi \wedge B_i\varphi \wedge I_i\neg B_j\varphi$$

Finally, the notion of half-truth consists in giving information that we believe to be true in order to mislead the other person by playing on an error of reasoning in the addressee. It is formally described by:

$$HT_{i,j}(\varphi,\psi) \triangleq (C_{i,j}\varphi \wedge B_i\varphi \wedge I_i B_j\varphi) \wedge \neg B_i B_j\psi \\ \wedge B_i B_j(\varphi \Rightarrow \psi) \wedge B_i\neg\psi \wedge \neg C_{i,j}\neg\psi \wedge I_i B_j\psi$$

The part of this predicate $(C_{i,j}\varphi \wedge B_i\varphi \wedge I_i B_j\varphi)$ is called the *intentional sincerity*. So, in their system, they show, for example, that lying about a proposition $\varphi$ while intending to be sincere is impossible and deduce the theorem $(C_{i,j}\varphi \wedge B_i\varphi \wedge I_i B_j\varphi) \wedge Lie^B_{i,j}(\varphi) \Rightarrow \bot$.

*3.1.2. A theory of persuasion and deception*

Bonnet et al. (2021) model notions such as influence, persuasion, deception and their logical relations. To do that they extend a propositional dynamic logic to a STIT logic and express a ex post knowledge $\mathsf{K}_i^{post}$ which 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. Here, a deliberate STIT $[\{i\}{:}\mathsf{dstit}]$ is defined as seeing to it that something is true, which is not necessary true in the other worlds in a current moment. There is also a temporal operator 'next' $\mathsf{X}$ defined as the applied joint action in the current world and a belief modality $\mathsf{B}_j$. In their approach, persuasion is defined as a deliberate action through which an agent (persuader) changes the beliefs of another agent's (persuadee):

$$\mathsf{Persuades}(i,j,\varphi) \triangleq \mathsf{K}_i^{post}[\{i\}{:}\mathsf{dstit}]\mathsf{X}\mathsf{B}_j\varphi$$

Based on this notion, they define deception as a persuasion in a proposition $\varphi$ under the assumption that the persuader believes that $\varphi$ is false.

$$\mathsf{Deceives}(i,j,\varphi) \triangleq \mathsf{Persuades}(i,j,\varphi) \wedge \mathsf{B}_i\mathsf{X}\neg\varphi$$

Here, truthfully telling is simply captured by persuasion, while deception always involves not telling the truth. This notion of deception is then refined to express more subtle notions, such as smooth-talking (persuading since the persuader is uncertain whether $\varphi$ is true or false) which is equivalent to Sakama et al. (2015)'s "bullshitting".

$$\mathsf{PersuadesBySmoothTalking}(i,j,\varphi) \triangleq \mathsf{Persuades}(i,j,\varphi) \\ \wedge \neg\mathsf{K}_i^{post}\mathsf{X}\neg\varphi \wedge \neg\mathsf{K}_i^{post}\neg\mathsf{X}\neg\varphi$$

Finally, Bonnet et al. (2021) distinguish benevolent, malevolent and reckless deception, which differ in how persuading an agent $j$ of something is good for $j$ (or not). In the malevolent form, the deceiver believes that believing $\varphi$ will have bad consequences for the deceived. On the contrary, in a benevolent deception the deceiver believes that believing $\varphi$ is good for the deceived. Finally, reckless deception consists for the deceiver to not know whether that $\varphi$ is good or bad for the deceived.

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

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

$$\mathsf{RecklessDeception}(i,j,\varphi) \triangleq \mathsf{Deceives}(i,j,\varphi) \wedge \neg\mathsf{K}_i^{post}\mathsf{X}(\mathsf{B}_j\varphi \rightarrow \mathsf{good}_j) \wedge \neg\mathsf{K}_i^{post}\mathsf{X}(\mathsf{B}_j\varphi \rightarrow \mathsf{bad}_j)$$

*3.1.3. A dynamic logic of lying*

While Bonnet et al. (2021) and Sakama et al. (2015) consider predicates to express lying, deception, or persuasion, another approach Van Ditmarsch et al. (2012) gets rid of representing those notions through predicates. Instead, they consider lying as a public announcement which is false, and they use a dynamic doxastic logic to describe the effects of lying through the semantics of the logical system. In this work, they consider a language $\mathcal{L}_{ll}$ generated by the following BNF:

$$\varphi ::= p|(\varphi_1 \wedge \varphi_2)|\neg\varphi|\top|\bot|B_i\varphi|[\ddagger\varphi_1]\varphi_2|[!\varphi_1]\varphi_2|[¡\varphi_1]\varphi_2$$

In this language, they consider a set of modal operators and public announcement operators as:

- $B_i\varphi$ means the agent $i$ believes that $\varphi$ is true,
- $[!\varphi]\psi$ means that $\psi$ is true after the public announcement of $\varphi$,
- $[¡\varphi]\psi$ means that $\psi$ is true after the lie $\varphi$,
- $[\ddagger\varphi]\psi$ means that $\psi$ is true after the public announcement of $\varphi$ but also true after the lie $\varphi$.


**Figure 1.** Event-driven model of the logic of lying (Van Ditmarsch et al., 2012)

To interpret these operators they first consider a doxastic model $\mathcal{M} = (\mathcal{W}, \{\mathcal{B}_i\}_{i\in\mathcal{N}}, V)$ with $\{\mathcal{B}_i\}_{i\in\mathcal{N}}$ a set of serial, transitive and euclidean relations. Secondly they define an action model $\mathcal{A}_p = (\mathcal{W}^\alpha, \{\mathcal{B}_i^\alpha\}_{i\in\mathcal{N}}, I^\alpha, P^\alpha)$ with $I^\alpha : \mathcal{W}^\alpha \rightarrow \mathcal{L}_{ll}$ the *precondition function* and $P^\alpha : \mathcal{W}^\alpha \rightarrow Sub_{\mathcal{L}_{ll}}$ a function called *postcondition function* which assigns to each possible event a substitution which represents the effects of actions on variables. If $\mathcal{A}(\mathcal{L}_{ll}) = \{p_1, \ldots, p_n\}$ is the set of propositional variables of $\mathcal{L}_{ll}$, elements of $Sub_{\mathcal{L}_{ll}}$ are represented by:

$$\{p_1 \mapsto \sigma_1, \ldots, p_n \mapsto \sigma_n | \sigma_1, \ldots, \sigma_n \in \mathcal{L}_{ll}\}$$

The action model they use is described by Figure 1 which expresses the effects of actions, here public announcements. For instance the event $1 : p$ represents the public announcement of $p$ where $p$ is true while $0 : \neg p$ represents the possible event when it is a lie i.e. $p$ is announced while $p$ is false. Furthermore let us notice that $p$ can be

substituted for any formula $\varphi$ which means that public announcement are not necessary related to atomic propositions. Then, to define the effect of public announcement on the beliefs of the agents, this is given by a standard product update defined as $(\mathcal{M}, U) \otimes (\mathcal{A}, S) = ((\mathcal{W}^{\otimes}, \{\mathcal{B}_i^{\otimes}\}_{i\in\mathcal{N}}, V^{\otimes}), U^{\otimes})$ where $U \subseteq \mathcal{W}$, $S \subseteq \mathcal{W}^{\alpha}$ and:

- $\mathcal{W}^{\otimes} = \{(v, f) \in \mathcal{W} \times \mathcal{W}^{\alpha} : \mathcal{M}, v \models I^{\alpha}(f)\}$
- $\forall (v, f) \in \mathcal{W}^{\otimes}, \forall i \in \mathcal{N}, \mathcal{B}_i^{\otimes}(v, f) = \{(u, g) \in \mathcal{W}^{\otimes} : u \in \mathcal{B}_i(v), g \in \mathcal{B}_i^{\alpha}(f)\}$
- $\forall (v, f) \in \mathcal{W}^{\otimes} : V^{\otimes}(v, f) = \{p \in \mathcal{A}(\mathcal{L}_B) | \mathcal{M}, w \models P(s)(p)\}$
- $U^{\otimes} := \{(v, f) | v \in U, f \in S, (v, f) \in \mathcal{W}^{\otimes}\}$

Finally, the semantics of the operators of $\mathcal{L}_{ll}$ is given by:

- $\mathcal{M}, w \models B_i\varphi$ iff $\forall v \in \mathcal{W} : w\mathcal{B}_i v, \mathcal{M}, v \models \varphi$
- $\mathcal{M}, w \models [!\varphi]\psi$ iff $(\mathcal{M}, \{w\}) \otimes (\mathcal{A}_{\varphi}, \{1\}), (w, 1) \models \psi$
- $\mathcal{M}, w \models [¡\varphi]\psi$ iff $(\mathcal{M}, \{w\}) \otimes (\mathcal{A}_{\varphi}, \{0\}), (w, 0) \models \psi$
- $\mathcal{M}, w \models [‡\varphi]\psi$ iff $(\mathcal{M}, \{w\}) \otimes (\mathcal{A}_{\varphi}, \{0, 1\}), (w, 0) \models \psi$ and $(\mathcal{M}, \{w\}) \otimes (\mathcal{A}_{\varphi}, \{0, 1\}), (w, 1) \models \psi$

This semantics then gives the axiomatic system summarized in Figure 2, the manipulative announcement is then described as $[‡\varphi]\psi$ and by **Ax0**, $[‡\varphi]\psi \Leftrightarrow [!\varphi]\psi \wedge [¡\varphi]\psi$. The rule **A1** describes that if a public announcement $\varphi$ brings the consequence $p$, then if $\varphi$ is true, then necessarily $p$ is true and vice versa. The formula $[!\varphi]p \Leftrightarrow \varphi \Rightarrow p$ is then a tautology of the system.

| | |
|---|---|
| (**CP**) | $\vdash \varphi$, for all PC theorems $\varphi$ |
| (**Ax0**) | $\vdash [‡\varphi]\psi \Leftrightarrow [!\varphi]\psi \wedge [¡\varphi]\psi$ |
| (**A1**) | $\vdash [!\varphi]p \Leftrightarrow \varphi \Rightarrow p$ |
| (**A2**) | $\vdash [!\varphi]\neg\psi \Leftrightarrow \varphi \Rightarrow \neg[!\varphi]\psi$ |
| (**A3**) | $\vdash [!\varphi](\psi_1 \wedge \psi_2) \Leftrightarrow [!\varphi]\psi_1 \wedge [!\varphi]\psi_2$ |
| (**A4**) | $\vdash [!\varphi]B_i\psi \Leftrightarrow \varphi \Rightarrow B_i[!\varphi]\psi$ |
| (**L1**) | $\vdash [¡\varphi]p \Leftrightarrow \neg\varphi \Rightarrow p$ |
| (**L2**) | $\vdash [¡\varphi]\neg\psi \Leftrightarrow \neg\varphi \Rightarrow [¡\varphi]\psi$ |
| (**L3**) | $\vdash [¡\varphi](\psi_1 \wedge \psi_2) \Leftrightarrow [¡\varphi]\psi_1 \wedge [¡\varphi]\psi_2$ |
| (**L4**) | $\vdash [¡\varphi]B_i\psi \Leftrightarrow \neg\varphi \Rightarrow B_i[¡\varphi]\psi$ |

**Figure 2.** Simplified axiomatic system of Van Ditmarsch et al. (2012)

Let us notice that other works using dynamic logic to model lying or deceiving exist. For instance, Sakama (2021) recently extends his previous work (Sakama et al., 2015) and proposes to model deception with a doxastic dynamic semantics.

### *3.2. Representing the effects of actions*

As manipulation is a deliberate effect in order to instrumentalize an agent, it is of interest to reason on actions, their intended consequences, their side-effects, and how they can influence the other agents. Many formalisms exist. For instance, dynamic logics (Harel, Kozen, & Tiuryn, 2001) and temporal logics (Alur, Henzinger, & Kupferman, 2002) consider several action modalities where each action modality is associated with a program and its outputs. Dynamic epistemic logics express the logical consequences generated by public or private announcements of agents (Van Ditmarsch, Van Der Hoek, & Kooi, 2007). Many other formalisms exist such as *fluent calculus*. For a

detailed survey, the interested reader may refer to (Segerberg, Meyer, & Kracht, 2009). Interestingly, (Giordano, Martelli, & Schwind, 2000) propose a formalism that catches all ramification effects of actions i.e. all direct consequences and side-effects of actions. However such formalism introduces a distinct modality for each possible action, while we saw in Table 1 that manipulation does not depend on particular actions or strategies (e.g. lying, rumor propagating, emotional blackmailing) but rather on its results. Thus, it is of interest to consider action logics which only represent abstract strategies that lead to a state-of-affair. Two approaches seem relevant: the STIT logics (Balbiani, Herzig, & Troquard, 2008; Belnap & Perloff, 1988; Lorini & Sartor, 2016) and the BIAT logics (Pörn, 1977; Santos & Carmo, 1996; Troquard, 2014), which both consider, in an abstract way, the fact of ensuring that something is done.

Both STIT and BIAT logics consider actions as the fruit of their consequences. This level of abstraction is well-adapted to define manipulation. BIAT considers a modality $E_i$ which means that the agent $i$ *brings it about*. Let us notice it is *side-effects free*, i.e. indirect consequences of actions are not considered as intended effects. For its part, STIT considers a modality $[STIT]_i$ which means 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 considers a S5 system[1] whereas BIAT is a non normal system based on neighborhood functions. Furthermore, STIT uses a notion of temporality while BIAT does not.

### *3.2.1. Representing influence*

In the literature about STIT, some approaches already defined a notion of influence. Here, influence consists in seeing to it that another agent will see to it that something becomes true. For instance, Lorini and Sartor (2016) define the fact that agent $i$ influences agent $j$ on $\varphi$ if, and only if, agent $i$ sees to it that in the future, agent $j$ rationally sees to it that $\varphi$, i.e.:

$$[infl]_{i,j}\varphi \triangleq [stit]_i X[rstit]_j\varphi$$

where $[rstit]$ is a primitive modal operator which represents the current choice as a rational choice. For its part, BIAT logics define influence similarly as the effect of an agent $i$ to bring it about that another agent $j$ brings about something. For instance, Bottazzi and Troquard (2015) considers the following predicate to define influence which is called *interpersonal control*:

$$[infl]_{i,j}\varphi \triangleq E_i(E_j\varphi \wedge \psi)$$

The modality $E_i$ (resp. $E_j$) refers to a non-normal modality which means that agent $i$ (resp. agent $j$) brings it about something. The formula $\psi$ represents any formula of the language that does not contradict $E_j\varphi$. This $\psi$ is explicitly included in the definition of the predicate of influence because of the semantics which does not allow us to have the property of normal logics $\Box(\varphi \wedge \psi) \equiv \Box\varphi \wedge \Box\psi$. If influence is defined only as $E_iE_j\varphi$ and if a formula $E_i(E_j\varphi \wedge \psi)$ is verified with $\psi \neq \top$, then since we do not have the property of normal logics, we could not be able to deduce the influence $E_iE_j\varphi$

[1] A S5 system for a $\Box$ modality is a system where the axioms: $(K)$ $\Box(\varphi \Rightarrow \psi) \Rightarrow (\Box\varphi \Rightarrow \Box\psi)$, $(T)$ $\Box\varphi \Rightarrow \varphi$, (4) $\neg\Box\varphi \Rightarrow \Box\neg\Box\varphi$ and (5) $\Box\varphi \Rightarrow \Box\Box\varphi$ are considered.

from $E_i(E_j\varphi \wedge \psi)$.

*3.2.2. Representing deliberate effects*

While STIT approaches define influence, they also define a notion of deliberate effect. Lorini and Sartor (2016) define deliberate effect by considering that something is done deliberately by another agent $i$ if, and only if, $i$ sees to it that something is true while it is not necessarily the case, that is to say:

$$[dstit]_i\varphi \stackrel{\triangle}{=} [stit]_i\varphi \wedge \neg\Box\varphi$$

However, this definition is not without problems. Let us imagine a situation in which one agent $i$ deliberately causes a car crash to take advantage of car insurance. But during this car crash, a person died. By following the formal definition of Lorini and Sartor (2016)'s deliberate STIT, we would deduce that the agent $i$ deliberately 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 $i$ deliberately 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 deliberately do or do not something, then we know what we are doing. Because they use standard STIT, Lorini and Sartor (2016) do not and cannot consider positive and negative introspection on knowledge. Let us notice that Broersen (2008) integrates a knowledge modality into the STIT language and defines a deliberate see to it such as:

$$[dxstit]_i\varphi \stackrel{\triangle}{=} K_i[xstit]_i\varphi \wedge K_i\neg\Box X\varphi$$

Here, $[xstit]_i\varphi$ denotes that in the future, agent $i$ sees to it that $\varphi$. Broersen (2008)'s deliberate STIT has the positive and negative introspection with knowledge. Furthermore it has also the side-effect free property since a consequence may not be known by the agent to consider it as a deliberate effect. A third important property that deliberate effects has to be verified which is *the conjunction of all deliberate effects forms a deliberate effect*. This property means that if agent $i$ deliberately sees to it that $\varphi$ and also deliberate sees to it that $\psi$, then he deliberately sees to it that $\varphi \wedge \psi$; and Broersen (2008)'s deliberate STIT has this property.

Concerning BIAT, as written previously, it explicitly defines a deliberate effect modality and it is side-effects free.

### *3.3. Representing revelation, concealment and awareness*

Lastly, we have seen that manipulation needs to conceal its effects. In the literature, this notion of concealment can be expressed either in terms of an agent's knowledge, or in terms of his level of awareness. In this section, we highlight some logical approaches that have represented those notions.

*3.3.1. A logic of revelation and concealment*

To the best of our knowledge, only van der Hoek et al. (2012) proposed a logic for representing revelation and concealment as a Propositional Dynamic Logic (PDL). They considered particular actions for revelation and concealment, denoted (resp) by $r(p,i)$ and $c(p,i)$. The formula $[r(p,i)]\varphi$ means that after revealing the value of $p$ to an agent $i$, the formula $\varphi$ is true, while $[c(p,i)]\varphi$ means that after concealing the value of $p$ to an agent $i$, the formula $\varphi$ is true. These modalities are described in a PDL logical frame. They combined to action a standard S5 epistemic operator $K_i$ with a Kripke relationship.

Their formalism makes possible to express and deduce interesting valid formulas as if a proposition $p$ is true, then after revealing $p$ to one agent $i$, then this agent will know that $p$ is true i.e.:

$$p \Rightarrow [r(p,i)]K_i p$$

In the same way, this formalism provides other valid formulas which combine effects of actions on the knowledge of agents. For instance, they can express a validity as e.g. if after doing an action $\alpha$, the variable $p$ is true, then if we reveal $p$ to one agent $i$ and then do $\alpha$, then $i$ will know that $p$ is true:

$$[\alpha]p \Rightarrow [r(p,i);\alpha]K_i p$$

In their logical framework, they are able to express interesting validities related to that action of concealment as e.g. after concealing to agent $i$ the value of $p$ it is possible that after some execution of one action $\alpha$, $\varphi$ is true, is equivalent to, after some execution of $\alpha$ and, after concealing to agent $i$ the value of $p$, makes $\varphi$ true, i.e.:

$$\langle c(p,i)\rangle\langle\alpha\rangle\varphi \Leftrightarrow \langle\alpha\rangle\langle c(p,i)\rangle\varphi$$

*3.3.2. Representing awareness*

Contrary to concealment, many works have taken an interest in formally representing the semantics of being aware of (Hill, 2010; Modica & Rustichini, 1994; Schipper, 2014; Van Ditmarsch, French, Velázquez-Quesada, & Wáng, 2018).

Fagin and Halpern (1987) consider Levesque's logic on implicit and explicit knowledge to model a logic of non-omniscience (Levesque, 1984). Modica and Rustichini (1994) consider that an agent $i \in \mathcal{N}$ is aware of something, denoted $A_i\varphi$, if, and only if, that agent knows $\varphi$ or it is not the case it knows $\varphi$ and it knows that it is not the case it knows $\varphi$, i.e. checks the predicate $A_i\varphi := K_i\varphi \vee (\neg K_i\varphi \wedge K_i\neg K_i\varphi)$ where $K_i$ is a knowledge modality. If $K_i$ is associated with a modal system S5, then the previous definition of consciousness becomes equivalent to $A_i\varphi := K_i\varphi \vee K_i\neg K_i\varphi$.

While previous mentioned approaches use knowledge or belief operators semantically defined by a Kripke relationship to define awareness, Schipper (2014) proposes to model awareness semantically with a function $\mathcal{A}_i : \mathcal{W} \to 2^{\mathcal{L}}$ that maps from a possible world, the set of formulas that agent $i$ is aware of. It defines a modal language $\mathcal{L}$ in which a modality $L_i$ refers to the implicit knowledge of an agent $i$, and a modality $A_i\varphi$ expresses the *awareness* of an agent $i$ on a proposition $\varphi$. The semantics of this modality $A_i$ is defined with a function $\mathcal{A}_i : \mathcal{W} \to 2^{\mathcal{L}}$. This function is called *Awareness Correspondence*. Thus, an agent $i$ is aware of a proposition $\varphi$ if, and only if, this formula $\varphi$ belongs to a set of formulas given by this correspondence function.

$\mathcal{A}_i : \mathcal{W} \to 2^{\mathcal{L}}$ associates for a given world, all the formulas that a $i$ agent is aware of. Knowledge is then defined by the predicate $K_i\varphi \stackrel{\triangle}{=} L_i\varphi \wedge A_i\varphi$. In this formalism, a model is a tuple, $\mathcal{M} = (\mathcal{W}, \{\mathcal{R}_i\}_{i\in\mathcal{N}}, \{\mathcal{A}_i\}_{i\in\mathcal{N}}, V)$ such that $(\mathcal{W}, \{\mathcal{R}_i\}_{i\in\mathcal{N}}, V)$ is a Kripke frame where $\{\mathcal{R}_i\}_{i\in\mathcal{N}}$ is a set of equivalence relations associated to the semantics of $L_i$ for each agent $i \in \mathcal{N}$. To define the correspondence function $\mathcal{A}_i$, it is then necessary to use a function that returns the set of atomic propositions of a formula, this function $At$ is such that:

- $At(\top) = \emptyset$
- $At(p) = \{p\}$ if $p$ is an atom of $\mathcal{L}$
- $At(\neg\varphi) = At(\varphi)$
- $At(\varphi \wedge \psi) = At(\varphi) \cup At(\psi)$
- $At(K_i\varphi) = At(\varphi)$
- $At(A_i\varphi) = At(\varphi)$

So, for each agent $i \in \mathcal{N}$ and all possible worlds $w \in \mathcal{W}$, the correspondence function is such that:

(1) $\varphi \in \mathcal{A}_i(w)$ if, and only if, $At(\varphi) \subseteq \mathcal{A}_i(w)$
(2) $\forall v \in \mathcal{W} : w\mathcal{R}_i v \Rightarrow \mathcal{A}_i(w) = \mathcal{A}_i(v)$

Property 1 means that, for an agent $i$ to be aware of a formula $\varphi$, it is necessary for that agent to be aware of all the propositional atoms contained in a formula. Property 2 corresponds to the fact that an agent knows what it is aware of. Finally, a pattern is defined in a standard way. We have, for every agent $i \in \mathcal{N}$ and for each world $w \in \mathcal{W}$, the following semantics:

- $\mathcal{M}, w \models L_i\varphi$ iff $\forall v \in \mathcal{W}, w\mathcal{R}_i v : \mathcal{M}, v \models \varphi$
- $\mathcal{M}, w \models A_i\varphi$ iff $\varphi \in \mathcal{A}_i(w)$
- $\mathcal{M}, w \models K_i\varphi$ iff $\mathcal{M}, w \models L_i\varphi$ and $\mathcal{M}, w \models A_i\varphi$

This semantics then gives us the axiomatic system of the figure 3. An immediate theorem of consciousness is that if an agent $i$ knows that something is true then necessarily this agent $i$ is aware of that something. Indeed:

- $\vdash K_i\varphi \Leftrightarrow L_i\varphi \wedge A_i\varphi$
- $\vdash L_i\varphi \wedge A_i\varphi \Rightarrow A_i\varphi$
- $\vdash K_i\varphi \Rightarrow A_i\varphi$

Finally, all these works have been extended in Van Ditmarsch et al. (2018) where they define a logic of speculative knowledge allowing to reason about the notion of non awareness, based on the logic of implicit and explicit knowledge of Fagin and Halpern (1987) and the logic of consciousness of Schipper (2014). This notion of speculative knowledge is distinct from that of implicit knowledge and explicit knowledge. It translates that an agent has speculative knowledge of a proposition $\varphi$ in a model $\mathcal{M}$ and a world $w$ if this formula is verified in all accessible worlds in all pointed models of the framework that are $\mathcal{A}_i(w)$-bisimilar[2] to the world $w$.

[2]A bisimulation is a relation between two models in which related states have identical atomic information and matching transition possibilities (Blackburn, De Rijke, & Venema, 2002).

| | |
|---|---|
| (**CP**) | $\vdash \varphi$, for all PC theorems $\varphi$ |
| (**KL**) | $\vdash K_i\varphi \Leftrightarrow L_i\varphi \wedge A_i\varphi$ |
| (**S5**$(L_i)$) | $\vdash \varphi$, for all theorems $\varphi$ of S5 associated with $L_i$ |
| (**AS**) | $\vdash A_i\varphi \Leftrightarrow A_i\neg\varphi$ |
| (**AC**) | $\vdash A_i(\varphi \wedge \psi) \Leftrightarrow A_i\varphi \wedge A_i\psi$ |
| (**AKR**) | $\vdash A_i\varphi \Leftrightarrow A_iK_i\varphi$ |
| (**AR**) | $\vdash A_i\varphi \Leftrightarrow A_iA_i\varphi$ |
| (**ALR**) | $\vdash A_i\varphi \Leftrightarrow A_iL_i\varphi$ |
| (**UL**) | $\vdash \neg A_i\varphi \Leftrightarrow L_i\neg A_i\varphi$ |
| (**MP**) | If $\vdash \varphi \Rightarrow \psi$ and $\vdash \varphi$ then $\vdash \psi$ |
| (**Nec**) | If $\vdash \varphi$ then $\vdash L_i\varphi$ |

**Figure 3.** Axiomatic system of the logic of awareness (Schipper, 2014)

## 4. A modal logic for manipulation

In this section, we propose a logic to represent the notion of manipulation as it has been defined in Section 2.1. Let us remark our goal is not to propose a definitive logic but rather to provide the basic elements to express such definition and disambiguate this concept.

To this end, we consider a logic with several modalities: deliberate effects, consequences of actions, belief and knowledge. One of the main element of this logic is the deliberate effect operator. Indeed, we require that this operator must have: (1) negative and positive introspection; (2) side-effect free; (3) the conjunction of all deliberate effects forms a deliberate effect. As written previously, STIT semantic catches all (direct and indirect) consequences of actions while BIAT semantics catches the deliberate effects. Moreover, since BIAT is side-effect free and makes it easy to express positive and negative introspection and also the conjunction of all deliberate effects forms a deliberate effect, we distinguish a modality of deliberate effects (expressed by a BIAT-like modality denoted $E_i^d$ in the sequel) from a modality that catches all consequences of performed actions (expressed by a STIT-like modality denoted $E_i$ in the sequel). Thanks to these modalities, combined with knowledge and belief, we are able to express instrumentalization and concealment. Furthermore distinguishing an epistemic modal operator from a belief operator allows us to define different levels of concealment. Then, we chose to not introduce an awareness operator explicitly in the language and prefer to keep the logic simple. Finally, we also do not want the modality of future which increases the complexity of the model.

### *4.1. Language*

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

$$\varphi ::= p \mid \neg\varphi \mid (\varphi \vee \varphi) \mid (\varphi \wedge \varphi) \mid (\varphi \Rightarrow \varphi) \mid K_i\varphi \mid B_i\varphi \mid E_i\varphi \mid E_i^d\varphi$$

The formula $E_i\varphi$ means that the actions of $i$ lead to $\varphi$. So $E_i$ represents the effects of actions which may have been deliberated or not such as side-effects. The formula $E_i^d\varphi$

means[3] that agent $i$ deliberately brings it about that $\varphi$. This modality is semantically represented with a neighborhood function which associates a set of possible worlds for each deliberate effect. Finally the formulas $K_i\varphi$ and $B_i\varphi$ mean respectively that the agent knows that $\varphi$, and the agent believes that $\varphi$. We note the dual operators as following: $\langle K_i\rangle\varphi = \neg K_i\neg\varphi$, $\langle B_i\rangle\varphi = \neg B_i\neg\varphi$, $\langle E_i\rangle\varphi = \neg E_i\neg\varphi$ and $\langle E_i^d\rangle\varphi = \neg E_i^d\neg\varphi$.

### *4.2. Associated semantics*

We consider the following logical frame:

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

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.:

$$\forall i \in \mathcal{N}, \mathcal{E}_i^d : \mathcal{W} \to 2^{2^{\mathcal{W}}}$$

The reasons why we consider neighborhood functions for the deliberate brings it about semantics are twofold. Firstly, neighborhood semantics allow us to be free from the necessitation since we do not want to express that the agent deliberately brings it about a tautology. Secondly, the neighborhood functions translate semantically sets of possible worlds that results from strategies deliberated by the agent. In this case, each set of possible worlds corresponds to the deliberate effects of those strategies.

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 $w \in \mathcal{W}$, $\varphi, \psi \in \mathcal{L}_{KBE}$, and $p \in \mathcal{P}$, we inductively define $\mathcal{M}, w \models \varphi$ as:

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

Let us detail the difference between rule (10) and rule (11). We do not want to express that an agent deliberately brings it about a tautology. Thus, since the rule (11) is defined on a neighborhood function, it allows $E_i^d$ to be freed from necessitation, while since the rule (10) characterizes a normal modality, the necessitation holds.

We also consider a dual notion of deliberate effects, denoted $\langle E_i^d\rangle$, which translates "it is possible that one agent deliberately brings it about something as he does not deliberately brings it about the contrary". Formally, we write $\mathcal{M}, w \models \langle E_i^d\rangle\varphi$ if, and only if, $\mathcal{W} \setminus ||\varphi|| \notin \mathcal{E}_i^d(w)$, with $||\varphi|| := \{v \in \mathcal{W} : \mathcal{M}, v \models \varphi\}$.

Finally, let us remind that $\varphi$ is valid in $\mathcal{M}$ (written $\mathcal{M} \models \varphi$) if, and only if, for all worlds $w \in \mathcal{W}$, $\varphi$ is satisfiable in $w$, i.e. $\mathcal{M}, w \models \varphi$ is true. A formula $\varphi$ is valid in a

[3] We also accept the expressions "$\varphi$ is a deliberate effect of agent $i$" or "agent $i$ deliberately sees to it that $\varphi$".

frame $\mathcal{C}$ (written $\models_{\mathcal{C}} \varphi$ or $\mathcal{C} \models \varphi$) if, and only if, for all models $\mathcal{M}$ built on $\mathcal{C}$, $\mathcal{M} \models \varphi$. In this case $\varphi$ is a tautology of $\mathcal{C}$, written $\models_{\mathcal{C}} \varphi$.

#### *4.2.1. Semantics for beliefs and knowledge*

For the modalities of knowledge and belief, we conventionally constrain our frame $\mathcal{C}$ so that, for any agent $i \in \mathcal{N}$, $\mathcal{K}_i$ is *reflexive*, *transitive* and *confluent*[4] and $\mathcal{B}_i$ is *serial*, *transitive* and *Euclidean*. The constraints and relations between these two modalities have already been well studied (Stalnaker, 2006). 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:

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

#### *4.2.2. Semantics for the effects of actions*

It is commonly accepted to consider the effects of action as an equivalence relationships, such as in STIT. Thus, for any agent $i \in \mathcal{N}$, $\mathcal{E}_i$ is a reflexive, transitive and Euclidean relationship[5]. Indeed, when an agent implements one or more actions, if he has carried them out, it means that this is the case and that the consequence $\varphi$ is true. The reflexivity, denoted $(E1)$, expresses the fact that once the actions leading to a certain $\varphi$ consequence have been performed by an agent $i$, then that consequence $\varphi$ is necessarily true in the current world. This constraint $(E1)$ gives the axiom T.

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

From this constraint, we immediately deduce that this relation $\mathcal{E}_i$ is also *serial*. This translates that if an agent $i$ ensures that one or several actions lead to a certain consequence $\varphi$, it is not the case that this action or series of actions can lead to the opposite in the current world. So we deduce in such a system the property D. The relation $\mathcal{E}_i$ is also *transitive* since when the actions of agent $i$ lead to $\varphi$, these actions also lead to the fact that these actions are done properly, i.e.:

$$\forall w, u, v \in \mathcal{W} : w\mathcal{E}_i u \wedge u\mathcal{E}_i v \Rightarrow w\mathcal{E}_i v \quad (E2)$$

---

[4] A binary relation $\mathcal{R}$ on $\mathcal{W}$ is *confluent* if, and only if, the following property is satisfied $\forall w, u, v \in \mathcal{W}, w\mathcal{R}u \wedge w\mathcal{R}v \rightarrow \exists z \in \mathcal{W} : u\mathcal{R}z \wedge v\mathcal{R}z$. Here we do not consider a S5 system with negative introspection but a S4.2 system. A S4.2 system is a S4 system – a system with the axioms for a $\Box$ modality $(K)$ $\Box(\varphi \Rightarrow \psi) \Rightarrow \Box\varphi \Rightarrow \Box\psi$, $(T)$ $\Box\varphi \Rightarrow \varphi$, and (4) $\neg\Box\varphi \Rightarrow \Box\neg\Box\varphi$ – with a 4.2 axiom, ie. $\Diamond\Box\varphi \Rightarrow \Box\Diamond\varphi$. The main reason is that since we would like to model also human agents' reasoning, we cannot accept that humans know everything they do not know. For more details, the interested reader may refer to Stalnaker (2006) who gave arguments to support S4.2 rather than S5 for modeling knowledge.

[5] An equivalence relationship is by definition a reflexive, transitive and symmetrical relationship, but equivalently we can consider any reflexive, transitive and Euclidean relationship.

Finally, if an agent $i$ does not perform actions that lead to some consequences $\varphi$, then agent $i$ indirectly performs actions that lead to not realize the actions that lead to $\varphi$. Thus the relation $\mathcal{E}_i$ is *Euclidean*, i.e.:

$$\forall w, u, v \in \mathcal{W} : w\mathcal{E}_i u \wedge w\mathcal{E}_i v \Rightarrow u\mathcal{E}_i v \quad (E3)$$

Let us notice that we do not consider positive and negative introspection with knowledge because since $E_i$ represents the effects of actions, some effects may not be known by the agent $i$ as side-effects.

*4.2.3. Semantic representation of deliberate effects*

While the effects of actions are represented as an equivalence relationship, we represent deliberate effects as a neighborhood function. This semantic difference is justified by the fact that when an agent deliberately brings it about something, the latter considers as many sets of possible worlds as deliberate effects which has been carried out. Moreover, Kripke's relations do not allow us to express the fact that an agent cannot deliberately bring it about something that he always knows to be true, such as tautologies. It is due to the principle of necessity which is valid in any Kripke frame.

Thus, the first semantic difference with the $E_i$ modality is that an agent $i$ cannot deliberately bring it about that a tautology to be true since he knows it is always true. This constraint, denoted $\overline{E^d_{Nec}}$, is semantically translated by the fact that the set of all possible worlds cannot belong to any neighborhood, i.e.:

$$\forall w \in \mathcal{W} : \mathcal{W} \notin \mathcal{E}^d_i(w) \quad (\overline{E^d_{Nec}})$$

Moreover, there is a logical link between the modality of deliberate effects and that of the effects of actions that may or may not be deliberate. Indeed, if an agent $i$ deliberately brings it about some action, then this agent does these actions. It is represented by the constraint $E^d E$ which is semantically translated by the fact that all the possible worlds reachable by the relation $\mathcal{E}_i$ are always included in all possible worlds sets of $\mathcal{E}^d_i$, i.e.:

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

When an agent $i$ deliberately brings it about that a proposition $\varphi$ to be true while deliberately bringing it about that another proposition $\psi$ to be true, then that agent $i$ deliberately brings it about that $\varphi \wedge \psi$ to be true, i.e.:

$$\forall w \in \mathcal{W} : S \in \mathcal{E}^d_i(w) \wedge T \in \mathcal{E}^d_i(w) \Longrightarrow S \cap T \in \mathcal{E}^d_i(w) \quad (E_{\Rightarrow,\wedge})$$

However, we cannot consider the reciprocal of $(E_{\Rightarrow,\wedge})$ because deliberate effects concern a whole and cannot be considered as the sum of its part. For example, when somebody decides to eat a hazelnut cake, he does not decide to deliberately eat the dough of the cake and eat the hazelnuts independently. Moreover, this reciprocal, defined by $\forall w \in \mathcal{W} : S \cap T \in \mathcal{E}^d_i(w) \Longrightarrow S \in \mathcal{E}^d_i(w) \wedge T \in \mathcal{E}^d_i(w)$ and associated with the theorem $E^d_i(\varphi \wedge \psi) \Rightarrow E^d_i\varphi \wedge E^d_i\psi$, cannot be considered for technical reasons. Indeed, an immediate result is that such constraint is equivalent to $\forall w \in \mathcal{W} : S \in \mathcal{E}^d_i(w) \wedge S \subseteq T \Longrightarrow T \in \mathcal{E}^d_i(w)$. Hence, if this reciprocal is considered, we would get an inconsistent logical system due to the constraint $(\overline{E^d_{Nec}})$.

Finally, an essential characteristic of deliberate effects is that it is introspective to the knowledge of agents. An agent always knows what it is doing and what it is not doing deliberately. Thus, the deliberate effect modality has positive ($E^d_{KP}$) and negative introspection ($E^d_{KN}$) with respect to the knowledge of the agent, i.e.:

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

$$\forall w, v \in \mathcal{W}, \forall S \in 2^{\mathcal{W}} : S \notin \mathcal{E}^d_i(w) \Longrightarrow (w\mathcal{K}_i v \Rightarrow S \notin \mathcal{E}^d_i(v)) \quad (E^d_{KN})$$

These constraints mean that when an agent $i$ deliberately brings it about a consequence, she knows what she is doing deliberately. The same is true when an agent $i$ does not deliberately bring it about that a consequence, then the agent $i$ knows that she did not do it deliberately. Furthermore, let us notice that ($E^d_{KN}$) is equivalent to:

$$\forall w \in \mathcal{W}, \forall S \in 2^{\mathcal{W}} : S \notin \mathcal{E}^d_i(w) \Longrightarrow S \notin \bigcup_{v \in \mathcal{W}: w\mathcal{K}_i v} \mathcal{E}^d_i(v) \quad (E^d_{KN})$$

Or simply by contraposition, ($E^d_{KN}$) can be rewritten as:

$$\forall w \in \mathcal{W}, \forall S \in 2^{\mathcal{W}} : \bigcup_{v \in \mathcal{W}: w\mathcal{K}_i v} \mathcal{E}^d_i(v) \subseteq \mathcal{E}^d_i(w) \quad (E^d_{KN})$$

Thus, we can deduce the following theorems, whose proofs are given in Appendix A.1:

**Proposition 4.1.** *If ($E^d_{KN}$) and ($E^d_{KP}$) hold at the same time, then it implies on the Kripke structure that the following property holds:*

$$\forall w \in \mathcal{W} : \mathcal{E}^d_i(w) = \bigcap_{v \in \mathcal{W}: w\mathcal{K}_i v} \mathcal{E}^d_i(v) = \bigcup_{v \in \mathcal{W}: w\mathcal{K}_i v} \mathcal{E}^d_i(v) \qquad (E^d_{KN}) + (E^d_{KP})$$

**Proposition 4.2.**

*(1) If ($E^d_{KP}$) holds and $\mathcal{K}_i$ is reflexive, then:*

$$\forall w \in \mathcal{W}, \mathcal{E}^d_i(w) = \bigcap_{v \in \mathcal{W}: w\mathcal{K}_i v} \mathcal{E}^d_i(v)$$

*(2) If ($E^d_{KN}$) holds and $\mathcal{K}_i$ is reflexive, then:*

$$\forall w \in \mathcal{W}, \mathcal{E}^d_i(w) = \bigcup_{v \in \mathcal{W}: w\mathcal{K}_i v} \mathcal{E}^d_i(v)$$

However considering for instance only the property (1) is not enough to verify that ($E^d_{KN}$) and ($E^d_{KP}$) hold at the same time. Thus, we can say that ($E^d_{KN}$) is stronger than considering an equivalence between intersections. Indeed, ($E^d_{KN}$) translates exactly that if an agent does not deliberate something, then it cannot be the case for all indistinguishable worlds by the epistemic relationship.

*4.2.4. Illustration of semantic relations*

Example 4.3 illustrates the fundamental differences between deliberate and non-deliberate effects. A deliberate effect is characterized as a calculated choice of an agent to achieve something, making the agent fully aware of the consequences, whereas a non-deliberate effect represents the set of all consequences of the actions of the agent which may be deliberate or not. Thus, the agent does not necessarily have knowledge of all these consequences.

**Example 4.3.** Suppose a situation in which there is a murderer $i$ that deliberately brings it about to kill a victim $j$. In order to represent this situation we consider three propositional variables: $p$ denotes that "the agent kills the victim by stabbing her", $q$ denotes that "the agent gets arrested by the police", and $r$ denotes that "there is no witness". For this situation, there are several possible worlds, for example, agent $i$ kills the victim and does not get stopped by the police, the victim was already dead before the agent stabs her, agent $i$ is remorseful and does not kill the victim, or the victim $j$ wakes up and makes agent $i$ run away, etc. Let us assume that a model $\mathcal{M}$ describes all of these possible worlds. Obviously we could consider a large number of possible worlds to describe this example, but for the sake of clarity, we consider only a small number of possible worlds. Let us assume $\mathcal{M}$ is such that:

- $\mathcal{W} = \{w, u, v, x, y, z, a\}$
- $V(p) = \{w, u, v\}, V(q) = \{v, y, z\}, V(r) = \{w, x, y, z, a\}$

The possible worlds are:

- $w$: "there is no witness, agent $i$ kills the victim $j$ and does not get arrested"
- $u$: "there is a witness and agent $i$ kills the victim $j$ but does not get arrested"
- $v$: "there is a witness and agent $i$ kills the victim $j$ and agent $i$ gets arrested"
- $x$: "the victim was already dead, no witness, and $i$ does not get arrested"
- $y$: "the victim was already dead, no witness, and agent $i$ gets arrested"
- $z$: "the killer is remorseful, no witness, and does not kill the victim, but agent $i$ still gets arrested for attempted murder"
- $a$: "no murder"

In the possible world $w$, the agent $i$ therefore has a deliberate effect to kill the victim. In addition, he also takes care that there is no witness. So the neighborhood function is such that $\mathcal{E}_i^d(w) = \{\{w, u, v\}, \{w, x, y, z, a\}\}$. It represents each independent strategy that agent $i$ intended to implement in $w$:

- $\{w, u, v\}$ represents the deliberate effect by agent $i$ to kill the victim;
- $\{w, x, y, z, a\}$ represents the deliberate effect by agent $i$ to ensure that there is no witness at the crime scene.

In the world $w$, agent $i$ successfully manages to kill the victim without getting caught, so $\mathcal{E}_i(w) = \{w\}$. Furthermore in $w$, since agent $i$ deliberately brings it about that there were no witness, we infer that the agent knows that $p$ and $r$ are true. So we have $\mathcal{K}_i(w) = \{w\}$, the only possible world where $p$ and $r$ are true simultaneously. The agent $i$ cannot discern in $w$ any other world than this one. Moreover in $w$, the agent knows that the police does not arrest him.

Let us remark that in this model semantics constraints, e.g. $\forall w \in \mathcal{W} : S \in \mathcal{E}_i^d(w) \Rightarrow \mathcal{E}_i(w) \subseteq S$, are naturally satisfied and illustrate the fact that since agent $i$ deliberately brings it about in $w$ to kill the victim, then this agent $i$ deliberately brings it about

that the victim is killed. Furthermore, since $||p|| \in \mathcal{E}_i^d(w)$ we have $\mathcal{M}, w \models E_i^d p$, i.e. agent $i$ deliberately brings it about to kill the victim. Since the agent's strategic choice was to succeed in killing the victim, he set up a strategy to make $p$ true.

### *4.3. Associated axiomatic system*

Given the constraints on our frame, the associated axiomatic system is given in Figure 4. Here, $\vdash \varphi$ means that $\varphi$ is a theorem. 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 $\vdash \varphi \Leftrightarrow \psi$, we infer $\vdash \Box\varphi \Leftrightarrow \Box\psi$. 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 $\vdash \varphi$, we infer $\vdash \Box\varphi$. Finally, we have duality ($DUAL$), i.e. for all $(\Box, \Diamond) \in \{(B_i, \langle B_i\rangle), (K_i, \langle K_i\rangle), (E_i, \langle E_i\rangle), (E_i^d, \langle E_i^d\rangle)\}$, $\vdash \Box\varphi \Leftrightarrow \neg\Diamond\neg\varphi$.

| | |
|---|---|
| $(\mathbf{PC})$ | All tautologies of Propositional Calculus |
| $(\mathbf{S4}_{K_i})$ | All S4-axioms for $K_i$ |
| $(\mathbf{4.2}_{K_i})$ | $\vdash \langle K_i\rangle K_i\varphi \Rightarrow K_i\langle K_i\rangle\varphi$ |
| $(\mathbf{KD45}_{B_i})$ | All KD45-axioms[6] for $B_i$ |
| $(\mathbf{S5}_{E_i})$ | All S5-axioms for $E_i$ |
| $(\mathbf{K}_i\mathbf{B}_i)$ | $\vdash K_i\varphi \Rightarrow B_i\varphi$ |
| $(\mathbf{4}_{K_i,B_i})$ | $\vdash B_i\varphi \Rightarrow K_iB_i\varphi$ |
| $(\mathbf{5}_{K_i,B_i})$ | $\vdash \neg B_i\varphi \Rightarrow K_i\neg B_i\varphi$ |
| $(\mathbf{E}_i^d\mathbf{E}_i)$ | $\vdash E_i^d\varphi \Rightarrow E_i\varphi$ |
| $(\mathbf{C}_{E_i^d})$ | $\vdash E_i^d\varphi \wedge E_i^d\psi \Rightarrow E_i^d(\varphi \wedge \psi)$ |
| $(\neg\mathbf{N}_{E_i^d})$ | $\vdash \neg E_i^d\top$ |
| $(\mathbf{4}_{K_i,E_i^d})$ | $\vdash E_i^d\varphi \Rightarrow K_iE_i^d\varphi$ |
| $(\mathbf{5}_{K_i,E_i^d})$ | $\vdash \neg E_i^d\varphi \Rightarrow K_i\neg E_i^d\varphi$ |

**Figure 4.** Axiomatic system KBE

From Figure 4, one may wonder if deliberate effect modality $E_i^d$ may be expressed with the non-deliberate modality $E_i$ and a combination of modalities $K_i$, $B_i$ or another modality expressing intention? It is not the case because, (1) as we consider a S4.2 system for knowledge, we do not have the negative introspection on $K_iE_i\varphi$ while we want it; (2) $B_iE_i\varphi$ does not satisfy the axiom T while $E_i^d$ satisfies it, which will be shown in Section 4.6; (3) an intention modality $I_i$ such has the one proposed by Sakama et al. (2015) will satisfy the necessitation on $I_iE_i\varphi$ which is a problem.

### *4.4. Soundness*

It is well known that the semantics of a normal modality of a S5 system that preserves validity is an equivalence relation (Blackburn et al., 2002). 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 as it is a S4.2 system. Concerning the inference rules between the modality $K_i$ and $B_i$, Stalnaker (2006) 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

[6]It corresponds to the axioms of the KD45-system for a $\Box$ modality, i.e. $(K)$ $\Box(\varphi \Rightarrow \psi) \Rightarrow (\Box\varphi \Rightarrow \Box\psi)$, $(D)$ $\Box\varphi \Rightarrow \neg\Box\neg\varphi$, (4) $\neg\Box\varphi \Rightarrow \Box\neg\Box\varphi$, and (5) $\Box\varphi \Rightarrow \Box\Box\varphi$ are considered.

section, we only focus on the non-normal properties associated with the neighborhood semantics $\mathcal{E}_i^d$. The following properties are in Pacuit (2017):

(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:

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

Other properties are standard to prove by using contraposition. Consequently it is straightforward to prove that our KBE system is sound. The complete proof is detailed in Appendix A.2.

**Theorem 4.4.** *The KBE system is sound.*

### *4.5. 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 (Pacuit, 2017).

**Theorem 4.5.** *The KBE system is complete.*

The complete proof can be found in Appendix A.3. Moreover the KBE system has the *deduction theorem*, is *strongly complete* and *strongly sound* (Blackburn et al., 2002; Pacuit, 2017). For the interested reader, the proofs are given in Appendix A.4.

### *4.6. Deductible theorems*

Let us remark that $(D)$ and $(T)$ can be derived for $E_i^d$. It means that one agent $i$ cannot deliberately bring about something and its opposite, and that when an agent $i$ deliberately brings about $\varphi$, he makes $\varphi$ to be true. In particular, it makes beliefs equivalent to knowledge for the special case of the deliberate effect modality. It means that an agent cannot have false beliefs concerning the formulas he deliberately brings about. It is given by the following theorem; the complete proofs are given in Appendix A.5.

**Theorem 4.6.**

*(1)* $\vdash \neg E_i^d \bot \quad (D_{E_i^d})$
*(2)* $\vdash E_i^d \varphi \Rightarrow \varphi \quad (T_{E_i^d})$
*(3)* $\vdash K_i E_i^d \varphi \Leftrightarrow E_i^d \varphi$
*(4)* $\vdash K_i \neg E_i^d \varphi \Leftrightarrow \neg E_i^d \varphi$
*(5)* $\vdash \neg K_i E_i^d \varphi \Leftrightarrow \neg E_i^d \varphi$
*(6)* $\vdash \neg K_i \neg E_i^d \varphi \Leftrightarrow E_i^d \varphi$
*(7)* $\vdash B_i E_i^d \varphi \Leftrightarrow E_i^d \varphi$
*(8)* $\vdash B_i \neg E_i^d \varphi \Leftrightarrow \neg E_i^d \varphi$
*(9)* $\vdash \neg B_i E_i^d \varphi \Leftrightarrow \neg E_i^d \varphi$
*(10)* $\vdash \neg B_i \neg E_i^d \varphi \Leftrightarrow E_i^d \varphi$

We can also show that any agent $i$, when it deliberately brings it about that another agent $j$ believes something, this agent $i$ also brings it about that the agent $j$ cannot believe that this agent $i$ can know the opposite. Such a theorem translates that when

an agent seeks to convey new beliefs, whether they are true or false in the case of lies, that agent always brings it about that he is credible to the other agent.

**Theorem 4.7.**

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

The complete proofs are given in Appendix A.5. Other interesting theorems can also be deduced. Moreover, when an agent deliberately brings it about that another agent believes something, then it cannot be the case that the agent deliberately brings it about the other to believe that a third-party agent can know the opposite.

**Theorem 4.8.**

*(1) An agent who deliberately influences the beliefs of an agent cannot have deliberated to show that other agents, including himself, may hold the opposite, i.e.:*

$$\vdash E_i^d B_j\varphi \Rightarrow \neg E_i^d B_j K_k\neg\varphi$$

*(2) An agent who deliberately brings it about that another agent does not believe a piece of information, cannot also deliberately bring it about that the agent knows that a third-party agent holds the opposite, i.e.:*

$$\vdash E_i^d\neg B_j\varphi \Rightarrow \neg E_i^d B_j K_k\varphi$$

As previously, the complete proofs are given in Appendix A.5. 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 $\vdash E_i^d\neg K_j\varphi \Rightarrow E_i\neg K_j K_k\varphi$ is also a theorem by following the same method as in 4.2. Moreover, as the contraposition of $\vdash K_k\varphi \Rightarrow B_k\varphi$ is $\vdash \neg B_k\varphi \Rightarrow \neg K_k\varphi$ and $\vdash E_i^d\varphi \Rightarrow E_i\varphi$ is $\vdash \neg E_i\varphi \Rightarrow \neg E_i^d\varphi$, we deduce two immediate corollaries to these theorems:

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

In the KBE system, we can also deduce the *qui facit per alium facit per se* principle, i.e. "he who acts through another does the act himself".

**Theorem 4.9** (Qui facit per alium facit per se)**.**

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

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

## 5. Modeling manipulations

In this section we first define formally what a manipulation is. Secondly we show our logical framework can also model coercion, persuasion and some forms of deception, and thus is consistent with "manipulation is not exactly coercion, not precisely persuasion, and not entirely similar to deception" (Handelman, 2009).

### *5.1. Different kinds of manipulation*

In terms of manipulation, a manipulator always intends to influence the intentions: by pushing his victim to do something, or by preventing his victim from doing something. It is the instrumentalization involved in Definition 2.1. Moreover, the manipulator always deliberately brings it about 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 brings it about to realize the manipulator's will; (3) how the manipulator intended to conceal. Hence, we consider *constructive manipulations* to be those where the manipulator brings about his victim to do 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 about another agent to do something in a deliberate way from doing it in an unintentional way. Thus, when the manipulator deliberately brings about the manipulated agent to deliberately bringing it about something, it relies on a *strong instrumentalization*. When the manipulator deliberately brings about the manipulated agent to do something in general, it relies on a *soft instrumentalization*. 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 from believing his intentions.

Tables 2 and 3 present different ways of expressing instrumentalization and concealment in the case of constructive manipulations and the case of destructive manipulations.

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

**Table 2.** Constructive forms of manipulation

Table 2 shows the different components of a constructive manipulation. For example, a *strong instrumentalization* is represented by the formula $E_i^d E_j^d \varphi$. Literally, this formula describes that the agent $i$ employed a strategy leading the agent $j$ deliberately bringing it about the consequence $\varphi$. A *soft instrumentalization* can be represented by the formula $E_i^d E_j \varphi$. Finally, in the case of constructive manipulations, an *epistemic concealment* can be represented by the formula $E_i^d \neg K_j E_i^d E_j \varphi$ and a *doxastic concealment* by the formula $E_i^d \neg B_j E_i^d E_j \varphi$.

Table 3 describes the different components when a manipulation is destructive. For example, in this case of destructive manipulations, *soft instrumentalization* is

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

**Table 3.** Destructive forms of manipulation

represented by the formula $E_i^d \neg E_j \varphi$, a *strong manipulation* is represented by the formula $E_i^d \neg E_j^d \varphi$, then an *epistemic concealment* by the formula $E_i^d \neg K_j E_i^d \neg E_j^d \varphi$ and finally, a *doxastic concealment* is represented by the formula $E_i^d \neg B_j E_i^d \neg E_j^d \varphi$.

Let us that notice that, for both constructive and destructive manipulations, an agent $i$ cannot manipulate another agent $j$ to deliberately brings about $\varphi$ while concealing $i$ deliberately brings $j$ about bringing about $\varphi$ as a side-effect (and inversely). This is due to the definition of manipulation we consider (see Definition 2.1): there is a manipulation when the effect of intrumentalization is the effect which is concealed. Thus, combining both a strong instrumentalization with the corresponding "soft" epistemic (or doxastic) and the soft instrumentalization with the corresponding "strong" epistemic (or doxastic) does not make sense.

### *5.1.1. The set of formulas* $\Sigma$

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. Since we use a non-normal modality for $E_i^d$ which does not have the theorem $\Box(\varphi \wedge \psi) \equiv \Box\varphi \wedge \Box\psi$, we have to consider all other possible formulas that this agent may deliberately bring about at the same time. Indeed, an agent can manipulate another one while deliberately bringing about something else. However, without this theorem, we cannot deduce manipulation alone in this situation. One way to deal with this problem is to consider an explicit set of formulas on which we (as designers) allow the logic to reason, and to use this set to define predicates for manipulation. However, this set must be finite, otherwise the formulas that will characterize manipulation cannot be well-formed.

Thus, we introduce a set of formulas $\Sigma$ which is finite, closed, and which contains $\mathcal{P} \cup \{\top, \bot\}$. We consider the closure as Blackburn et al. (2002): a set of formulas $\Sigma$ is said to be *closed* if, and only if, (1) if $\sigma \in \Sigma$ and $\theta$ is a subformula of $\sigma$, then $\theta \in \Sigma$ and (2) if $\sigma \in \Sigma$ and $\sigma$ is not of the form $\neg\theta$, then $\neg\sigma \in \Sigma$. While $\Sigma$ may be arbitrarily instantiated, it makes sense to ground it on a action-effect knowledge base: $\Sigma$ shall be the closed set of all subformulas of this knowledge base. In order to manipulate the predicates that will use in the next sections a $\Sigma$-set, we introduce a new operator $\oslash$ to define a conjunctive Cartesian product between two sets of formulas, i.e. $\Sigma \oslash \Gamma := \{\sigma \wedge \gamma : \sigma \in \Sigma, \gamma \in \Gamma\}$. This operator will allow us to rewrite these predicates in order to highlight logical relations between related concepts, as for instance between strong instrumentalization and the persuasion predicate.

### *5.1.2. Soft constructive manipulations*

A *soft constructive manipulation with epistemic concealment*, denoted $MCEK_{i,j}^{\Sigma}(\varphi)$, is when a manipulator deliberately brings the victim about doing something while

making sure that the victim does not know the deliberate effects of the manipulator. Here, $\varphi \in \Sigma$ and must be consistent. A *soft constructive manipulation with doxastic concealment* – denoted $MCEB^{\Sigma}_{i,j}(\varphi)$ below – is defined similarly but, in this case, the manipulator deliberately brings it about that the victim does not believe his deliberate effects. Formally, we define these manipulation forms such as:

$$MCEK^{\Sigma}_{i,j}(\varphi) \stackrel{\triangle}{=} \bigvee_{\psi \in \Sigma} E^d_i(E_j\varphi \wedge \neg K_j E^d_i E_j\varphi \wedge \psi)$$

$$MCEB^{\Sigma}_{i,j}(\varphi) \stackrel{\triangle}{=} \bigvee_{\psi \in \Sigma} E^d_i(E_j\varphi \wedge \neg B_j E^d_i E_j\varphi \wedge \psi)$$

Let us notice that $\psi$ represents all formulas from $\Sigma$ which do not contradict $E_j\varphi \wedge \neg K_j E^d_i E_j\varphi$. Indeed, if $\psi$ contradicts $E_j\varphi \wedge \neg K_j E^d_i E_j\varphi$, then we immediately deduce that $E^d_i\bot$. However it is necessarily false due to the theorem $\vdash \neg E^d_i\bot$.

### *5.1.3. Strong constructive manipulations*

A *strong constructive manipulation with epistemic concealment* is denoted with $MCE^dK^{\Sigma}_{i,j}(\varphi)$, and is when the manipulator 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 $MCE^dB^{\Sigma}_{i,j}(\varphi)$ below – is similar but, in this case, the manipulator makes sure that the victim does not believe his effects. Formally,

$$MCE^dK^{\Sigma}_{i,j}(\varphi) \stackrel{\triangle}{=} \bigvee_{\psi \in \Sigma} E^d_i(E^d_j\varphi \wedge \neg K_j E^d_i E^d_j\varphi \wedge \psi)$$

$$MCE^dB^{\Sigma}_{i,j}(\varphi) \stackrel{\triangle}{=} \bigvee_{\psi \in \Sigma} E^d_i(E^d_j\varphi \wedge \neg B_j E^d_i E^d_j\varphi \wedge \psi)$$

**Example 5.1.** Let us illustrate this predicate with an example related to advertising. An advertiser always intends to bring new customers about buying a product. These intentions are not concealed and therefore we cannot speak of manipulation. On the other hand, it becomes manipulation when the advertiser uses concealed sales techniques such as, for example, the use of subliminal images. Thus, the advertiser does not seek to conceal his intention to bring the customer about buying the product, but to deliberately conceal the technique he uses to bring the customer about buying. If the agent $i$ is the advertiser, $E^d_i$ represents his strategy of using subliminal images to get the customer $j$ to buy a product. The customer does not know that the advertiser has used these images. Thus, if $\varphi :=$ "the product is bought", then this situation of manipulation is fully described by the formula $E^d_i(E_j\varphi \wedge \neg K_j E^d_i E_j\varphi)$.

### *5.1.4. 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,

$$MDEK^{\Sigma}_{i,j}(\varphi) \triangleq \bigvee_{\psi\in\Sigma} E^d_i(\neg E_j\varphi \wedge \neg K_j E^d_i \neg E_j\varphi \wedge \psi)$$

$$MDE^dK^{\Sigma}_{i,j}(\varphi) \triangleq \bigvee_{\psi\in\Sigma} E^d_i(\neg E^d_j\varphi \wedge \neg K_j E^d_i \neg E^d_j\varphi \wedge \psi)$$

$$MDEB^{\Sigma}_{i,j}(\varphi) \triangleq \bigvee_{\psi\in\Sigma} E^d_i(\neg E_j\varphi \wedge \neg B_j E^d_i \neg E_j\varphi \wedge \psi)$$

$$MDE^dB^{\Sigma}_{i,j}(\varphi) \triangleq \bigvee_{\psi\in\Sigma} E^d_i(\neg E^d_j\varphi \wedge \neg B_j E^d_i \neg E^d_j\varphi \wedge \psi)$$

**Example 5.2.** An example of destructive manipulation is illustrated by the case of eclipse attacks (Singh, Ngan, Druschel, & Wallach, 2006). These attacks consist in isolating an agent in order to exclude it from a network. This type of manipulation is captured by destructive manipulation. Indeed, the hacker makes sure at the moment of the attack that the target node can no longer communicate with other nodes in the network (i.e. $E^d_i\neg E_j\varphi$ with $\varphi$ any communicable information) while making sure that he does not know that he is at that moment under the instance of an attack (i.e. $E^d_i\neg K_j E^d_i\neg E_j\varphi$).

#### *5.1.5. A general definition of manipulation*

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

$$M^{\Sigma}_{i,j}(\varphi) \triangleq \bigvee_{\Box\in\{B,K\}} MCE\Box^{\Sigma}_{i,j}(\varphi) \vee MCE^d\Box^{\Sigma}_{i,j}(\varphi) \vee MDE\Box^{\Sigma}_{i,j}(\varphi) \vee MDE^d\Box^{\Sigma}_{i,j}(\varphi)$$

### *5.2. Some properties of manipulation*

In this section, we exhibit some properties of the above definitions.

#### *5.2.1. Believing being influenced*

Obvioulsy, if an agent is a victim of an epistemic (resp. doxastic) manipulation, he cannot know (resp. believe) he is instrumentalized. However, an agent can believe he is instrumentalized while being a victim of an epistemic manipulation. It means that the victim may believe to be instrumentalized while being unable to prove it, i.e. it is not the case the agent knows he is instrumentalized. Indeed, when a child cries to get a new toy from his parents, the parents may believe the child deliberately puts himself in this state but cannot be sure it is the case. It is given by the following theorem for the soft epistemic constructive manipulation.

**Theorem 5.3.** $\not\models MCEK^{\Sigma}_{i,j}(\varphi) \wedge B_j E^d_i E_j \varphi \Rightarrow \bot$

This property is obvious because there is no contradiction between $\neg K_j E^d_i E_j \varphi$ and $B_j E^d_i E_j \varphi$. The same property holds for the other forms of manipulation with epistemic concealment, i.e. strong epistemic and destructive manipulation.

*5.2.2. Knowing being influenced*

Interestingly, an agent can be victim of a manipulation while knowing he is influenced. In this case, the agent does not know (resp. believe) he is instrumentalized, i.e. the manipulator deliberately brings the victim about bringing about something, but he can know the manipulator influenced him. Indeed, the manipulation conceals the deliberate intention of the the manipulator, but not the influence in itself. It is given by the following theorem.

**Theorem 5.4.** $\not\models MCEK^{\Sigma}_{i,j}(\varphi) \wedge K_j E_i E_j \varphi \Rightarrow \bot$

This property is obvious as there is no contradiction between $\neg K_j E^d_i E_j \varphi$ and $K_j E_i E_j \varphi$. The same property holds for the other forms of manipulation. Let us consider the previous example: when the child cries to get a new toy from his parents, the parents can know the cries influence them while not knowing the child deliberately puts himself in this state to affect them.

*5.2.3. Knowing or believing the effects of a manipulation*

If an agent is strongly manipulated in a constructive way to bring himself about a formula $\varphi$, he knows that $\varphi$, and cannot know or believe $\neg\varphi$. Indeed, while a manipulation conceals the instrumentalization, it does not conceal its effect, i.e. the fact that the victim brings about $\varphi$.

**Theorem 5.5.**

*(1)* $\vdash (MCE^d K^{\Sigma}_{i,j}(\varphi) \vee MCE^d B^{\Sigma}_{i,j}(\varphi)) \wedge K_j \neg\varphi \Rightarrow \bot$
*(2)* $\vdash (MCE^d K^{\Sigma}_{i,j}(\varphi) \vee MCE^d B^{\Sigma}_{i,j}(\varphi)) \wedge B_j \neg\varphi \Rightarrow \bot$
*(3)* $\vdash (MCE^d K^{\Sigma}_{i,j}(\varphi) \vee MCE^d B^{\Sigma}_{i,j}(\varphi)) \Rightarrow K_j \varphi$

The proofs are given in Appendix A.6. Let us notice that those properties do not hold for the soft or destructive manipulations.

*5.2.4. Interactions between deliberate and non-deliberate effects*

When the goal $\varphi$ of a manipulation consists in bringing the victim about another formula $\varphi'$, some form of manipulation may be reduced while some others allow us to express particular situations depending if it is a soft or strong manipulation.

In the case of soft manipulation:

**Theorem 5.6.**

*(1)* $\vdash MCEK^{\Sigma}_{i,j}(E_j \varphi) \Leftrightarrow MCEK^{\Sigma}_{i,j}(\varphi)$
*(2)* $\vdash MCEK^{\Sigma}_{i,j}(\neg E_j \varphi) \Leftrightarrow MDEK^{\Sigma}_{i,j}(\varphi)$
*(3)* $\vdash MCEB^{\Sigma}_{i,j}(E_j \varphi) \Leftrightarrow MCEB^{\Sigma}_{i,j}(\varphi)$
*(4)* $\vdash MCEB^{\Sigma}_{i,j}(\neg E_j \varphi) \Leftrightarrow MDEB^{\Sigma}_{i,j}(\varphi)$

*(5)* $\vdash MDEK^{\Sigma}_{i,j}(E_j\varphi) \Leftrightarrow MDEK^{\Sigma}_{i,j}(\varphi)$
*(6)* $\vdash MDEK^{\Sigma}_{i,j}(\neg E_j\varphi) \Leftrightarrow MCEK^{\Sigma}_{i,j}(\varphi)$
*(7)* $\vdash MDEB^{\Sigma}_{i,j}(E_j\varphi) \Leftrightarrow MDEB^{\Sigma}_{i,j}(\varphi)$
*(8)* $\vdash MDEB^{\Sigma}_{i,j}(\neg E_j\varphi) \Leftrightarrow MCEB^{\Sigma}_{i,j}(\varphi)$

The previous properties obviously hold because we have $\models E_iE_i\varphi \equiv E_i\varphi$, $\models \neg E_iE_i\varphi \equiv \neg E_i\varphi$, and $\models E_j\neg E_j\varphi \Leftrightarrow \neg E_j\varphi$ i.e. $\models \neg E_j\neg E_j\varphi \Leftrightarrow E_j\varphi$. Interestingly such properties do not hold for deliberate effects, i.e. when $\varphi = E^d_j\varphi'$. However, it allows us to express manipulation where the goal is to let a victim obliges or forbids herself by side-effect to apply a given strategy, e.g. to make a mistake that obliges or forbids herself to deliberately bring her about something. For instance for constructive manipulation, such situations are expressed by:

- $E^d_i(E_jE^d_j\varphi' \wedge \neg K_jE^d_iE_jE^d_j\varphi')$
- $E^d_i(E_jE^d_j\varphi' \wedge \neg B_jE^d_iE_jE^d_j\varphi')$
- $E^d_i(E_j\neg E^d_j\varphi' \wedge \neg K_jE^d_iE_j\neg E^d_j\varphi')$
- $E^d_i(E_j\neg E^d_j\varphi' \wedge \neg B_jE^d_iE_j\neg E^d_j\varphi')$

In the case of strong manipulation, if we consider $\varphi = E_j\varphi'$, we can express manipulation where the goal is to instrumentalize a victim to make her to insure producing (or not producing) something by side-effects, e.g. to jeopardize (or insuring not to jeopardize) another strategy. For instance for constructive manipulation, such situations are expressed by:

- $E^d_i(E^d_jE_j\varphi' \wedge \neg K_jE^d_iE^d_jE_j\varphi')$
- $E^d_i(E^d_jE_j\varphi' \wedge \neg B_jE^d_iE^d_jE_j\varphi')$
- $E^d_i(E^d_j\neg E_j\varphi' \wedge \neg K_jE^d_iE^d_j\neg E_j\varphi')$
- $E^d_i(E^d_j\neg E_j\varphi' \wedge \neg B_jE^d_iE^d_j\neg E_j\varphi')$

Finally, we have the case where $\varphi = E^d_j\varphi'$. In this case, it correspond to manipulation where the goal is to instrumentalize a victim in order she puts herself a situation where she deliberately brings about something or deliberately forbid herself to bring about something, e.g. pushing a drug-addict to attach himself in order to not be able to take drug. For instance for constructive manipulation, such situations are expressed by:

- $E^d_i(E^d_jE^d_j\varphi' \wedge \neg K_jE^d_iE^d_jE^d_j\varphi')$
- $E^d_i(E^d_jE^d_j\varphi' \wedge \neg K_jE^d_iE^d_jE^d_j\varphi')$
- $E^d_i(E^d_j\neg E^d_j\varphi' \wedge \neg K_jE^d_iE^d_j\neg E^d_j\varphi')$
- $E^d_i(E^d_j\neg E^d_j\varphi' \wedge \neg B_jE^d_iE^d_j\neg E^d_j\varphi')$

*5.2.5. Interactions between soft and strong manipulation*

An agent can strongly manipulate another one while not softly manipulating the victim. Conversely, an agent can softly manipulate another one while not strongly manipulating the victim. Obviously, an agent can both softly and strongly manipulate another one.

**Theorem 5.7.**

*(1)* $\not\models \neg MCEK^{\Sigma}_{i,j}(\varphi) \wedge MCE^dK^{\Sigma}_{i,j}(\varphi) \Rightarrow \bot$
*(2)* $\not\models MCEK^{\Sigma}_{i,j}(\varphi) \wedge \neg MCE^dK^{\Sigma}_{i,j}(\varphi) \Rightarrow \bot$

*(3)* $\not\models MCEK^{\Sigma}_{i,j}(\varphi) \wedge MCE^{d}K^{\Sigma}_{i,j}(\varphi) \Rightarrow \bot$

Those properties can be extended to the other forms of manipulation. While it may be seen as counterintuitive, it expressed some particular situations. For instance, if $\neg MCEK^{\Sigma}_{i,j}(\varphi) \wedge MCE^{d}K^{\Sigma}_{i,j}(\varphi)$ is true, then it means that the manipulator aims at concealing the way he instrumentalized the victim, i.e. he brings the victim to *deliberately* bring about something, while not concealing the fact that he brings the victim about something (e.g. by side-effects).

#### *5.2.6. Manipulating oneself*

As defined in Section 2.2, manipulation consists in instrumentalizing *another agent*. However, the KBE system only partially prevents an agent to manipulate himself as shown by the following properties.

**Theorem 5.8.**

*(1)* $\vdash E^{d}_{i}E_{i}\varphi \wedge E^{d}_{i}\neg K_{i}E^{d}_{i}E_{i}\varphi \Rightarrow \bot$
*(2)* $\not\models MCEK^{\Sigma}_{i,i}(\varphi) \Rightarrow \bot$ *and* $\not\models MCEB^{\Sigma}_{i,i}(\varphi) \Rightarrow \bot$

Both properties can be extended to the other forms of manipulation. Property (1) holds because of $\models \neg K_{i}E^{d}_{i}\varphi \Leftrightarrow \neg E^{d}_{i}\varphi$ while properties in (2) are due to $\not\models E^{d}_{i}(\varphi \wedge \psi) \implies E^{d}_{i}\varphi \wedge E^{d}_{i}\psi$. Property (1) means that manipulating oneself by independently deliberately bringing oneself about each part of the manipulation is inconsistent. However, properties in (2) mean that an agent can manipulate himself when the strategy of concealment is intrinsic to the strategy of instrumentalization. It makes particularly sense for artificial agents, e.g. in the case where a hacker changes the normal behavior of an artificial agent to delete or prevent log recording.

### *5.3. Other notions related to manipulation*

We can also express related notions like coercion, persuasion and deception. We exhibit some links between those notions and manipulation. As previously, the proofs are given in Appendix A.6.

#### *5.3.1. Coercion*

The coercion is an influence of an agent over another agent by means of pressure without any dissimulation. For instance, 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 or believes that he is under pressure (by pointing the gun). As manipulation, coercion can be expressed in a constructive or a destructive form. Thus, coercion can be defined formally with several predicates, i.e. constructive epistemic coercion, constructive doxastic coercion, destructive epistemic coercion and destructive doxastic coercion:

$$CKCoe^{\Sigma}_{i,j}(\varphi) \stackrel{\triangle}{=} \bigvee_{\psi \in \Sigma} E^{d}_{i}(E^{d}_{j}\varphi \wedge K_{j}E^{d}_{i}E^{d}_{j}\varphi \wedge \psi)$$

$$CBCoe_{i,j}^{\Sigma}(\varphi) \stackrel{\triangle}{=} \bigvee_{\psi \in \Sigma} E_i^d(E_j^d\varphi \wedge B_j E_i^d E_j^d \varphi \wedge \psi)$$

$$DKCoe_{i,j}^{\Sigma}(\varphi) \stackrel{\triangle}{=} \bigvee_{\psi \in \Sigma} E_i^d(\neg E_j^d\varphi \wedge K_j E_i^d \neg E_j^d \varphi \wedge \psi)$$

$$DBCoe_{i,j}^{\Sigma}(\varphi) \stackrel{\triangle}{=} \bigvee_{\psi \in \Sigma} E_i^d(\neg E_j^d\varphi \wedge B_j E_i^d \neg E_j^d \varphi \wedge \psi)$$

Obviously, coercion is a particular case of strong instrumentalization. Let us notice the following theorems use the $\oslash$ operator defined in Section 5.1.1 in order to better highlight some interesting properties. We recall the reader this operator is a conjunctive Cartesian product between two sets of formulas.

**Theorem 5.9.**

*(1)* $\vdash CKCoe_{i,j}^{\Sigma}(\varphi) \Leftrightarrow \bigvee_{\psi \in \Sigma \oslash \{K_j E_i^d E_j^d \varphi\}} E_i^d(E_j^d\varphi \wedge \psi)$

*(2)* $\vdash CBCoe_{i,j}^{\Sigma}(\varphi) \Leftrightarrow \bigvee_{\psi \in \Sigma \oslash \{B_j E_i^d E_j^d \varphi\}} E_i^d(E_j^d\varphi \wedge \psi)$

*5.3.2. Persuasion*

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

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

Persuasion is view as a particular case of influence, i.e. an influence on the belief state of the persuadee. Let us notice that, as the KBE system is not based on dynamic logic, this notion of persuasion does not capture the case of belief revision. However, if the persuadee is persuaded to deliberately bring about something, then the persuasion is equivalent to strong instrumentalization. It is expressed by the following theorem.

**Theorem 5.10.**

*(1)* $\vdash per_{i,j}^{\Sigma}(\neg E_j^d\varphi) \Leftrightarrow \bigvee_{\psi \in \Sigma} E_i^d(\neg E_j^d\varphi \wedge \psi)$

*(2)* $\vdash per_{i,j}^{\Sigma}(E_j^d\varphi) \Leftrightarrow \bigvee_{\psi \in \Sigma} E_i^d(E_j^d\varphi \wedge \psi)$

When an agent persuades another one to deliberately bring about something while ensuring that this agent knows or believes that he is persuaded, we can deduce a coercion.

**Theorem 5.11.**

*(1)* $\vdash CKCoe_{i,j}^{\Sigma}(\varphi) \Leftrightarrow per_{i,j}^{\Sigma \oslash \{K_j E_i^d E_j^d \varphi\}}(E_j^d\varphi)$

*(2)* $\vdash CBCoe^{\Sigma}_{i,j}(\varphi) \Leftrightarrow per^{\Sigma \oslash \{B_j E^d_i E^d_j \varphi\}}_{i,j}(E^d_j \varphi)$
*(3)* $\vdash DKCoe^{\Sigma}_{i,j}(\varphi) \Leftrightarrow per^{\Sigma \oslash \{K_j E^d_i \neg E^d_j \varphi\}}_{i,j}(\neg E^d_j \varphi)$
*(4)* $\vdash DBCoe^{\Sigma}_{i,j}(\varphi) \Leftrightarrow per^{\Sigma \oslash \{B_j E^d_i \neg E^d_j \varphi\}}_{i,j}(\neg E^d_j \varphi)$

Interestingly, when an agent persuades another agent to bring about something while concealing the persuader deliberate intention, we can deduce a strong manipulation.

**Theorem 5.12.**

*(1)* $\vdash per^{\Sigma \oslash \{\neg K_j E^d_i E^d_j \varphi\}}_{i,j}(E^d_j \varphi) \Leftrightarrow MCE^d K^{\Sigma}_{i,j}(\varphi)$
*(2)* $\vdash per^{\Sigma \oslash \{\neg B_j E^d_i E^d_j \varphi\}}_{i,j}(E^d_j \varphi) \Leftrightarrow MCE^d B^{\Sigma}_{i,j}(\varphi)$
*(3)* $\vdash per^{\Sigma \oslash \{\neg K_j E^d_i \neg E^d_j \varphi\}}_{i,j}(\neg E^d_j \varphi) \Leftrightarrow MDE^d K^{\Sigma}_{i,j}(\varphi)$
*(4)* $\vdash per^{\Sigma \oslash \{\neg B_j E^d_i \neg E^d_j \varphi\}}_{i,j}(\neg E^d_j \varphi) \Leftrightarrow MDE^d B^{\Sigma}_{i,j}(\varphi)$

Let us notice the previous properties does not hold for soft manipulation.

### *5.3.3. 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 (Sakama et al., 2015). Let us 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). Interestingly, both cases can be defined as a special case of persuasion.

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 (Aggarwal & Wu, 2006). Thus, it can be characterized by the fact that an agent makes sure to conceal – from an epistemic (resp. doxastic) point-of-view – his deliberate effects to make someone believes something. Thus, we can define epistemic (resp. doxastic) source concealment as follows:

$$KSoCo^{\Sigma}_{i,j}(\varphi) \triangleq \bigvee_{\psi \in \Sigma} E^d_i(B_j\varphi \wedge \neg K_j E^d_i B_j \varphi \wedge \psi)$$

$$BSoCo^{\Sigma}_{i,j}(\varphi) \triangleq \bigvee_{\psi \in \Sigma} E^d_i(B_j\varphi \wedge \neg B_j E^d_i B_j \varphi \wedge \psi)$$

Mahon 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)" (Mahon, 2008). Thus, a statement is a lie if there is also a deliberate effect to conceal the intended effects to lie, from an epistemic or a doxastic point-of-view. We call such a lie an epistemic (reps. doxastic) credible lie:

$$KCrLie^{\Sigma}_{i,j}(\varphi) \stackrel{\triangle}{=} B_i\neg\varphi \wedge (\bigvee_{\psi\in\Sigma} E^d_i(B_j\varphi \wedge \neg K_j B_i\neg\varphi \wedge \psi))$$

$$BCrLie^{\Sigma}_{i,j}(\varphi) \stackrel{\triangle}{=} B_i\neg\varphi \wedge (\bigvee_{\psi\in\Sigma} E^d_i(B_j\varphi \wedge \neg B_j B_i\neg\varphi \wedge \psi))$$

One property of credible lie is that an agent cannot lie to another one in order to make the victim believes he deliberately brings about something. Indeed, due to Theorem 5.10(1), believing to deliberately bring about $\varphi$ implies deliberately bringing about $\varphi$, and the persuader would believe it while he should not. It is given by the next theorem.

**Theorem 5.13.**

*(1)* $\vdash KCrLie^{\Sigma}_{i,j}(E^d_j\varphi) \Rightarrow \bot$
*(2)* $\vdash BCrLie^{\Sigma}_{i,j}(E^d_j\varphi) \Rightarrow \bot$
*(3)* $\vdash KSoCo^{\Sigma}_{i,j}(E^d_j\varphi) \Leftrightarrow MCE^d K^{\Sigma}_{i,j}(\varphi)$
*(4)* $\vdash BSoCo^{\Sigma}_{i,j}(E^d_j\varphi) \Leftrightarrow MCE^d B^{\Sigma}_{i,j}(\varphi)$
*(5)* $\vdash KCrLie^{\Sigma}_{i,j}(\varphi) \Rightarrow Per^{\Sigma \oslash \{\neg K_j B_i\neg\varphi\}}_{i,j}(\varphi)$
*(6)* $\vdash BCrLie^{\Sigma}_{i,j}(\varphi) \Rightarrow Per^{\Sigma \oslash \{\neg B_j B_i\neg\varphi\}}_{i,j}(\varphi)$

We can compare the persuasion and deception definitions we propose here to the literature presented in Section 3.1. Firstly, we do not need to introduce explicit communication modality as Sakama et al. (2015) did. Here, the communication is expressed (more precisely reduced) to a deliberate effect of having an influence on the mental state of another agent, e.g. making the agent believing something, or making him bringing about something. Obviously, communication in a whole cannot be reduced to just that, e.g. when an agent asks a question, or apologize, or publicly commit himself to do something, etc. Our notion of persuasion is syntactically close to the one proposed by Bonnet et al. (2021) – deliberately seeing to it that the persuadee believes something – except we do not consider a temporal modality. Our notion of deception is however different as we explicitly consider concealment, which is not the case in Bonnet et al. (2021). The same remark holds for lie when compared to the one proposed by Sakama et al. (2015).

## 6. Application of KBE

The purpose of this section is to instantiate the KBE system in a situation where it is possible for one agent to manipulate another.

### *6.1. The story*

We consider an e-commerce website in which two agents perform a commercial transaction. Let $i$ be the seller and $j$ be the customer, and $i$ says to $j$: "You can trust me on the quality of the product. You will not find a better product anywhere else. You are

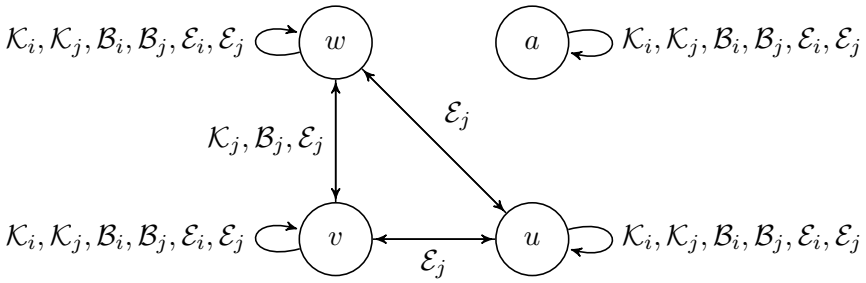


**Figure 5.** Mental states of the agents

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 in the theory of free will compliance. Indeed, 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 (Joule, Girandola, & Bernard, 2007).

### *6.2. Variables and possible worlds*

To represent this situation, we consider two propositional variables $p$ and $q$:

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

For the sake of readability, we do not consider all possible scenarios such as agent $j$ does not buy the product but trusts $i$ on the product quality. We represent the possible scenarios we consider as a set of possible worlds $\mathcal{W} = \{a, w, v, u\}$ where:

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

### *6.3. Defining the set* $\Sigma$

Let us define the set $\Sigma$ (see Section 5.1.1). Here, $\Sigma = Cl(\Gamma)$ be a finite and closed set of formulas where:

$$\begin{aligned}\Gamma = \{&E_i^d(E_i^d p \Rightarrow E_i^d E_j q),\\ &E_i^d E_j q \wedge E_i^d \neg K_j E_i^d E_j q \Rightarrow E_i^d(E_j q \wedge \neg K_j E_i^d E_j q),\\ &E_i^d E_j^d q \wedge E_i^d K_j E_i^d E_j^d q \Rightarrow E_i^d(E_j^d q \wedge K_j E_i^d E_j^d q),\\ &\top, \bot\}\end{aligned}$$

Hence for example, the formula $E_i^d(E_i^d p \Rightarrow E_i^d E_j q)$ allows us 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. As another example, 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 us to deduce a *soft constructive manipulation* if agent $i$ manipulates agent $j$. 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 if there is coercion in a world. Furthermore, as we consider the closure of $\Gamma$, we also express all subformulas and their single negation[7].

### *6.4. The KBE model*

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 5 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) $\mathcal{E}_j^d(w) = \{\{w, u, v\}\}$, $\mathcal{E}_j^d(v) = \{\{w, u, v\}\}$, $\mathcal{E}_j^d(u) = \{\{w, u, v\}\}$, $\mathcal{E}_j^d(a) = \{\{w, a\}\}$

Informally:

(1) describes the fact that the agent $i$ knows agent $j$ trusts him about the product quality and knows that agent $j$ knows he buys a product and trusts $i$ about the product quality, if it is the case.
(2) describes the agent $j$ that buy the product while trusting the agent $j$ does not distinguish the worlds where he is instrumentalized and where he is not.
(3) and (4) describe that the agents believe what they know and vice versa, i.e. $\mathcal{K}_i = \mathcal{B}_i$ and $\mathcal{K}_j = \mathcal{B}_j$.
(5) means that in the world $w$, agent $i$ ensures that $p$ and $q$.
(6) says that agent $j$ buys the product in $\{w, u, v\}$ but does not necessarily trust $i$ on the product quality.
(7) means the agent $i$ in $w$ deliberately brings agent $j$ about trusting him and he deliberately brings it about if agent $j$ trusts him, then agent $j$ buys the product while $i$ makes sure to hide his strategy to get $j$ to buy the product.
(8) means the agent $j$ in $\{w, u, v\}$ only intended to buy the product.

### *6.5. Deductions*

We give now some deductions. We show according to the given KBE model that there is a soft instrumentalization, a deliberate concealment in $w$, a possible manipulation in $w$ and a coercion in the world $u$.

#### *6.5.1. Soft instrumentalization in w*

Let us notice that in $w$, this model expresses that agent $i$ has deliberately brought $j$ about buying the product by building trust. Indeed, we have $||E_i^d p \Rightarrow E_i^d E_j q|| =$

[7] A set of formulas $\Sigma$ is *closed under single negation* iff if $\sigma \in \Sigma$ and $\sigma$ is not of the form $\neg\theta$, then $\neg\sigma \in \Sigma$.

$\{w, u, a\}$ 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 $\vdash E_i^d \varphi \Rightarrow \varphi$, 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 $\mathcal{M}, w \models E_i^d E_j q$.

*6.5.2. Deliberate concealment in w*

In addition, we can notice that in $w$ agent $i$ also ensures to hide his strategy to get $j$ to buy the product. Indeed, we have in $v$ that $\mathcal{M}, v \models \neg E_i^d E_j q$, since $||E_j q|| = \{w, u, v\}$ and $\{w, u, v\} \notin \mathcal{E}_i^d(v)$. Thus, since 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 $||\neg K_j E_i^d E_j q|| = \{w, v, a\}$[8] 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 $\mathcal{M}, w \models E_i^d E_j q \wedge E_i^d \neg K_j E_i^d E_j q$.

*6.5.3. Possible manipulation in w*

Now, by the tautology $\models E_i^d \varphi \wedge E_i^d \psi \Rightarrow E_i^d(\varphi \wedge \psi)$, 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 agent $i$ manipulates 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 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 dissimulation strategy 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 $\models E_i^d \varphi \wedge E_i^d \psi \Rightarrow E_i^d(\varphi \wedge \psi)$. These sets of possible worlds reflect the fact that when an agent sets up different plans, this agent also considers all combinations of all the different plans as a possible plan. For example, since $\{w, a\} = \{w, u, a\} \cap \{w, v, a\}$, $\{w, a\}$ is the combination of the respective plans $||E_i^d p \Rightarrow E_i^d E_j q||$ and $||\neg K_j E_i^d E_j q||$.

*6.5.4. Coercion in the world u*

Finally let us notice that in the world $u$, agent $i$ coerced agent $j$ to push him to buy the product. Indeed, since we have $||E_j^d q|| = \{w, u, v\}$ and $||K_j E_i^d E_j^d q|| = \{u\}$[9] 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 $\mathcal{M}, u \models coe_{i,j}^{\Sigma}(q)$.

---

[8] 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 $u \notin ||\neg K_j E_i^d E_j q||$.

[9] 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 $x = u$.

## 7. Conclusion and future works

In this article, we provide a broad state-of-the-art about manipulation in social science. We used this state-of-the-art to define manipulation as the deliberate effect to instrumentalize a victim while making sure to conceal that effect. We proposed a logical framework to reason about this notion. To this end, we defined a new deliberate BIAT modality for deliberate effects. We then considered logical interactions between knowledge, belief, deliberate effects and consequences of actions and proved that our system is strongly sound and complete. Furthermore we deduced 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, future works are twofold. Firstly, it may be of interest to formalize deliberated intention in a deeper way. Indeed, KBE neither consider temporal nor dynamic operators while they are necessary to express actions and their consequences. Secondly, expressing more kind of manipulation is of interest. For instance, it should be interesting to extend the framework with awareness so as to define new manipulation forms with *awareness concealment*. As another example, extending KBE with other mental state can be of interest to model manipulation strategies such as presented in Table 1.

### *7.1. Extending KBE to awareness logics*

When we defined manipulation such as the deliberate effect to instrumentalize an agent while making sure to conceal that effect, we defined concealment as a lack of knowledge or belief about the manipulator's effects. However, in many situations, a manipulated agent is unaware that he or she has been manipulated. In section 3.3, we have presented work on the representation of awareness. One perspective is to integrate awareness into manipulation.

We could for instance consider a modality $A_i$ to represent that agent $i$ is aware of something, as proposed in Schipper (2014). Let $\Sigma$ be a set of formulas that is closed and finite such as $\{\top, \bot\} \subseteq \Sigma$, $\varphi \in \Sigma$ and $\mathcal{A}_i : \mathcal{W} \to 2^{\Sigma}$ be the awareness function associated with the modality $A_i$. We could then define new forms of manipulations with absence of awareness. Thus, a *soft constructive manipulation with absence of awareness* would be defined as:

$$MCEA^{\Sigma}_{i,j}(\varphi) \triangleq \bigvee_{\psi \in \Sigma} E^d_i(E_j\varphi \wedge \neg A_j E^d_i E_j \varphi \wedge \psi)$$

A *strong constructive manipulation with absence of awareness* would be defined as:

$$MCE^dA^{\Sigma}_{i,j}(\varphi) \triangleq \bigvee_{\psi \in \Sigma} E^d_i(E^d_j\varphi \wedge \neg A_j E^d_i E^d_j \varphi \wedge \psi)$$

A *soft destructive manipulation with absence of awareness* would be defined as:

$$MDEA^{\Sigma}_{i,j}(\varphi) \triangleq \bigvee_{\psi \in \Sigma} E^d_i(\neg E_j\varphi \wedge \neg A_j E^d_i \neg E_j \varphi \wedge \psi)$$

A *strong destructive manipulation with absence of awareness* would be defined as:

$$MDE^d A^{\Sigma}_{i,j}(\varphi) \triangleq \bigvee_{\psi \in \Sigma} E^d_i(\neg E^d_j \varphi \wedge \neg A_j E^d_i \neg E^d_j \varphi \wedge \psi)$$

Furthermore, note that in this case the function $\mathcal{A}_i$ is such that for any $w \in \mathcal{W}$, $\mathcal{A}_i(w) \subseteq \Sigma$. That means that, in such a system, we assume agents are never aware of formulas that are not in $\Sigma$.

### *7.2. Extending KBE with other mental states*

We also could consider mental states other than awareness, in particular mental states on which a manipulation strategy relies (see Table 1). Indeed, while our KBE system can express strategies based on an agent's beliefs and knowledge as with deception or lying, we cannot express other forms of strategies when they are based on the agents' desires, norms, or when they are based on trust in the sincerity that one agent gives to another agent.

In a previous work, we propose a logic – called TB system – in which a modality $T^s_{j,i}$ expresses the fact that an agent $i$ trusts the sincerity of an agent $i$ about something (Leturc & Bonnet, 2018). A perspective would consist in merging the KBE system and the TB system. By merging them we would be able to deduce new theorems such as if an agent $j$ trusts the sincerity of another agent $i$ about the fact that $i$ does not instrumentalize him, then it cannot be the case that agent $j$ believes that $i$ can instrumentalize him. This theorem would be represented by the following formula:

$$\vdash T^s_{j,i} \neg E^d_i E_j \varphi \Rightarrow \neg B_j E^d_i E_j \varphi$$

We would then have new forms of manipulation whose concealment would be based on trust between two agents. In the same way, it could be interesting to extend the KBE system with notions of desires, norms or obligations so as to describe other kinds of manipulation strategies presented in Table 1.

## References


Abell, P. (1977). The many faces of power and liberty: Revealed preference, autonomy, and teleological explanation. *Sociology*, *11*(1), 3–24.

Ackerman, F. N. (1995). The concept of manipulativeness. *Philosophical Perspectives*, *9*, 335–340.

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

Aïmeur, E., & Sahnoune, Z. (2020). Privacy, trust, and manipulation in online relationships. *Journal of Technology in Human Services*, *38*(2), 159-183.

Akopova, A. S. (2013). Linguistic manipulation: Definition and types. *International Journal of Cognitive Research in Science, Engineering and Education*, *1*(2), 78–82.

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

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

Barnhill, A. (2014). What is manipulation. In C. Coons & M. Weber (Eds.), *Manipulation: Theory and practice* (pp. 51–72). Oxford University Press.

Baron, M. (2003). Manipulativeness. *Addresses of the American Philosophical Association*, *77*(2), 37–54.

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

Blackburn, P., De Rijke, M., & Venema, Y. (2002). *Modal logic: Graph. darst* (Vol. 53). Cambridge University Press.

Bonnet, G., Leturc, C., Lorini, E., & Sartor, G. (2021). Influencing choices by changing beliefs: A logical theory of influence, persuasion, and deception. In S. Sarkadi, B. Wright, P. Masters, & M. P. (Eds.), *Deceptive AI. DeceptECAI 2020, DeceptAI 2021. Communications in Computer and Information Science* (Vol. 1296, pp. 124–141). Springer.

Bottazzi, E., & Troquard, N. (2015). On help and interpersonal control. In A. Herzig & E. Lorini (Eds.), *The cognitive foundations of group attitudes and social interaction* (pp. 1–23). Springer.

Bowers, L. (2003). Manipulation: description, identification and ambiguity. *Journal of Psychiatric and Mental Health Nursing*, *10*(3), 323–328.

Broersen, J. (2008). A complete STIT logic for knowledge and action, and some of its applications. In *International workshop on declarative agent languages and technologies* (pp. 47–59).

Calo, R. (2013). Digital market manipulation. *George Washington Law Review*, *82*(4), 995–1051.

Castelfranchi, C., & Falcone, R. (2010). *Trust theory: A socio-cognitive and computational model*. John Wiley & Sons.

Cialdini, R. B. (2001). Harnessing the science of persuasion. *Harvard Business Review*, *79*(9), 72–81.

Cialdini, R. B. (2012). *Influence and manipulation*. First Editions.

Cohen, S. (2017). Manipulation and deception. *Australasian Journal of Philosophy*, *96*(3), 1–15.

de Saussure, L., & Schulz, P. J. (2005). *Manipulation and ideologies in the twentieth century: Discourse, language, mind*. John Benjamins Publishing.

Ettinger, D., & Jehiel, P. (2010). A theory of deception. *Microeconomics*, *2*(1), 1–20.

Faden, R. R., & Beauchamp, T. L. (1986). *A history and theory of informed consent*. Oxford University Press.

Fagin, R., & Halpern, J. Y. (1987). Belief, awareness, and limited reasoning. *Artificial Intelligence*, *34*(1), 39–76.

Festinger, L. (1962). *A theory of cognitive dissonance* (Vol. 2). Stanford University Press.

Gärdenfors, P. (1976). Manipulation of social choice functions. *Journal of Economic Theory*, *13*(2), 217–228.

Gibbard, A. (1973). Manipulation of voting schemes: a general result. *Econometrica*, *41*(4), 587–601.

Giordano, L., Martelli, A., & Schwind, C. (2000). Ramification and causality in a modal action logic. *Journal of Logic and Computation*, *10*(5), 625–662.

Goodin, R. E. (1980). *Manipulatory politics*. Yale University Press.

Gunderson, J. G. (1984). *Borderline personality disorder*. SUNY Press.

Handelman, S. (2009). *Thought manipulation: the use and abuse of psychological trickery*. Praeger.

Harel, D., Kozen, D., & Tiuryn, J. (2001). Dynamic logic. In *Handbook of philosophical logic* (pp. 99–217). Springer.

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

Hoffman, K., Zage, D., & Nita-Rotaru, C. (2009). A survey of attack and defense techniques for reputation systems. *ACM Computing Surveys*, *42*(1), 1–17.

Hoffman, K., Zage, D., & Nita-Rotaru, C. (2009). A survey of attack and defense techniques for reputation systems. *ACM Computing Surveys (CSUR)*, *42*(1), 1.

Josang, A., & Golbeck, J. (2009). Challenges for robust trust and reputation systems. In *5th International workshop on security and trust management*.

Joule, R.-V., Beauvois, J.-L., & Deschamps, J.-C. (2002). *Concise handbook of manipulation in honest people's favour (french edition).* Grenoble University Press.

Joule, R.-V., Girandola, F., & Bernard, F. (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), 493–505.

Kahneman, D. (2011). *Thinking, fast and slow.* Macmillan.

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

Leturc, C., & Bonnet, G. (2018). A normal modal logic for trust in the sincerity. In *17th International conference on autonomous agents and multiagent systems* (pp. 175–183).

Leturc, C., & Bonnet, G. (2020). A deliberate BIAT logic for modeling manipulations. In *20th International conference on autonomous agents and multiagent systems* (pp. 699–707).

Levesque, H. J. (1984). A logic of implicit and explicit belief. In *AAAI Conference on artificial intelligence* (pp. 198–202).

Lorini, E., & Sartor, G. (2016). A STIT logic for reasoning about social influence. *Studia Logica*, *104*(4), 773–812.

Luhmann, N. (1979). *Trust and power.* Wiley.

Mahon, J. E. (2008). Two definitions of lying. *International Journal of Applied Philosophy*, *22*(2), 211–230.

Maillat, D., & Oswald, S. (2009). Defining manipulative discourse: The pragmatics of cognitive illusions. *International Review of Pragmatics*, *1*(2), 348–370.

Maoz, Z. (1990). Framing the national interest: The manipulation of foreign policy decisions in group settings. *World Politics*, *43*(1), 77–110.

Masters, P., Smith, W., Sonenberg, L., & Kirley, M. (2021). Characterising deception in AI: A survey. In S. Sarkadi, B. Wright, P. Masters, & M. P. (Eds.), *Deceptive AI. DeceptECAI 2020, DeceptAI 2021. Communications in Computer and Information Science* (Vol. 1296, pp. 3–16). Springer.

McCloskey, H. J. (1980). Coercion: its nature and significance. *The Southern Journal of Philosophy*, *18*(3), 335-351.

McGinn, C. (2014). *Mindfucking: A critique of mental manipulation.* Routledge.

Mills, C. (1995). Politics and manipulation. *Social Theory and Practice*, *21*(1), 97–112.

Mintzberg, H. (1987). The strategy concept I: Five Ps for strategy. *California Management Review*, *30*(1), 11–24.

Mobasher, B., Burke, R., Bhaumik, R., & Sandvig, J. J. (2007). Attacks and remedies in collaborative recommendation. *IEEE Intelligent Systems*, *22*(3), 56–63.

Modica, S., & Rustichini, A. (1994). Awareness and partitional information structures. *Theory and Decision*, *37*(1), 107–124.

Noggle, R. (1996). Manipulative actions: a conceptual and moral analysis. *American Philosophical Quarterly*, *33*(1), 43–55.

O'Keefe, D. J. (2015). *Persuasion: Theory and research (third edition).* Sage Publications.

Pacuit, E. (2017). *Neighborhood semantics for modal logic.* Springer.

Parkes, D. C., & Ungar, L. H. (2000). Preventing strategic manipulation in iterative auctions: Proxy agents and price-adjustment. In *AAAI Conference on artificial intelligence* (pp. 82–89).

Pörn, I. (1977). *Action theory and social science: Some formal models.* Springer.

Poulin, R. (2010). Parasite manipulation of host behavior: an update and frequently asked questions. *Advances in the Study of Behavior*, *41*, 151–186.

Raz, J. (1986). *The morality of freedom.* Clarendon Press.

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

Rigotti, E. (2005). Towards a typology of manipulative processes. In L. de Saussure & P. Schulz (Eds.), *Manipulation and ideologies in the twentieth century: discourse, language, mind* (pp. 61–83). John Benjamins Publishing Company.

Robinson, M. S. (1985). Collusion and the choice of auction. *RAND Journal of Economics*,

*16*(1), 141–145.
Rosenberg, M., & Pearlin, L. I. (1962). Power-orientations in the mental hospital. *Human Relations*, *15*(4), 335–349.
Ruan, Y., & Durresi, A. (2016). A survey of trust management systems for online social communities – trust modeling, trust inference and attacks. *Knowledge-Based Systems*, *106*, 150–163.
Rudinow, J. (1978). Manipulation. *Ethics*, *88*(4), 338–347.
Saint Clair, H. R. (1966). Manipulation. *Comprehensive Psychiatry*, *7*(4), 248–258.
Sakama, C. (2021). Deception in epistemic causal logic. In S. Sarkadi, B. Wright, P. Masters, & M. P. (Eds.), *Deceptive AI. DeceptECAI 2020, DeceptAI 2021. Communications in Computer and Information Science* (Vol. 1296, pp. 105–123). Springer.
Sakama, C., Caminada, M., & Herzig, A. (2015). A formal account of dishonesty. *Logic Journal of the IGPL*, *23*(2), 259–294.
Sanghvi, S., & Parkes, D. (2004). *Hard-to-manipulate VCG-based auctions* (Tech. Rep.). Division of Engineering and Applied Sciences: Harvard University.
Santos, F., & Carmo, J. (1996). Indirect action, influence and responsibility. In *Deontic logic, agency and normative systems* (pp. 194–215). Springer.
Schipper, B. (2014). Awareness. In H. van Ditmarsch, J. Y. Halpern, W. van der Hoek, & B. Kooi (Eds.), *Handbook of epistemic logic* (pp. 77–146). College Publications.
Segerberg, K., Meyer, J.-J., & Kracht, M. (2009). *The logic of action.* Stanford Library of Philosophy.
Singh, A., Ngan, T.-W., Druschel, P., & Wallach, D. (2006). Eclipse attacks on overlay networks: Threats and defenses. In *25th International Conference on Computer Communications.*
Sorlin, S. (2017). The pragmatics of manipulation: Exploiting im/politeness theories. *Journal of Pragmatics*, *121*, 132–146.
Stalnaker, R. (2006). On logics of knowledge and belief. *Philosophical Studies*, *128*(1), 169–199.
Strauss, N. (2006). *The game.* Canongate Books.
Sunstein, C. R. (2015). Fifty shades of manipulation. *Journal of Marketing Behavior*, *213*.
Todd, P. (2013). *Manipulation.* International Encyclopedia of Ethics.
Troquard, N. (2014). Reasoning about coalitional agency and ability in the logics of "bringing-it-about". *Autonomous Agents and Multiagent Systems*, *28*(3), 381–407.
Turner, J. A., Deyo, R. A., Loeser, J. D., Von Korff, M., & Fordyce, W. E. (1994). The importance of placebo effects in pain treatment and research. *Journal of the American Medical Association*, *271*(20), 1609–1614.
Vallée, T., & Bonnet, G. (2015). Using KL divergence for credibility assessment. In *14th International conference on autonomous agents and multiagent Systems* (pp. 1797–1798).
Van Dijk, T. A. (2006). Discourse and manipulation. *Discourse & Society*, *17*(3), 359–383.
Van Ditmarsch, H., Van Der Hoek, W., & Kooi, B. (2007). *Dynamic epistemic logic.* Springer Science & Business Media.
van der Hoek, W., Iliev, P., & Wooldridge, M. J. (2012). A logic of revelation and concealment. In *Aamas* (pp. 1115–1122).
Van Ditmarsch, H., French, T., Velázquez-Quesada, F. R., & Wáng, Y. N. (2018). Implicit, explicit and speculative knowledge. *Artificial Intelligence*, *256*, 35–67.
Van Ditmarsch, H., Van Eijck, J., Sietsma, F., & Wang, Y. (2012). On the logic of lying. In J. van Eijck & R. Verbrugge (Eds.), *Games, actions and social software* (pp. 41–72). Springer.
Wagner, A. R., & Arkin, R. C. (2009). Robot deception: recognizing when a robot should deceive. In *International symposium on computational intelligence in robotics and automation* (pp. 46–54).
Ware, A. (1981). The concept of manipulation: its relation to democracy and power. *British Journal of Political Science*, *11*(2), 163–181.
Whiten, A., & Byrne, R. W. (1988). Tactical deception in primates. *Behavioral and Brain*

*Sciences*, *11*(2), 233–244.
Wilkinson, T. M. (2013). Nudging and manipulation. *Political Studies*, *61*(2), 341–355.
Wood, A. W. (2014). Coercion, manipulation, exploitation. In C. Coons & M. Weber (Eds.), *Manipulation: Theory and practice* (pp. 17–50). Oxford University Press.

## Appendix A. Soundness, completeness and theorems of KBE

In this appendix, we firstly prove some theorems on the KBE semantics constraints. Then we demonstrate that the axiomatic system presented in Figure 4 is correct. Then, we demonstrate that this axiomatic system is complete. We show that it verifies the deduction theorems and that it is strongly correct and strongly complete. Finally, we give the proofs of the theorems with a Hilbert system.

### *A.1. Theorems on semantic constraints*

**Proposition 4.1.** *If $(E^d_{KN})$ and $(E^d_{KP})$ hold at the same time, then it implies on the Kripke structure that the following property holds:*

$$\forall w \in \mathcal{W} : \mathcal{E}^d_i(w) = \bigcap_{v \in \mathcal{W}: w\mathcal{K}_i v} \mathcal{E}^d_i(v) = \bigcup_{v \in \mathcal{W}: w\mathcal{K}_i v} \mathcal{E}^d_i(v) \qquad (E^d_{KN}) + (E^d_{KP})$$

***Proof.*** Let assume a frame $\mathcal{C}$ s.t. $(E^d_{KP})$ and $(E^d_{KN})$ hold. Let us remark that $(E^d_{KP})$ obviously implies that:

$$\forall w \in \mathcal{W} : \mathcal{E}^d_i(w) \subseteq \bigcap_{v \in \mathcal{W}: w\mathcal{K}_i v} \mathcal{E}^d_i(v) \subseteq \bigcup_{v \in \mathcal{W}: w\mathcal{K}_i v} \mathcal{E}^d_i(v) \quad (E^d_{KP})$$

Then, since for all sets $E, F$, $(E \subseteq F \wedge \forall e, e \notin E \Rightarrow e \notin F) \Longleftrightarrow E = F$[10], and since $(E^d_{KN})$ holds, we thus deduce:

$$\forall w \in \mathcal{W} : \mathcal{E}^d_i(w) = \bigcup_{v \in \mathcal{W}: w\mathcal{K}_i v} \mathcal{E}^d_i(v)$$

Then, with $(E^d_{KP})$ we prove the complete theorem :

$$\forall w \in \mathcal{W} : \bigcup_{v \in \mathcal{W}: w\mathcal{K}_i v} \mathcal{E}^d_i(v) = \mathcal{E}^d_i(w) \subseteq \bigcap_{v \in \mathcal{W}: w\mathcal{K}_i v} \mathcal{E}^d_i(v)$$

Consequently:

$$\forall w \in \mathcal{W} : \mathcal{E}^d_i(w) = \bigcap_{v \in \mathcal{W}: w\mathcal{K}_i v} \mathcal{E}^d_i(v) = \bigcup_{v \in \mathcal{W}: w\mathcal{K}_i v} \mathcal{E}^d_i(v) \qquad (E^d_{KN}) + (E^d_{KP})$$

□

**Proposition 4.2.**

*(1) If $(E^d_{KP})$ holds and $\mathcal{K}_i$ is reflexive, then:*

$$\forall w \in \mathcal{W}, \mathcal{E}^d_i(w) = \bigcap_{v \in \mathcal{W}: w\mathcal{K}_i v} \mathcal{E}^d_i(v)$$

[10]It is an obvious theorem of the theory of set. Let $E, F$ be two sets. ($\Rightarrow$) Let assume $(E \subseteq F \wedge \forall e, e \notin E \Rightarrow e \notin F)$. W have $E \subseteq F$ and so, let us show that $F \subseteq E$. Let $f \in F$. Since $\forall e, e \notin E \Rightarrow e \notin F$ is equivalent, by contraposition, to $\forall e, e \in F \Rightarrow e \in E$, we deduce that $f \in E$. So $E = F$. ($\Leftarrow$) Let us assume $E = F$, thus $E \subseteq F$ and $F \subseteq E$. Thus $\forall e, e \in F \Rightarrow e \in E$ and by contraposition, we deduce $\forall e, e \notin E \Rightarrow e \notin F$.

*(2) If $(E^d_{KN})$ holds and $\mathcal{K}_i$ is reflexive, then:*

$$\forall w \in \mathcal{W}, \mathcal{E}^d_i(w) = \bigcup_{v \in \mathcal{W}: w\mathcal{K}_i v} \mathcal{E}^d_i(v)$$

***Proof.***

(1) Let assume a frame $\mathcal{C}$ s.t. $(E^d_{KP})$ holds and $\mathcal{K}_i$ is reflexive, then:

$$\forall w \in \mathcal{W}, \mathcal{E}^d_i(w) \subseteq \bigcap_{v \in \mathcal{W}: w\mathcal{K}_i v} \mathcal{E}^d_i(v) \subseteq \mathcal{E}^d_i(w)$$

Consequently, $\forall w \in \mathcal{W}, \mathcal{E}^d_i(w) = \bigcap_{v \in \mathcal{W}: w\mathcal{K}_i v} \mathcal{E}^d_i(v)$.

(2) Let assume a frame $\mathcal{C}$ s.t. $(E^d_{KN})$ holds and $\mathcal{K}_i$ is reflexive, then:

$$\forall w \in \mathcal{W}, \bigcup_{v \in \mathcal{W}: w\mathcal{K}_i v} \mathcal{E}^d_i(v) \subseteq \mathcal{E}^d_i(w) \subseteq \bigcup_{v \in \mathcal{W}: w\mathcal{K}_i v} \mathcal{E}^d_i(v)$$

Consequently, $\forall w \in \mathcal{W}, \mathcal{E}^d_i(w) = \bigcup_{v \in \mathcal{W}: w\mathcal{K}_i v} \mathcal{E}^d_i(v)$.

□

### *A.2. Soundness*

This section aims to demonstrate the correctness of our KBE system. Thus, in the sequel we consider any frame $\mathcal{C} = (\mathcal{W}, \{\mathcal{B}_i\}_{i\in\mathcal{N}}, \{\mathcal{K}_i\}_{i\in\mathcal{N}}, \{\mathcal{E}_i\}_{i\in\mathcal{N}}, \{\mathcal{E}^d_i\}_{i\in\mathcal{N}})$.

#### *A.2.1. Normal modalities*

It is well known that the semantics of a normal modality of a S5 system that preserves validity is an equivalence relation Blackburn et al. (2002). 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. Let us recall a proof method to show that *confluence* corresponds to the axiom **4.2**.

**Proposition A.1.** *$\mathcal{K}_i$ is* confluent *if, and only if, $\mathcal{C} \models \langle K_i\rangle K_i p \Rightarrow K_i\langle K_i\rangle\varphi$*

***Proof.*** ($\Rightarrow$) Let us suppose a frame $\mathcal{C} \not\models \langle K_i\rangle K_i p \Rightarrow K_i\langle K_i\rangle p$, i.e. there exists a model $\mathcal{M}$ on $\mathcal{C}$ and a world $w \in \mathcal{W}$ s.t. $\mathcal{M}, w \models \langle K_i\rangle K_i p \wedge \neg K_i\langle K_i\rangle p$. Thus, there exists $u \in \mathcal{W}, w\mathcal{K}_i u$ and $v \in \mathcal{W}, w\mathcal{K}_i v$ such that $\mathcal{M}, u \models K_i p$ and $\mathcal{M}, v \models \neg\langle K_i\rangle p$. So for all $z_1 \in \mathcal{W}, u\mathcal{K}_i z_1, \mathcal{M}, z_1 \models p$ and for all $z_2 \in \mathcal{W}, v\mathcal{K}_i z_2, \mathcal{M}, z_2 \models \neg p$. Therefore $z_1 \neq z_2$ and $\mathcal{K}_i$ is not confluent.

($\Leftarrow$) By contraposition, let us suppose a frame $\mathcal{C}$ s.t. $\mathcal{K}_i$ is not confluent i.e.

$$\exists w, u, v \in \mathcal{W} : w\mathcal{K}_i u \wedge w\mathcal{K}_i v \wedge \nexists z \in \mathcal{W} : u\mathcal{K}_i z \wedge v\mathcal{K}_i z$$

Let us assume $\mathcal{M}$ on $\mathcal{C}$ s.t. $\mathcal{M}, w \models \langle K_i\rangle K_i p$, $V(p) = \mathcal{W} \setminus \mathcal{K}_i(u)$. Thus, $\mathcal{M}, u \models K_i\neg p$ i.e. $\mathcal{M}, u \models \neg\langle K_i\rangle p$. Since $w\mathcal{K}_i u$, we have $\mathcal{M}, w \models \langle K_i\rangle\neg\langle K_i\rangle p$ i.e. $\mathcal{M}, w \models \neg K_i\langle K_i\rangle p$.

Consequently, we have proved, there exists $\mathcal{M}$ on $\mathcal{C}$ s.t. $\mathcal{M}, w \models \langle K_i \rangle K_i p \wedge \neg K_i \langle K_i \rangle p$ i.e. $\mathcal{C} \not\models \langle K_i \rangle K_i p \Rightarrow K_i \langle K_i \rangle \varphi$

□

Concerning the inference rules between the modality $K_i$ and $B_i$, Stalnaker (2006) 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 $B_i$.

*A.2.2. Non standard properties*

In this section, we focus on the non-standard properties associated with our neighborhood semantics for the relation $\mathcal{E}_i^d$. Some of these properties may be found in Pacuit (2017) but not all. Thus, in the interest of rigor, we demonstrate all non standard properties in this section.

Firstly the axiom $(C)$ is valid in KBE.

**Proposition A.2.**

$$\mathcal{C} \models E_i^d \varphi \wedge E_i^d \psi \Rightarrow E_i^d(\varphi \wedge \psi)$$

*if, and only if,*

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

***Proof.*** ($\Rightarrow$) By contraposition, let be a frame $\mathcal{C}$ such that : $\exists w \in \mathcal{W} : S \in \mathcal{E}_i^d(w) \wedge T \in \mathcal{E}_i^d(w) \wedge S \cap T \notin \mathcal{E}_i^d(w)$. Let us assume $\mathcal{M}, w$ s.t. $\mathcal{M}, w \models E_i^d p \wedge E_i^d q$, $V(p) = S$ and $V(q) = T$. Thus, $\mathcal{M}, w \models E_i^d p$ i.e. $||p|| \in \mathcal{E}_i^d(w)$ and $\mathcal{M}, w \models E_i^d q$ i.e. $||p|| \in \mathcal{E}_i^d(w)$. However since $S \cap T \notin \mathcal{E}_i^d(w)$, we have $||p|| \cap ||q|| \notin \mathcal{E}_i^d(w)$ and so $||p \wedge q|| \notin \mathcal{E}_i^d(w)$ i.e. $\mathcal{M}, w \not\models E_i^d(p \wedge q)$.

($\Leftarrow$) Let us assume that $\mathcal{C} \not\models E_i^d p \wedge E_i^d q \Rightarrow E_i^d(p \wedge q)$, i.e. there exists a model $\mathcal{M}$ and a world $w \in \mathcal{W}$ such that $\mathcal{M}, w \models E_i^d p \wedge E_i^d q \wedge \neg E_i^d(p \wedge q)$. So $||p|| \in E_i(w)$ and $||q|| \in E_i(w)$. Moreover, as $\mathcal{M}, w \models \neg E_i^d(p \wedge q)$, we have $||p \wedge q|| \notin \mathcal{E}_i^d(w)$ i.e $||p|| \cap ||q|| \notin \mathcal{E}_i^d(w)$. Thus, we prove that there exists a world $w \in \mathcal{W}$, $S \in \mathcal{E}_i(w) \wedge T \in \mathcal{E}_i(w) \wedge S \cap T \notin \mathcal{E}_i(w)$. □

The necessitation does not hold for the modality of deliberate effects.

**Proposition A.3.**

$$\mathcal{C} \models \neg E_i^d \top$$

*if, and only if,*

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

***Proof.*** Let us suppose a frame $\mathcal{C} \models \neg E_i^d \top$. As $||\top|| = \mathcal{W}$, it follows $\forall \mathcal{M}, \forall w \in \mathcal{W} : \mathcal{M}, w \models \neg E_i^d \top$ if, and only if, $\forall w \in \mathcal{W}, ||\top|| \notin \mathcal{E}_i(w)$ i.e. $\forall w \in \mathcal{W}, \mathcal{W} \notin \mathcal{E}_i(w)$. □

The deliberate effects has positive introspection with knowledge.

**Proposition A.4.**

$$\mathcal{C} \models E_i^d p \Rightarrow K_i E_i^d p$$

*if, and only if,*

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

***Proof.*** ($\Rightarrow$) By contraposition, let us assume a frame $\mathcal{C}$ such that $\exists w \in \mathcal{W} : \mathcal{E}_i^d(w) \not\subseteq \bigcap_{v \in \mathcal{W}: w\mathcal{K}_i v} \mathcal{E}_i^d(v)$. First, let us notice that necessary $\mathcal{E}_i^d(w) \neq \emptyset$ since $\emptyset \subseteq \bigcap_{v \in \mathcal{W}: w\mathcal{K}_i v} \mathcal{E}_i^d(v)$. Thus there exists $X \in \mathcal{E}_i^d(w) \setminus \bigcap_{v \in \mathcal{W}: w\mathcal{K}_i v} \mathcal{E}_i^d(v)$ i.e. $X \in \mathcal{E}_i^d(w)$ and $X \notin \bigcap_{v \in \mathcal{W}: w\mathcal{K}_i v} \mathcal{E}_i^d(v)$. Second, let $\mathcal{M}$ be a model on $\mathcal{C}$ such that $X = V(p) \in \mathcal{E}_i^d(w) \setminus \bigcap_{v \in \mathcal{W}: w\mathcal{K}_i v} \mathcal{E}_i^d(v)$. As $||p|| \in \mathcal{E}_i^d(w)$, we have $\mathcal{M}, w \models E_i^d p$. Furthermore since $X \notin \bigcap_{v \in \mathcal{W}: w\mathcal{K}_i v} \mathcal{E}_i^d(v)$, we have there exists $v \in \mathcal{K}_i(w), X \notin \mathcal{E}_i^d(v)$[11] i.e. $\mathcal{M}, v \models \neg E_i^d p$. Thus, $\mathcal{M}, w \models \langle K_i \rangle \neg E_i^d p$. Consequently, $\mathcal{M}, w \models \neg K_i E_i^d p$ and so, $\mathcal{C} \not\models E_i^d p \Rightarrow K_i E_i^d p$.

($\Leftarrow$) Let us assume $\mathcal{C} \not\models E_i^d p \Rightarrow K_i E_i^d p$, i.e. there exists a model $\mathcal{M}$ and a world $w \in \mathcal{W}$ such that $\mathcal{M}, w \models E_i^d p \wedge \neg K_i E_i^d p$. So, we have that $||p|| \in \mathcal{E}_i^d(w)$. Moreover, there exists $v \in \mathcal{W}$ such that: $w\mathcal{K}_i v$ and $\mathcal{M}, v \models \neg E_i^d p$, i.e. $||p|| \notin \mathcal{E}_i^d(v)$. So, we deduce that :

$$||p|| \notin \bigcap_{v \in \mathcal{W}: w\mathcal{K}_i v} \mathcal{E}_i^d(v)$$

Consequently, we have proved that:

$$\exists w \in \mathcal{W} : \mathcal{E}_i^d(w) \not\subseteq \bigcap_{v \in \mathcal{W}: w\mathcal{K}_i v} \mathcal{E}_i^d(v)$$

□

The deliberate effects has negative introspection with knowledge.

**Proposition A.5.**

$$\mathcal{C} \models \neg E_i^d p \Rightarrow K_i \neg E_i^d p$$

*if, and only if,*

$$\forall w, v \in \mathcal{W}, \forall S \in 2^{\mathcal{W}} : S \notin \mathcal{E}_i^d(w) \Longrightarrow (w\mathcal{K}_i v \Rightarrow S \notin \mathcal{E}_i^d(v))$$

***Proof.*** ($\Rightarrow$) By contraposition, let us assume a frame $\mathcal{C}$ such that :

$$\exists w, v \in \mathcal{W}, \exists S \in 2^{\mathcal{W}} : S \notin \mathcal{E}_i^d(w) \wedge w\mathcal{K}_i v \wedge S \in \mathcal{E}_i^d(v)$$

Let us define a model $\mathcal{M}$ on $\mathcal{C}$ s.t. $V(p) = S$. As $||p|| \notin \mathcal{E}_i^d(w)$, we directly have $\mathcal{M}, w \models \neg E_i^d p$. As $||p|| \in \mathcal{E}_i^d(v)$, we have $\mathcal{M}, v \models E_i^d p$. Furthermore, since $w\mathcal{K}_i v$,

[11] $X \in \bigcap_{v \in \mathcal{W}: w\mathcal{K}_i v} \mathcal{E}_i^d(v)$ iff $\forall v \in \mathcal{K}_i(w), X \in \mathcal{E}_i^d(v)$. Thus, $X \notin \bigcap_{v \in \mathcal{W}: w\mathcal{K}_i v} \mathcal{E}_i^d(v)$ iff $\exists v \in \mathcal{K}_i(w), X \notin \mathcal{E}_i^d(v)$

we deduce that $\mathcal{M}, w \models \langle K_i \rangle E_i^d p$ i.e. $\mathcal{M}, w \models \neg K_i \neg E_i^d p$. Consequently, $\mathcal{M}, w \models \neg E_i^d p \wedge \neg K_i \neg E_i^d p$. We have proved that there exists a model $\mathcal{M} \not\models \neg E_i^d p \Rightarrow K_i \neg E_i^d p$ and so $\mathcal{C} \not\models \neg E_i^d p \Rightarrow K_i \neg E_i^d p$.

($\Leftarrow$) Let us suppose by contraposition $\mathcal{C} \not\models \neg E_i^d p \Rightarrow K_i \neg E_i^d p$, i.e. there exists a model $\mathcal{M}$ and a world $w \in \mathcal{W}$ such that $\mathcal{M}, w \models \neg E_i^d p \wedge \neg K_i \neg E_i^d p$. So $||p|| \notin \mathcal{E}_i^d(w)$. Moreover, there exists $v \in \mathcal{W}$ such that $w \mathcal{K}_i v$ and $\mathcal{M}, v \models E_i^d p$, i.e $||p|| \in \mathcal{E}_i^d(v)$. Consequently, for $S = ||p||$, we can conclude that:

$$\exists w, v \in \mathcal{W}, \exists S \in 2^{\mathcal{W}} : S \notin \mathcal{E}_i^d(w) \wedge w \mathcal{K}_i v \wedge S \in \mathcal{E}_i^d(v)$$

□

There is logical link between deliberate effects and effects of actions.

**Proposition A.6.**

$$\mathcal{C} \models E_i^d p \Rightarrow E_i p$$

*if, and only if,*

$$\forall w \in \mathcal{W}, \forall S \in 2^{\mathcal{W}} : S \in \mathcal{E}_i^d(w) \Longrightarrow \mathcal{E}_i(w) \subseteq S$$

***Proof.*** ($\Rightarrow$) By contraposition, let us suppose a frame $\mathcal{C}$ such that : $\exists w \in \mathcal{W}, \exists S \in 2^{\mathcal{W}} : S \in \mathcal{E}_i^d(w) \wedge \mathcal{E}_i(w) \not\subseteq S$. Let us consider a model $\mathcal{M}$ on $\mathcal{C}$ s.t. $V(p) = S$. So, we have, $||p|| \in \mathcal{E}_i^d(w)$ and so $\mathcal{M}, w \models E_i^d p$. Furthermore, as $\mathcal{E}_i(w) \not\subseteq S$, there exists $v \in \mathcal{E}_i(w) \setminus S$, and so $\mathcal{M}, v \models \neg p$ (since $S = ||p||$). We deduce that since $w \mathcal{E}_i v$ we have $\mathcal{M}, w \models \langle E_i \rangle \neg p$ i.e. $\mathcal{M}, w \models \neg E_i p$. We have $\mathcal{M}, w \models E_i^d p \wedge \neg E_i p$. Consequently, we can conclude that there exists $\mathcal{M}$ satisfying the constraint of the frame $\mathcal{C}$ such that $\mathcal{M} \not\models E_i^d p \Rightarrow E_i p$, and so $\mathcal{C} \not\models E_i^d p \Rightarrow E_i p$.

($\Leftarrow$) Let us suppose $\mathcal{C} \not\models E_i^d p \Rightarrow K_i E_i^d p$, i.e. there exists a model $\mathcal{M}$ and a world $w \in \mathcal{W}$ such that $\mathcal{M}, w \models E_i^d p \wedge \neg E_i p$. So $||p|| \in \mathcal{E}_i^d(w)$ and there exists $v \in \mathcal{W}$ s.t. $w \mathcal{E}_i v$ and $\mathcal{M}, v \models \neg p$. By definition, we have $v \notin ||p||$. Consequently, for $S = ||p||$, we prove that $\exists (w, S) \in \mathcal{W} \times 2^{\mathcal{W}} : S \in \mathcal{E}_i^d(w) \wedge \mathcal{E}_i(w) \not\subseteq S$. □

*A.2.3. Soundness properties*

In this section, we give the complete proof of soundness.

**Theorem 4.4.** *The KBE system is sound.*

***Proof.*** In order to show soundness, we have to show that substitution, modus ponens, necessitation, (RE) and (DUAL) preserve the validity for all modalities.

(1) Let us prove that substitution preserves validity. Let $\varphi \in \mathcal{L}_{KBE}$ be a formula and $p_{a_1}, \ldots, p_{a_n} \in \mathcal{P}$ be propositional atoms that are contained in $\varphi$. We define $\theta = \varphi(\psi_1/p_{a_1}, \ldots, \psi_n/p_{a_n})$ the obtained formula after an uniform substitution on $\varphi$ and $\psi_1, \ldots, \psi_n \in \mathcal{L}_{KBE}$ formulas. We want to prove that if $\varphi$ is valid in $\mathcal{C}$, then $\theta$ is also valid in $\mathcal{C}$. By contraposition let us assume that $\mathcal{C} \not\models \theta$. So there exists a model $\mathcal{M} = (\mathcal{C}, h)$ and a world $w \in \mathcal{W}$ such that: $\mathcal{M}, w \not\models \theta$. Let us build a model for $\varphi$, $\mathcal{M}' = (\mathcal{C}, h')$ such that:

- $\forall j \in \mathbb{N} : 1 \geq j \geq n, \mathcal{M}, w \models \psi_j \Longrightarrow w \in h'(p_{a_j})$

- $\forall j \in \mathbb{N} : 1 \geq j \geq n, \mathcal{M}, w \not\models \psi_j \Longrightarrow w \notin h'(p_{a_j})$
- $\forall p \in \mathcal{P} : \forall j \in \mathbb{N} : p \neq p_{a_j}, w \notin h'(p)$[12]

Since $\mathcal{M}, w \not\models \theta$, we have that the combination of $\psi_j$ invalidate formula $\theta$ in $\mathcal{M}$, $w$. Since for all $\psi_j$ we associate an atom $p_{a_j}$ language with the same truth value as the formula as $\psi_j$. The combination of $p_{a_j}$ invalidate the formula $\varphi$ in $\mathcal{M}'$, $w$. Thus, we have $\mathcal{M}', w \not\models \varphi$. It has therefore been well proven by contraposition that substitution preserves validity, i.e. if $\varphi$ is valid in $\mathcal{C}$ then its substitution $\psi$ is also valid in $\mathcal{C}$.

(2) Let us prove that modus ponens preserves validity. Let us suppose $\mathcal{C} \models \varphi$ and $\mathcal{C} \models (\varphi \Rightarrow \psi)$. So for all models $\mathcal{M}$ and for all worlds $w$, we have $\mathcal{M}, w \models \varphi$ and $\mathcal{M}, w \models \varphi \Rightarrow \psi$, i.e. $\mathcal{M}, w \models \varphi \wedge (\neg\varphi \vee \psi)$, i.e $\mathcal{M}, w \models (\varphi \wedge \neg\varphi) \vee (\varphi \wedge \psi)$, so $\mathcal{M}, w \models (\varphi \wedge \psi)$, i.e. $\mathcal{M}, w \models \varphi$ and $\mathcal{M}, w \models \psi$. So $\mathcal{M}, w \models \psi$. Therefore it has been proven that for all models $\mathcal{M}$ and for all worlds $w$, $\psi$ is valid, i.e. $\mathcal{C} \models \psi$.

(3) Let us prove that necessitation preserves validity for all $(\Box, \mathcal{R}) \in \{(B_i, \mathcal{B}_i), (K_i, \mathcal{K}_i), (E_i, \mathcal{E}_i)\}$. Let $\varphi$ be a tautology, so for all models $\mathcal{M}$ and $\forall v \in \mathcal{W}, \mathcal{M}, v \models \varphi$. So $\forall w, v \in \mathcal{W}, w\mathcal{R}v : \mathcal{M}, v \models \varphi$, i.e. $\forall w \in \mathcal{W}, \mathcal{M}, w \models \Box\varphi$. We prove that if $\mathcal{C} \models \varphi$, then $\mathcal{C} \models \Box\varphi$. Consequently, necessitation preserves validity.

(4) Let us prove that the rule (RE) preserves validity for $E_i^d$ i.e.:

$$\text{If } \models \varphi \Leftrightarrow \psi \text{ then } \models E_i^d\varphi \Leftrightarrow E_i^d\psi$$

Let us assume $\models \varphi \Leftrightarrow \psi$. We have $\models E_i^d\varphi$ if, and only if, for all $(\mathcal{M}, w), \mathcal{M}, w \models E_i^d\varphi$ i.e. for all $(\mathcal{M}, w), ||\varphi||_{\mathcal{M}} \in \mathcal{E}_i^d(w)$ if, and only if, for all $(\mathcal{M}, w), ||\psi||_{\mathcal{M}} \in \mathcal{E}_i^d(w)$ (since $\models \varphi \Leftrightarrow \psi$, we have for all $(\mathcal{M}', w'), ||\varphi||_{\mathcal{M}'} = ||\psi||_{\mathcal{M}'}$) iff $\models E_i^d\psi$. Consequently, we prove that if $\models \varphi \Leftrightarrow \psi$ then $\models E_i^d\varphi \Leftrightarrow E_i^d\psi$ i.e. (RE) preserves validity.

(5) Let us prove that the rule (DUAL) preserves validity for $E_i^d$ i.e.:

$$\models E_i^d\varphi \Leftrightarrow \neg\langle E_i^d\rangle\neg\varphi$$

For all models $\mathcal{M}$ and for all worlds $w \in \mathcal{W}$ such that $\mathcal{M}, w \models E_i^d\varphi$. We just have to notice that $||\varphi||_{\mathcal{M}} = \mathcal{W} \setminus (\mathcal{W} \setminus ||\varphi||_{\mathcal{M}})$ and $||\neg\varphi||_{\mathcal{M}} = \mathcal{W} \setminus ||\varphi||_{\mathcal{M}}$. Then, we have $\mathcal{M}, w \models E_i^d\varphi$ iff $||\varphi||_{\mathcal{M}} \in \mathcal{E}_i^d(w)$ iff $\mathcal{W} \setminus (\mathcal{W} \setminus ||\varphi||_{\mathcal{M}}) \in \mathcal{E}_i^d(w)$ iff $\mathcal{M}, w \not\models \langle E_i^d\rangle\neg\varphi$ iff $\mathcal{M}, w \models \neg\langle E_i^d\rangle\neg\varphi$. Consequently (DUAL) preserves validity.

□

### *A.3. Completeness*

In this section we prove the completeness of the logic KBE by using maximal consistent sets. Firstly, we define maximal consistent sets for KBE. Secondly, we give the proof of the completeness.

#### *A.3.1. Definition of a deductible system*

We recall in this section the notion of deductible system which is used in this article to prove later the strongly completeness.

[12]Since $p$ is not an atom involved in the substitution, regardless if $w \notin h'(p)$ or $w \in h'(p)$, it will not affect the demonstration. We are just making sure we have got the right model here.

*A.3.2. Deduction theorems*

We talk about KBE-deductibility when a formula $\varphi$ can be deduced from a set $\Sigma$ of formulas.

**Definition A.7** (KBE-deductibility)**.** Let $\Sigma$ be a set formulas in KBE and $\varphi$ be a formula. We say that $\varphi$ is a *KBE-deduction from* $\Sigma$ and we write $\Sigma \vdash \varphi$ if, and only if:

(1) If $\Sigma = \emptyset$, then $\vdash \varphi$ ;
(2) Otherwise $\exists \psi_1, \ldots, \psi_n \in \Sigma$ with $n \in \mathbb{N}^*$ and $\vdash \psi_1 \wedge \ldots \wedge \psi_n \Rightarrow \varphi$.

If $\Sigma \vdash \varphi$ is verified, then $\varphi$ is *KBE-deductible*.

The KBE-deductibility has various fundamental properties such as reflexivity, transitivity, and left weakening.

**Proposition A.8.** *Let $\Sigma$ and $\Gamma$ be two sets of formulas in KBE. It holds:*

*(1)* Reflexivity *holds i.e. if $\varphi \in \Sigma$, then $\Sigma \vdash \varphi$*
*(2)* Transitivity *holds i.e. if $\Sigma \vdash \varphi$ and $\{\varphi\} \vdash \psi$, then $\Sigma \vdash \psi$*
*(3)* Left weakening *holds i.e. if $\Sigma \vdash \varphi$ and $\Sigma \subseteq \Gamma$, then $\Gamma \vdash \varphi$*

***Proof.*** Let $\Sigma$ and $\Gamma$ be two non-empty sets of formulas in KBE.

(1) If $\varphi \in \Sigma$, then since $\vdash \varphi \Rightarrow \varphi$, by definition of KBE-deductibility, we have $\Sigma \vdash \varphi$.
(2) If $\Sigma \vdash \varphi$ and $\{\varphi\} \vdash \psi$, then there exists $\psi_1, \ldots, \psi_n \in \Sigma$ with $n \in \mathbb{N}^*$ such that:

$$\vdash \Big( \bigwedge_{k \in \{1,\ldots,n\}} \psi_k \Big) \Rightarrow \varphi$$

Furthermore since $\{\varphi\} \vdash \psi$, by definition we have $\vdash \varphi \Rightarrow \psi$. Thus, we deduce:

$$\vdash \Big( \bigwedge_{k \in \{1,\ldots,n\}} \psi_k \Big) \Rightarrow \psi$$

Consequently, we prove that $\Sigma \vdash \psi$.
(3) If $\Sigma \vdash \varphi$ and $\Sigma \subseteq \Gamma$, then there exists $\psi_1, \ldots, \psi_n \in \Sigma$ with $n \in \mathbb{N}^*$ such that:

$$\vdash \Big( \bigwedge_{k \in \{1,\ldots,n\}} \psi_k \Big) \Rightarrow \varphi$$

Since $\psi_1, \ldots, \psi_n \in \Sigma$ and $\Sigma \subseteq \Gamma$, we have $\psi_1, \ldots, \psi_n \in \Gamma$. Thus, we prove that $\Gamma \vdash \varphi$.

□

We prove now the deduction theorem of KBE. We first recall the definition of a model of a set of formula.

**Definition A.9.** Let $\Sigma$ be a set of formulas and $\varphi$ be a formula. $\Sigma$ *semantically entails* $\varphi$, written $\Sigma \models \varphi$ if, and only if, for all models $\mathcal{M}$ of $\Sigma$ (i.e. $\forall \psi \in \Sigma, \mathcal{M} \models \psi$), $\mathcal{M} \models \varphi$.

**Theorem A.10** (Deduction theorems)**.**

*If* $\Gamma$ *is a set of formulas in KBE,* $\varphi$ *and* $\psi$ *be two formulas of KBE, then:*

$$(1) \quad \Gamma \cup \{\psi\} \vdash \varphi \text{ if, and only if, } \Gamma \vdash \psi \Rightarrow \varphi$$

*If* $\Gamma$ *is a set of formulas in KBE,* $\varphi$ *and* $\psi$ *be two formulas of KBE, then:*

$$(2) \quad \Gamma \cup \{\psi\} \models \varphi \text{ if, and only if, } \Gamma \models \psi \Rightarrow \varphi$$

***Proof.*** Let $\Gamma \subseteq \mathcal{L}_{KBE}$ be a set of formulas, $\varphi$ and $\psi$ be two formulas of $\mathcal{L}_{KBE}$.

($\Rightarrow$) Let us assume that $\Gamma \cup \{\psi\} \vdash \varphi$. By definition of the KBE-deductibility, we have that there exists $\Sigma = \{\psi_1, \ldots, \psi_n\}, \Sigma \subseteq \Gamma \cup \{\psi\}$ such that:

$$\vdash \bigwedge_{i \in \{1,\ldots,n\}} \psi_i \Rightarrow \varphi$$

We have two cases to consider $\psi \in \Sigma$ and when $\psi \notin \Sigma$.

(1) If $\psi \in \Sigma$, then there exists $i \in \{1, \ldots, n\}$ such that $\psi = \psi_i$. So $\vdash (\bigwedge_{k \in \{1,\ldots,n\} \setminus \{i\}} \psi_k) \wedge \psi \Rightarrow \varphi$ by commutativity and by associativity of $\wedge$. Then, $\vdash \bigwedge_{k \in \{1,\ldots,n\} \setminus \{i\}} \psi_k \Rightarrow (\psi \Rightarrow \varphi)$. However for all $k \in \{1, \ldots, n\} \setminus \{i\}, \psi_k \in \Sigma$ and so $\psi_k \in \Gamma$ by inclusion. Consequently, we prove that $\Gamma \vdash \psi \Rightarrow \varphi$.

(2) If $\psi \notin \Sigma$, then for all $i \in \{1, \ldots, n\}, \psi_i \neq \psi$. We have

$$\vdash \bigwedge_{k \in \{1,\ldots,n\}} \psi_k \Rightarrow \varphi$$

Or since $\vdash \varphi \Rightarrow (\psi \Rightarrow \varphi)$ is a PC axiom, we immediately deduce:

$$\vdash \bigwedge_{k \in \{1,\ldots,n\}} \psi_k \Rightarrow (\psi \Rightarrow \varphi)$$

Consequently, since for all $k \in \{1, \ldots, n\}, \psi_k \in \Sigma$ and so $\psi_k \in \Gamma$ by inclusion. Thus, $\Gamma \vdash \psi \Rightarrow \varphi$.

($\Leftarrow$) Let us assume that $\Gamma \vdash \psi \Rightarrow \varphi$. So there exists $\Sigma = \{\psi_1, \ldots, \psi_n\}$, $\Sigma \subseteq \Gamma$ such that:

$$\vdash \bigwedge_{i \in \{1,\ldots,n\}} \psi_i \Rightarrow (\psi \Rightarrow \varphi)$$

Thus,

$$\vdash \bigwedge_{i \in \{1,\ldots,n\}} \psi_i \wedge \psi \Rightarrow \varphi$$

is a theorem. So we deduce that $\Sigma \cup \{\psi\} \vdash \varphi$. But since $\Sigma \subseteq \Gamma$, by left weakening, we immediately deduce that $\Gamma \cup \{\psi\} \vdash \varphi$.

The semantic version (2) of the deduction theorem immediately follows:

$\Gamma \models \psi \Rightarrow \varphi$ iff ( $\forall \mathcal{M}:$ if $\mathcal{M} \models \Gamma$, then $\mathcal{M} \models \psi \Rightarrow \varphi$) iff $\forall \mathcal{M} : \mathcal{M} \models \Gamma \Rightarrow (\psi \Rightarrow \varphi)$ iff $\forall \mathcal{M} : \mathcal{M} \models \neg\Gamma \vee \neg\psi \vee \varphi$ iff $\forall \mathcal{M} : \mathcal{M} \models \neg(\Gamma \wedge \psi) \vee \varphi$ iff $\forall \mathcal{M} : \mathcal{M} \models (\Gamma \wedge \psi) \Rightarrow \varphi$ iff ($\forall \mathcal{M}:$ if $\mathcal{M} \models \Gamma \cup \{\psi\}$, then $\mathcal{M} \models \varphi$) iff $\Gamma \cup \{\psi\} \models \varphi$. □

*A.3.3. Maximal KBE-consistent sets*

First of all let us recall some well-known results about maximal consistent sets. A set of formulas $\Sigma$ is *inconsistent* if, and only if, $\exists \psi_1, \ldots, \psi_n \in \Sigma :\vdash \neg \bigwedge_{i=1}^{n} \psi_i$. A set of formulas $\Sigma$ is *consistent* iff $\Sigma$ is not inconsistent. A set of formulas $\Gamma$ is a maximal consistent iff $\nexists \Gamma' : \Gamma \subsetneq \Gamma', \Gamma'$ consistent. It leads us to the well-known Lindenbaum's lemma: *for all consistent sets of formulas* $\Gamma$ *, there exists a set of formulas* $\Gamma'$ *s.t* $\Gamma \subseteq \Gamma'$ *and* $\Gamma'$ *is a maximal consistent set* (MCS). Let us consider a MCS $\Gamma$ and $\varphi, \psi \in \mathcal{L}_{KBE}$ two formulas:

(1) MCS1: $\Gamma \vdash \varphi \Longrightarrow \varphi \in \Gamma$
(2) MCS2: $(\varphi \in \Gamma \vee \neg\varphi \in \Gamma) \wedge \neg(\varphi \in \Gamma \wedge \neg\varphi \in \Gamma)$
(3) MCS3: $(\varphi \vee \psi \in \Gamma) \Longleftrightarrow \varphi \in \Gamma$ or $\psi \in \Gamma$
(4) MCS3': $(\varphi \wedge \psi \in \Gamma) \Longleftrightarrow \varphi \in \Gamma$ and $\psi \in \Gamma$
(5) MCS4: $[(\varphi \Rightarrow \psi \in \Gamma) \wedge (\varphi \in \Gamma)] \Longrightarrow \psi \in \Gamma$
(6) MCS5: $\vdash \varphi$ iff $\forall \Gamma$ is a MCS , $\varphi \in \Gamma$

*A.3.4. Canonical models*

In order to prove that our system is complete, we are doing a Henkin-like proof. Thus, we need to define the canonical model for our non normal system Pacuit (2017).

**Definition A.11.** A model $\mathcal{M}^c = (\mathcal{W}^c, \{\mathcal{B}_i^c\}_{i\in\mathcal{N}}, \{\mathcal{K}_i^c\}_{i\in\mathcal{N}}, \{\mathcal{E}_i^c\}_{i\in\mathcal{N}}, \{\mathcal{E}_i^{d^c}\}_{i\in\mathcal{N}}, V^c)$ is the *canonical model* for KBE if, and only if $\mathcal{M}^c$ is such that:

- $\mathcal{W}^c$ is a non-empty set of worlds where each world is a MCS,
- For all $(\mathcal{R}^c, \Box) \in \{(\mathcal{B}_i^c, B_i), (\mathcal{K}_i^c, K_i), (\mathcal{E}_i^c, E_i)\}_{i\in\mathcal{N}}$:

$$\forall i \in \mathcal{N}, \forall w, v \in \mathcal{W} : w\mathcal{R}^c v \text{ if, and only if, } \Box\varphi \in w \Rightarrow \varphi \in v$$

- $\{\mathcal{E}_i^{d^c}\}_{i\in\mathcal{N}}$ is a set of neighborhood functions such that:

$$\forall i \in \mathcal{N}, \forall w \in \mathcal{W}^c : \mathcal{E}_i^{d^c}(w) := \{|\varphi| : E_i^d\varphi \in w\} \text{ with } |\varphi| := \{w | w \in \mathcal{W}^c \wedge \varphi \in w\}$$

- $V^c : \mathcal{P} \to 2^{\mathcal{W}}$ is an interpretation function such that $\forall p \in \mathcal{P}, w \in V^c(p)$ if, and only if, $p \in w$, i.e.:

$$\forall p \in \mathcal{P}, V^c(p) = |p| \text{ with } |p| := \{w | w \in \mathcal{W}^c \wedge p \in w\}$$

The part $\mathcal{E}_i^{d^c}$ of the canonical model for KBE corresponds to the notion of *minimal canonical model for neighborhood semantics* described in Pacuit (2017). In the sequel, we use the following notations, for all $i \in \mathcal{N}, w \in \mathcal{W}^c$:

- $\mathcal{K}_i^*(w) := \{\varphi | K_i\varphi \in w\}$
- $\mathcal{B}_i^*(w) := \{\varphi | B_i\varphi \in w\}$
- $\mathcal{E}_i^*(w) := \{\varphi | E_i\varphi \in w\}$

*A.3.5. Truth lemma*

Let $\mathcal{M}^c = (\mathcal{W}^c, \{\mathcal{B}_i^c\}_{i\in\mathcal{N}}, \{\mathcal{K}_i^c\}_{i\in\mathcal{N}}, \{\mathcal{E}_i^c\}_{i\in\mathcal{N}}, \{\mathcal{E}_i^{d^c}\}_{i\in\mathcal{N}}, V^c)$ be a minimal canonical model for KBE.

**Lemma A.12.** *Let $i, j \in \mathcal{N}$ and $\varphi \in \mathcal{L}_{KBE}$ be a formula,*

*(1)* $\forall w \in \mathcal{W}^c : \neg K_i\varphi \in w \Rightarrow \mathcal{K}_i^*(w) \cup \{\neg\varphi\}$ *is KBE-consistent*
*(2)* $\forall w \in \mathcal{W}^c : \neg B_i\varphi \in w \Rightarrow \mathcal{B}_i^*(w) \cup \{\neg\varphi\}$ *is KBE-consistent*
*(3)* $\forall w \in \mathcal{W}^c : \neg E_i\varphi \in w \Rightarrow \mathcal{E}_i^*(w) \cup \{\neg\varphi\}$ *is KBE-consistent*

***Proof.*** Let $w \in \mathcal{W}^c$, $i, j \in \mathcal{N}$, $\mathcal{R} \in \{\mathcal{K}_i^c, \mathcal{B}_i^c\}$, $\Box \in \{K_i, B_i, E_i\}$ and $\varphi \in \mathcal{L}_{KBE}$ be a formula. Let us assume by contraposition that $\mathcal{R}^*(w) \cup \{\neg\varphi\}$ is KBE-inconsistent. There exists $n \in \mathbb{N}$ and $\psi_1, \ldots, \psi_n \in \mathcal{R}^*(w)$ such that:

(1) $\vdash \neg(\bigwedge_{k=1}^n \psi_k \wedge \neg\varphi)$
(2) $\vdash \neg \bigwedge_{k=1}^n \psi_k \vee \neg\neg\varphi$
(3) $\vdash \bigwedge_{k=1}^n \psi_k \Rightarrow \varphi$
(4) $\vdash \Box(\bigwedge_{k=1}^n \psi_k \Rightarrow \varphi)$
(5) $\vdash (\Box \bigwedge_{k=1}^n \psi_k \Rightarrow \Box\varphi)$
(6) $\vdash (\bigwedge_{k=1}^n \Box\psi_k \Rightarrow \Box\varphi)$
(7) $\vdash \neg(\bigwedge_{k=1}^n \Box\psi_k \wedge \neg\Box\varphi)$

So $\{\Box\psi_1, \ldots, \Box\psi_n, \neg\Box\varphi\}$ is KBE-inconsistent. However $\forall k \in \{1, \ldots, n\}$, $\psi_k \in \mathcal{R}^*(w)$ if, and only if, $\Box\psi_k \in w$ and $w$ is a maximal KBE-consistent set. Thus, $\bigwedge_{k=1}^n \Box\psi_k \in w$ (MCS3') and so $\{\Box\psi_1, \ldots, \Box\psi_n\}$ is KBE-consistent. But $\{\Box\psi_1, \ldots, \Box\psi_n\} \cup \{\neg\Box\varphi\}$ is KBE-inconsistent. So $\neg\Box\varphi$ cannot belong to the set of maximal KBE-consistent formulas $w$. Indeed by *reductio ad absurdum*, if we have $\neg\Box\varphi \in w$, we would also have that $\bigwedge_{k=1}^n \Box\psi_k \wedge \neg\Box\varphi \in w$ (by MCS3'), and $\{\Box\psi_1, \ldots, \Box\psi_n, \neg\Box\varphi\}$ would be KBE-consistent, which is a contradiction. Thus, $\neg\Box\varphi \notin w$. Consequently we prove that if $\neg\Box\varphi \in w$, then $\mathcal{R}^*(w) \cup \{\neg\varphi\}$ is KBE-consistent. □

**Lemma A.13.** *Let $i, j \in \mathcal{N}$ and $\varphi, \psi \in \mathcal{L}_{KBE}$ be a formula,*

*(1)* $|\varphi \wedge \psi| = |\varphi| \cap |\psi|$
*(2)* $|\neg\varphi| = \mathcal{W}^c \setminus |\varphi|$
*(3)* $|\varphi \vee \psi| = |\varphi| \cup |\psi|$
*(4)* $|\varphi| \subseteq |\psi|$ *iff* $\vdash \varphi \Rightarrow \psi$
*(5)* $|\varphi| = |\psi|$ *iff* $\vdash \varphi \Leftrightarrow \psi$

***Proof.*** See Pacuit (2017). □

**Lemma A.14.** *Let $w \in \mathcal{W}^c$, $\varphi, \psi \in \mathcal{L}_{KBE}$.*
*If $|\varphi| = |\psi|$ and $|\varphi| \in \mathcal{E}_i^{d^c}(w)$ then $E_i^d\psi \in w$*

***Proof.*** See Pacuit (2017). □

We need a second lemma to demonstrate the completeness of our system. This lemma shows that any valid formula of the canonical model is a formula of a maximal KBE-consistent set corresponding to a world in which it is verified and reciprocally.

The following proofs are based on the degree of formulas. We recall the basic notion of a degree of a formula.

**Definition A.15.** We define $deg : \mathcal{L}_{KBE} \rightarrow \mathbb{N}$ the degree function iff for all $\varphi, \psi \in \mathcal{L}_{KBE}$:

(1) $\forall p \in \mathcal{P}, deg(p) = 0$
(2) $deg(\neg\varphi) = deg(\varphi) + 1$
(3) $\forall \Box \in \{K_i, B_i, E_i, E_i^d\}, deg(\Box\varphi) = deg(\varphi) + 1$
(4) $deg(\varphi \wedge \psi) = max\{deg(\varphi), deg(\psi)\} + 1$
(5) $deg(\varphi \vee \psi) = max\{deg(\varphi), deg(\psi)\} + 1$
(6) $deg(\varphi \Rightarrow \psi) = max\{deg(\neg\varphi), deg(\psi)\} + 1$

We say that the degree of a formula $\varphi$ is $n \in \mathbb{N}^*$ iff $deg(\varphi) = n$.

**Lemma A.16.** *Let $\varphi \in \mathcal{L}_{KBE}$ be a formula, for all $w \in \mathcal{W}^c$ :*

$$\mathcal{M}^c, w \models \varphi \text{ if, and only if, } \varphi \in w$$

***Proof.*** Let us reason by recurrence on the degree of a formula.

(Initialization) If $\varphi \in \mathcal{L}_{KBE}$ is a formula such that $deg(\varphi) = 0$, i.e. there exists $p \in \mathcal{P}, \varphi = p$. By definition of a canonical model we have $\forall w \in \mathcal{W}^c, w \in V(p)$ if, and only if, $p \in w$.

(Heredity) Let us assume for all formulas $\varphi \in \mathcal{L}_{KBE}$ and $n \in \mathbb{N}^*$ such that $deg(\varphi) < n$, we have for all $w \in \mathcal{W}^c, \mathcal{M}^c, w \models \varphi$ if, and only if, $\varphi \in w$.

Let $\psi, \theta \in \mathcal{L}_{KBE}$ such that $max(deg(\psi), deg(\theta)) = n - 1$. So we have for all worlds $w \in \mathcal{W}^c$, $\mathcal{M}^c, w \models \psi$ iff $\psi \in w$ and $\mathcal{M}^c, w \models \theta$ iff $\theta \in w$. Furthermore we have $\mathcal{M}^c, w \models \neg\psi$ iff $\mathcal{M}^c, w \nvDash \psi$ iff $\psi \notin w$. Then $\mathcal{M}^c, w \models \psi \wedge \theta$ iff $\mathcal{M}^c, w \models \psi$ and $\mathcal{M}^c, w \models \theta$ iff $\psi \in w$ and $\theta \in w$ iff $\psi \wedge \theta \in w$ (MCS3'). Then $\mathcal{M}^c, w \models \psi \vee \theta$ iff $\mathcal{M}^c, w \models \psi$ or $\mathcal{M}^c, w \models \theta$ iff $\psi \in w$ or $\theta \in w$ iff $\psi \vee \theta \in w$ (MCS3). Finally $\mathcal{M}^c, w \models \psi \Rightarrow \theta$ iff $\mathcal{M}^c, w \models \neg\psi$ or $\mathcal{M}^c, w \models \theta$ iff $\psi \notin w$ or $\theta \in w$ iff $\psi \Rightarrow \theta \in w$.

Let us consider $(\mathcal{R}, \Box) \in \{(\mathcal{B}_i, B_i), (\mathcal{K}_i, K_i), (\mathcal{E}_i, E_i)\}$ and a world $w \in \mathcal{W}^c$. Let us show that the equivalence for normal modalities by double implication.

($\Rightarrow$) Let us assume by contraposition that $\Box\psi \notin w$ and since $w$ is a maximal KBE-consistent set, we have $\neg\Box\psi \in w$. By the previous lemma, we have $\mathcal{R}^*(w) \cup \{\neg\psi\}$ is KBE-consistent and so, by the Lindenbaum's lemma, there exists $v \in \mathcal{W}^c : \mathcal{R}^*(w) \cup \{\neg\psi\} \subseteq v$ and $v$ is maximal KBE-consistent set. So we have $\neg\psi \in v$ and, by definition of $\mathcal{R}^c$, we have $w\mathcal{R}^c v$. Furthermore, we have $\psi \notin v$ and, by induction hypothesis $\mathcal{W}^c, v \nvDash \psi$. So since there exists $v \in \mathcal{W}^c : w\mathcal{R}^c v : v \models \neg\psi$, we have $\mathcal{M}^c, w \models \neg\Box\psi$, i.e. $\mathcal{M}^c, w \nvDash \Box\psi$.

($\Leftarrow$) By contraposition, let us assume that $\mathcal{M}^c, w \nvDash \Box\psi$, i.e. $\mathcal{M}^c, w \models \neg\Box\psi$. So there exists $v \in \mathcal{W} : w\mathcal{R}^c v, \mathcal{M}^c, v \models \neg\psi$. So $\mathcal{M}^c, v \nvDash \varphi$ and by induction hypothesis, we have $\varphi \notin v$. However, since $\varphi \notin v$, by definition of $\mathcal{R}^c$, we have $\Box\varphi \notin w$.

Let us now show that equivalence is also verified when the formula is of the form $E_i^d\varphi$ and such that $deg(\varphi) = n$. Let us assume that $\mathcal{M}^c, w \models E_i^d\varphi$. In other words, this is equivalent to $\{v | v \in \mathcal{W}^c \wedge \mathcal{M}^c, v \models \varphi\} \in \mathcal{E}_i^{d^c}(w)$. Applying the induction hypothesis, we therefore have by equivalence that $\{v | v \in \mathcal{W}^c \wedge \varphi \in v\} \in \mathcal{E}_i^{d^c}(w)$, i.e. $|\varphi| \in \mathcal{E}_i^{d^c}(w)$ with $|\varphi| := \{v | v \in \mathcal{W}^c \wedge \varphi \in v\}$. Finally, by definition of the canonical model, this is equivalent to $E_i^d\varphi \in w$.

(Conclusion) So we have shown by recurrence that:

$$\forall \varphi \in \mathcal{L}_{KBE}, \forall w \in \mathcal{W}^c : \mathcal{M}^c, w \models \varphi \text{ if, and only if, } \varphi \in w$$

□

Now that the link between validity and KBE-consistent maximum sets has been demonstrated, we can prove the link between our canonical model and the proven formulas in our axiomatic system.

**Lemma A.17** (Truth lemma)**.** *Let $\varphi \in \mathcal{L}_{KBE}$ be a formula and $w \in \mathcal{W}^c$,*

$$\mathcal{M}^c, w \models \varphi \text{ if, and only if, } \vdash \varphi$$

***Proof.*** Let $\varphi \in \mathcal{L}_{KBE}$ be a formula and $w \in \mathcal{W}^c$. We have $\mathcal{M}^c, w \models \varphi$ iff (by lemma A.16) $\varphi \in w$ iff (by MCS5) $\vdash \varphi$.

□

*A.3.6. Canonical model properties*

Let us prove that the canonical model preserves the semantic constraints of KBE. Let $\mathcal{M}^c = (\mathcal{W}^c, \{\mathcal{B}_i^c\}_{i\in\mathcal{N}}, \{\mathcal{K}_i^c\}_{i\in\mathcal{N}}, \{\mathcal{E}_i^c\}_{i\in\mathcal{N}}, \{\mathcal{E}_i^{d^c}\}_{i\in\mathcal{N}}, V^c)$ be the canonical model.

**Lemma A.18.** *The canonical model $\mathcal{M}^c$ is a KBE model i.e. the following properties for the canonical model hold:*

*(1)* $\forall w \in \mathcal{W}^c : X \in \mathcal{E}_i^{d^c}(w) \wedge Y \in \mathcal{E}_i^{d^c}(w) \Longrightarrow X \cap Y \in \mathcal{E}_i^{d^c}(w)$
*(2)* $\forall w \in \mathcal{W}^c : \mathcal{E}_i^{d^c}(w) \subseteq \bigcap_{v\in\mathcal{W}:w\mathcal{K}_i^c v} \mathcal{E}_i^{d^c}(v)$
*(3)* $\forall w, v \in \mathcal{W}^c, \forall X \in 2^{\mathcal{W}} : X \notin \mathcal{E}_i^{d^c}(w) \Longrightarrow (w\mathcal{K}_i^c v \Rightarrow X \notin \mathcal{E}_i^{d^c}(v))$
*(4)* $\forall w \in \mathcal{W}^c : \mathcal{W}^c \notin \mathcal{E}_i^{d^c}(w)$
*(5)* $\forall w \in \mathcal{W} : X \in \mathcal{E}_i^{d^c}(w) \Longrightarrow (\mathcal{E}_i(w) \subseteq X)$
*(6)* $\mathcal{M}^c$ *respects the constraints for* $\mathcal{K}_i$, $\mathcal{B}_i$ *and* $\mathcal{E}_i$

***Proof.*** (1) Let us first show that:

$$\forall w \in \mathcal{W}^c : X \in \mathcal{E}_i^{d^c}(w) \wedge Y \in \mathcal{E}_i^{d^c}(w) \Longrightarrow X \cap Y \in \mathcal{E}_i^{d^c}(w)$$

Let $w \in \mathcal{W}^c$, $X \in \mathcal{E}_i^c(w)$ and $Y \in \mathcal{E}_i^{d^c}(w)$. By definition of the minimal canonical model, there exists $\varphi, \psi \in \mathcal{L}_{KBE}$ such that $X = |\varphi|$ and $Y = |\psi|$. Thus, $|\varphi| \in \mathcal{E}_i^{d^c}(w)$ and $|\psi| \in \mathcal{E}_i^{d^c}(w)$. Then, by definition, we have $E_i^d\varphi \in w$ and $E_i^d\psi \in w$. Furthermore by MCS3' we have $E_i^d\varphi \wedge E_i^d\psi \in w$. However $\vdash E_i^d\varphi \wedge E_i^d\psi \Rightarrow E_i^d(\varphi \wedge \psi)$. Thus, by MCS5 $E_i^d\varphi \wedge E_i^d\psi \Rightarrow E_i^d(\varphi \wedge \psi) \in w$ and by MCS4, we deduce that $E_i^d(\varphi \wedge \psi) \in w$. So $|\varphi \wedge \psi| \in \mathcal{E}_i^{d^c}(w)$. However $|\varphi \wedge \psi| = |\varphi| \cap |\psi| = X \cap Y$. Consequently, we prove that $X \cap Y \in \mathcal{E}_i^{d^c}(w)$.

(2) Let us show that:

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

Let $w \in \mathcal{W}^c$ and $X \in \mathcal{E}_i^{d^c}(w)$. By definition of $\mathcal{E}_i^{d^c}$, there exists $\varphi \in \mathcal{L}_{KBE}$, $X = |\varphi|$. So $E_i^d\varphi \in w$. However, $\vdash E_i^d\varphi \Rightarrow K_iE_i^d\varphi$ and, by MCS5, we deduce that $E_i^d\varphi \Rightarrow K_iE_i^d\varphi \in w$ and, by MCS4, $K_iE_i^d\varphi \in w$. Thus, by definition of the canonical model, we have for all $v \in \mathcal{W}^c, w\mathcal{K}_i^c v : \mathcal{M}^c, v \models E_i^d\varphi$. So $E_i^d\varphi \in v$, which means that $|\varphi| \in \mathcal{E}_i^{d^c}(v)$, i.e. $X \in \mathcal{E}_i^{d^c}(v)$. Consequently, we prove that $\forall v \in \mathcal{W}^c, w\mathcal{K}_i^c v : X \in \mathcal{E}_i^{d^c}(v)$, i.e. $X \in \bigcap_{v\in\mathcal{W}:w\mathcal{K}_i^c v} \mathcal{E}_i^{d^c}(v)$.

(3) Let us show that:

$$\forall w, v \in \mathcal{W}^c, \forall X \in 2^{\mathcal{W}} : X \notin \mathcal{E}_i^{d^c}(w) \Longrightarrow (w\mathcal{K}_i^c v \Rightarrow X \notin \mathcal{E}_i^{d^c}(v))$$

Let $w, v \in \mathcal{W}^c$ s.t. $w\mathcal{K}_i^c v$ and $X \notin \mathcal{E}_i^{d^c}(w)$. By definition of $\mathcal{E}_i^{d^c}(w) = \{|\varphi| : E_i^d\varphi \in w\}$, since $X \notin \mathcal{E}_i^{d^c}(w)$, there is no $\varphi \in \mathcal{L}_{KBE}$ such that $X = |\varphi|$ and $E_i^d\varphi \in w$. So, this means that for all $\varphi \in \mathcal{L}_{KBE}$, $X = |\varphi| \Rightarrow E_i^d\varphi \notin w$, and by MCS2, we have for all $\varphi \in \mathcal{L}_{KBE}$, $X = |\varphi| \Rightarrow \neg E_i^d\varphi \in w$. Thus, there are two possible cases:

- If $X$ is of the form $X = |\varphi|$, then $\neg E_i^d\varphi \in w$. However, $\vdash \neg E_i^d\varphi \Rightarrow K_i\neg E_i^d\varphi$ and, by MCS5, $\neg E_i^d\varphi \Rightarrow K_i\neg E_i^d\varphi \in w$ then, by MCS4, $K_i\neg E_i^d\varphi \in w$. Thus, $\forall u \in \mathcal{W}^c$ s.t. $w\mathcal{K}_i^c u, \neg E_i^d\varphi \in u$. Finally, $\forall u \in \mathcal{W}^c$ s.t. $w\mathcal{K}_i^c u, |\varphi| \notin \mathcal{E}_i^{d^c}(u)$, and thus $X \notin \mathcal{E}_i^{d^c}(v)$.
- If $X$ cannot be written as $X = |\varphi|$, then by using the definition of the minimal canonical model, i.e. $\mathcal{E}_i^{d^c}(w) = \{|\varphi| : E_i^d\varphi \in w\}$, we deduce that $\forall u \in \mathcal{W}^c : X \notin \mathcal{E}_i^{d^c}(u)$. So $X \notin \mathcal{E}_i^{d^c}(v)$.

Consequently, we prove that:

$$\forall w, v \in \mathcal{W}^c, \forall X \in 2^{\mathcal{W}} : X \notin \mathcal{E}_i^{d^c}(w) \Longrightarrow (w\mathcal{K}_i^c v \Rightarrow X \notin \mathcal{E}_i^{d^c}(v))$$

(4) Let us show that:

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

Let $w \in \mathcal{W}^c$. We have $\vdash \neg E_i^d\top$ and, by MCS5, we have $\neg E_i\top \in w$, i.e. $|\top| \notin \mathcal{E}_i^{d^c}(w)$. However, since $|\top| = \mathcal{W}^c$, we deduce that $\mathcal{W}^c \notin \mathcal{E}_i^{d^c}(w)$.

(5) Let us show that:

$$\forall w \in \mathcal{W} : X \in \mathcal{E}_i^{d^c}(w) \Longrightarrow (\mathcal{E}_i(w) \subseteq X)$$

Let $w \in \mathcal{W}^c$ and $X \in \mathcal{E}_i^{d^c}(w)$. By definition of $\mathcal{E}_i^{d^c}(w) = \{|\varphi| : E_i\varphi \in w\}$, there exists $\varphi \in \mathcal{L}_{KBE}$ s.t. $X = |\varphi|$. So, $E_i^d\varphi \in w$. However, $\vdash E_i^d\varphi \Rightarrow E_i\varphi$ and, by MCS5, $E_i^d\varphi \Rightarrow E_i\varphi \in w$, then, by MCS4, we have $E_i\varphi \in w$, i.e. $\mathcal{M}^c, w \models E_i\varphi$ and so $\forall u \in \mathcal{W}^c : u \in \mathcal{E}_i^c(w), \mathcal{M}^c, u \models \varphi$. But since $X = |\varphi|$, then we have $\forall u \in \mathcal{E}_i^c(w), u \in |\varphi|$. Thus, if $v \in \mathcal{E}_i^c(w)$, then $v \in |\varphi|$ i.e. $v \in X$. So we prove that $\forall w \in \mathcal{W} : X \in \mathcal{E}_i^{d^c}(w) \Longrightarrow (\mathcal{E}_i(w) \subseteq X)$.

(6) The other canonical properties are easy to show and standard for $\mathcal{K}_i$, $\mathcal{B}_i$ and $\mathcal{E}_i$ (see Blackburn et al. (2002)).

□

*A.3.7. Completeness proof*

In this section, we give the complete proof of completeness.

**Theorem 4.5.** *The KBE system is complete.*

***Proof.*** (Completeness) By definition, we have that for all valid formulas $\varphi$ in the frame $\mathcal{C}$, $\varphi$ is valid in all models $\mathcal{M}$ on $\mathcal{C}$. Thus, since the canonical model is a KBE model from Lemma A.18, $\varphi$ is also valid in the canonical model $\mathcal{M}^c$. So, from the

Lemma A.17, we have $\vdash \varphi$. Consequently, we have proved that the KBE system is complete. □

### *A.4. Strong properties of KBE*

We prove here the strong soundness and strong completeness of KBE.

#### *A.4.1. Strong soundness*

In the following, we demonstrate the strong correction of KBE. The strong correction is an almost immediate consequence of the correction and the semantic weakening that we demonstrate in the proof of the theorem.

**Theorem A.19** (Strongly soundness of KBE)**.** *Let $\Gamma$ be a set of formulas of KBE, $\varphi$ be a theorem of KBE, we have that the system KBE is strongly sound, i.e. if $\Gamma \vdash \varphi$ then $\Gamma \models \varphi$.*

***Proof.*** Let $\Gamma$ be a set of formulas of KBE, $\varphi$ be a formula of KBE. Let us assume that $\Gamma \vdash \varphi$, so there exists $\psi_1, \ldots, \psi_n \in \Gamma$ such that $\vdash \psi_1 \wedge \ldots \wedge \psi_n \Rightarrow \varphi$. Thus, by soundness we have $\models \psi_1 \wedge \ldots \wedge \psi_n \Rightarrow \varphi$ i.e. $\models \neg\psi_1 \vee \ldots \vee \neg\psi_n \vee \varphi$. So:

$$\models \neg\psi_1 \vee \ldots \vee \neg\psi_n \vee \neg \bigwedge_{\theta \in \Gamma} \theta \vee \varphi$$

Thus, by applying the rule of De Morgan, we have:

$$\models \neg \bigwedge_{\theta \in \Gamma \cup \{\psi_1, \ldots, \psi_n\}} \theta \vee \varphi$$

However since $\{\psi_1, \ldots, \psi_n\} \subseteq \Gamma$, we have $\Gamma \cup \{\psi_1, \ldots, \psi_n\} = \Gamma$ and so (semantic weakening):

$$\models \neg \bigwedge_{\theta \in \Gamma} \theta \vee \varphi$$

Consequently, we have $\forall \mathcal{M},$ if $\mathcal{M} \models \Gamma,$ then $\mathcal{M} \models \varphi$. Thus, we prove that $\Gamma \models \varphi$. □

#### *A.4.2. Strong completeness*

The KBE system is strongly complete. In order to demonstrate the strong completeness of the system, we need the following lemma A.20.

**Lemma A.20.** *For all KBE-consistent sets $\Gamma$ of formulas, there exists a world $w \in \mathcal{W}^c$ in the canonical model $\mathcal{M}^c$ such that $\mathcal{M}^c, w \models \Gamma$, i.e. $\forall \varphi \in \Gamma : \mathcal{M}^c, w \models \varphi$.*

***Proof.*** Let $\mathcal{M}^c = (\mathcal{W}^c, (\mathcal{K}_i^c)_{i \in \mathcal{N}}, (\mathcal{B}_i^c)_{i \in \mathcal{N}}, (\mathcal{E}_i^c)_{i \in \mathcal{N}}, (\mathcal{E}_i^{d^c})_{i \in \mathcal{N}}, V^c)$ be the canonical model. Let $\Gamma$ be a KBE-consistent set of formulas. By applying the lemma of Lindenbaum, there exists a maximal consistent set of formulas $\Gamma'$ such that $\Gamma \subseteq \Gamma'$ and $\Gamma' \in \mathcal{W}^c$. Let $w = \Gamma'$ denotes the possible world in $\mathcal{W}^c$. We have $\forall \varphi \in \Gamma' : \mathcal{M}^c, \Gamma' \models \varphi$, and so $\forall \varphi \in \Gamma : \mathcal{M}^c, \Gamma \models \varphi$ i.e. $\forall \varphi \in \Gamma : \mathcal{M}^c, w \models \varphi$. Thus, we have proved that for

all KBE-consistent sets $\Gamma$, there exists a world $w \in \mathcal{W}^c$ satisfying all formulas of $\Gamma$ in the canonical model $\mathcal{M}^c$. □

Finally we give the proof of the strong completeness of KBE.

**Theorem A.21** (Strong completeness of KBE)**.** *Let $\varphi \in \mathcal{L}_{KBE}$ be a formula and for all sets $\Gamma \subseteq \mathcal{L}_{KBE}$ of formulas, we have that the system KBE is strongly complete i.e. if $\Gamma \models \varphi$, then $\Gamma \vdash \varphi$.*

***Proof.*** By contraposition, let $\Gamma \subseteq \mathcal{L}_{KBE}$ be a set of formulas such that $\Gamma \nvdash \varphi$, we have that $\Gamma \cup \{\neg\varphi\}$ is a KBE-consistent set. Indeed, by absurdity, if we have $\Gamma \cup \{\neg\varphi\}$ is KBE-inconsistent, then we would have that there exists $\psi_1, \ldots, \psi_n \in \Gamma$ such that $\vdash \neg(\psi_1 \wedge \ldots \wedge \psi_n \wedge \neg\varphi)$, and so we would also have $\vdash (\psi_1 \wedge \ldots \wedge \psi_n) \Rightarrow \varphi$. However by deduction theorem, we would deduce that $\Gamma \cup \{\psi_1, \ldots, \psi_n\} \vdash \varphi$, i.e. $\Gamma \vdash \varphi$, which contradicts the hypothesis $\Gamma \nvdash \varphi$. Thus, by lemma A.20, there exists a model $\mathcal{M}$ (the canonical model) and a world $w$ such that $\mathcal{M}, w \models \Gamma \cup \{\neg\varphi\}$, i.e. $\mathcal{M}, w \models \Gamma$ and $\mathcal{M}, w \models \neg\varphi$. Consequently we have proved that there exists a model $\mathcal{M}$ such that $\mathcal{M}, \Gamma \not\models \varphi$. □

### *A.5. Theorems of KBE*

In Section 4.6, we gave theorems that can be deduced with KBE such as the property D, T, concealing contrary beliefs or knowledge and the *qui facit per alium facit per se* principle. In this section we give complete Hilbert-style proof of these theorems.

**Theorem 4.6** (Statements 1 and 2)**.**

*(1)* $\vdash \neg E_i^d \bot \quad (D_{E_i^d})$
*(2)* $\vdash E_i^d \varphi \Rightarrow \varphi \quad (T_{E_i^d})$

***Proof.***

| | | |
|---|---|---|
| (1) | (a) $\vdash E_i^d \bot \Rightarrow E_i \bot$ | [Ax. $(E_i^d E_i)$] |
| | (b) $\vdash (E_i^d \bot \Rightarrow E_i \bot) \Rightarrow (\neg E_i \bot \Rightarrow \neg E_i^d \bot)$ | [Contra. on (a)] |
| | (c) $\vdash \neg E_i \bot \Rightarrow \neg E_i^d \bot$ | [MP (a), (b)] |
| | (d) $\vdash \neg E_i \bot$ | [Ax. $(D_{E_i})$] |
| | (e) $\vdash \neg E_i^d \bot$ | [MP (d), (c)] |
| (2) | (a) $\vdash E_i^d \varphi \Rightarrow E_i \varphi$ | [Ax. $(E_i^d E_i)$] |
| | (b) $\vdash E_i \varphi \Rightarrow \varphi$ | [Ax. $(T_{E_i})$] |
| | (c) $\vdash E_i^d \varphi \Rightarrow E_i \varphi \Rightarrow \varphi$ | [Aug. (b)] |
| | (d) $\vdash (E_i^d \varphi \Rightarrow E_i \varphi \Rightarrow \varphi) \Rightarrow ((E_i^d \varphi \Rightarrow E_i \varphi) \Rightarrow (E_i^d \varphi \Rightarrow \varphi))$ | [Syll. (c)] |
| | (e) $\vdash E_i^d \varphi \Rightarrow \varphi$ | [MP (c) (a) (d)] |

□

**Theorem 4.6** (Statements 3 to 10)**.**

*(3)* $\vdash K_i E_i^d \varphi \Leftrightarrow E_i^d \varphi$
*(4)* $\vdash K_i \neg E_i^d \varphi \Leftrightarrow \neg E_i^d \varphi$
*(5)* $\vdash \neg K_i E_i^d \varphi \Leftrightarrow \neg E_i^d \varphi$
*(6)* $\vdash \neg K_i \neg E_i^d \varphi \Leftrightarrow E_i^d \varphi$
*(7)* $\vdash \neg B_i \neg E_i^d \varphi \Leftrightarrow E_i^d \varphi$
*(8)* $\vdash B_i \neg E_i^d \varphi \Leftrightarrow \neg E_i^d \varphi$

*(9)* $\vdash B_i E_i^d \varphi \Leftrightarrow E_i^d \varphi$
*(10)* $\vdash \neg B_i E_i^d \varphi \Leftrightarrow \neg E_i^d \varphi$

***Proof.***

(3) (a) $(\Rightarrow)$ $\vdash K_i E_i^d \varphi \Rightarrow E_i^d \varphi$ [Ax. $(T_{K_i})$]
(b) $(\Leftarrow)$ $\vdash E_i^d \varphi \Rightarrow K_i E_i^d \varphi$ [Ax. $(4_{K_i, E_i^d})$]
(c) $\vdash K_i E_i^d \varphi \Leftrightarrow E_i^d \varphi$ [Def. $(\Leftrightarrow)$]
(4) (a) $(\Rightarrow)$ $\vdash K_i \neg E_i^d \varphi \Rightarrow \neg E_i^d \varphi$ [Ax. $(T_{K_i})$]
(b) $(\Leftarrow)$ $\vdash \neg E_i^d \varphi \Rightarrow K_i \neg E_i^d \varphi$ [Ax. $(5_{K_i, E_i^d})$]
(c) $\vdash K_i \neg E_i^d \varphi \Leftrightarrow \neg E_i^d \varphi$ [Def. $(\Leftrightarrow)$]
(5) (a) $\vdash \neg K_i E_i^d \varphi \Leftrightarrow \neg E_i^d \varphi$ [Contrap. (1.c)]
(6) (a) $\vdash \neg K_i \neg E_i^d \varphi \Leftrightarrow E_i^d \varphi$ [Contrap. (2.c)]
(7) (a) $(\Rightarrow) \vdash \neg B_i \neg E_i^d \varphi \Rightarrow \neg K_i \neg E_i^d \varphi$ [Contrap. $(K_i B_i)$]
(b) $(\Rightarrow) \vdash \neg B_i \neg E_i^d \varphi \Rightarrow E_i^d \varphi$ [RE (3.a) in (a)]
(c) $(\Rightarrow) \vdash \neg B_i \neg E_i^d \varphi \Rightarrow \neg B_i \neg E_i^d \varphi \Rightarrow E_i^d \varphi$ [Aug. (b)]
(d) $(\Rightarrow) \vdash (\neg B_i \neg E_i^d \varphi \Rightarrow \neg B_i \neg E_i^d \varphi \Rightarrow E_i^d \varphi) \Rightarrow ((\neg B_i \neg E_i^d \varphi \Rightarrow \neg B_i \neg E_i^d \varphi) \Rightarrow (\neg B_i \neg E_i^d \varphi \Rightarrow E_i^d \varphi))$ [Syll. (c)]
(e) $(\Rightarrow) \vdash \neg B_i \neg E_i^d \varphi \Rightarrow E_i^d \varphi$ [MP (c) (a) (b)]
(f) $(\Leftarrow) \vdash E_i^d \varphi \Rightarrow K_i E_i^d \varphi$ [Ax. $(4_{K_i, E_i^d})$]
(g) $(\Leftarrow) \vdash K_i E_i^d \varphi \Rightarrow B_i E_i^d \varphi$ [Ax. $(K_i B_j)$]
(h) $(\Leftarrow) \vdash E_i^d \varphi \Rightarrow K_i E_i^d \varphi \Rightarrow B_i E_i^d \varphi$ [Aug. (g)]
(i) $(\Leftarrow) \vdash (E_i^d \varphi \Rightarrow K_i E_i^d \varphi \Rightarrow B_i E_i^d \varphi) \Rightarrow (E_i^d \varphi \Rightarrow K_i E_i^d \varphi) \Rightarrow (E_i^d \varphi \Rightarrow B_i E_i^d \varphi)$ [Syll. (h)]
(j) $(\Leftarrow) \vdash E_i^d \varphi \Rightarrow B_i E_i^d \varphi$ [MP (h) (i) (g)]
(k) $(\Leftarrow) \vdash B_i E_i^d \varphi \Rightarrow \neg B_i \neg E_i^d \varphi$ [Ax. $(D_{B_i})$]
(l) $(\Leftarrow) \vdash E_i^d \varphi \Rightarrow K_i E_i^d \varphi \Rightarrow B_i E_i^d \varphi \Rightarrow \neg B_i \neg E_i^d \varphi$ [Aug. (k)]
(m) $(\Leftarrow) \vdash (E_i^d \varphi \Rightarrow K_i E_i^d \varphi \Rightarrow B_i E_i^d \varphi \Rightarrow \neg B_i \neg E_i^d \varphi) \Rightarrow (E_i^d \varphi \Rightarrow K_i E_i^d \varphi) \Rightarrow (E_i^d \varphi \Rightarrow B_i E_i^d \varphi) \Rightarrow (E_i^d \varphi \Rightarrow \neg B_i \neg E_i^d \varphi)$ [Syll. (l)]
(n) $(\Leftarrow) \vdash E_i^d \varphi \Rightarrow \neg B_i \neg E_i^d \varphi$ [MP (l) (f) (j) (m)]
(o) $\vdash E_i^d \varphi \Leftrightarrow \neg B_i \neg E_i^d \varphi$ [Def. $(\Leftrightarrow)$]
(8) (a) $\vdash \neg E_i^d \varphi \Leftrightarrow B_i \neg E_i^d \varphi$ [Contrap. (5.o)]
(9) (a) $(\Rightarrow) \vdash B_i E_i^d \varphi \Rightarrow \neg B_i \neg E_i^d \varphi$ [Ax. $(D_{B_i})$]
(b) $(\Rightarrow) \vdash B_i E_i^d \varphi \Rightarrow E_i^d \varphi$ [RE (5.o)]
(c) $(\Leftarrow) \vdash E_i^d \varphi \Rightarrow B_i E_i^d \varphi$ [Th. (5.j)]
(d) $\vdash E_i^d \varphi \Leftrightarrow B_i E_i^d \varphi$ [Def. $(\Leftrightarrow)$]
(10) (a) $\vdash \neg E_i^d \varphi \Leftrightarrow \neg B_i E_i^d \varphi$ [Contrap. (7.$d$)]

□

**Theorem 4.7.**

*(1)* $\vdash E_i^d B_j \varphi \Rightarrow E_i \neg B_j K_k \neg \varphi$
*(2)* $\vdash E_i^d \neg B_j \varphi \Rightarrow E_i \neg B_j K_k \varphi$

***Proof.*** Here is the proof of (1).

(1) $\vdash B_j \varphi \wedge B_j K_k \neg \varphi \Rightarrow B_j \varphi$ [Left Elim. $\wedge$]
(2) $\vdash B_j (K_k \neg \varphi \Rightarrow \neg \varphi)$ [Nec. $B_j$ on $(T_{K_k})$]
(3) $\vdash B_j (K_k \neg \varphi \Rightarrow \neg \varphi) \Rightarrow (B_j K_k \neg \varphi \Rightarrow B_j \neg \varphi)$ [Ax. $(K)$]
(4) $\vdash B_j K_k \neg \varphi \Rightarrow B_j \neg \varphi$ [MP (2), (3)]
(5) $\vdash B_j \neg \varphi \Rightarrow \neg B_j \varphi$ [Ax. (D)]

(6) $\vdash B_j K_k \neg\varphi \Rightarrow B_j \neg\varphi \Rightarrow \neg B_j\varphi$ [Aug. (5)]
(7) $\vdash (B_j K_k \neg\varphi \Rightarrow B_j \neg\varphi \Rightarrow \neg B_j\varphi) \Rightarrow ((B_j K_k \neg\varphi \Rightarrow B_j \neg\varphi) \Rightarrow (B_j K_k \neg\varphi \Rightarrow \neg B_j\varphi))$ [Syll. (6)]
(8) $\vdash B_j K_k \neg\varphi \Rightarrow \neg B_j\varphi$ [MP (6),(4),(7)]
(9) $\vdash B_j\varphi \wedge B_j K_k \neg\varphi \Rightarrow (B_j K_k \neg\varphi \Rightarrow \neg B_j\varphi)$ [Aug. (8)]
(10) $\vdash B_j\varphi \wedge B_j K_k \neg\varphi \Rightarrow B_j K_k \neg\varphi$ [Right Elim. ($\wedge$) ]
(11) $\vdash (B_j\varphi \wedge B_j K_k \neg\varphi \Rightarrow (B_j K_k \neg\varphi \Rightarrow \neg B_j\varphi)) \Rightarrow ((B_j\varphi \wedge B_j K_k \neg\varphi \Rightarrow B_j K_k \neg\varphi) \Rightarrow (B_j\varphi \wedge B_j K_k \neg\varphi \Rightarrow \neg B_j\varphi)))$ [Syll. (9)]
(12) $\vdash B_j\varphi \wedge B_j K_k \neg\varphi \Rightarrow \neg B_j\varphi$ [MP (9),(10),(11)]
(13) $\vdash ((B_j\varphi \wedge B_j K_k \neg\varphi \Rightarrow B_j\varphi)) \Rightarrow ((B_j\varphi \wedge B_j K_k \neg\varphi \Rightarrow \neg B_j\varphi) \Rightarrow \neg(B_j\varphi \wedge B_j K_k \neg\varphi))$ [*Reductio ad absurdum*]
(14) $\vdash \neg(B_j\varphi \wedge B_j K_k \neg\varphi)$ [MP (1), (12), (13)]
(15) $\vdash \neg(B_j\varphi \wedge B_j K_k \neg\varphi) \equiv (B_j\varphi \Rightarrow \neg B_j K_k \neg\varphi)$ [De Morgan Th. (14)]
(16) $\vdash B_j\varphi \Rightarrow \neg B_j K_k \neg\varphi$ [MP (14), (15)]
(17) $\vdash E_i(B_j\varphi \Rightarrow \neg B_j K_k \neg\varphi)$ [Nec. (16)]
(18) $\vdash E_i(B_j\varphi \Rightarrow \neg B_j K_k \neg\varphi) \Rightarrow (E_i B_j\varphi \Rightarrow E_i \neg B_j K_k \neg\varphi)$ [Ax. (K)]
(19) $\vdash E_i B_j\varphi \Rightarrow E_i \neg B_j K_k \neg\varphi$ [MP (17), (18)]
(20) $\vdash E_i^d B_j\varphi \Rightarrow E_i B_j\varphi$ [Ax. ($E_i^d E_i$)]
(21) $\vdash E_i^d B_j\varphi \Rightarrow E_i B_j\varphi \Rightarrow E_i \neg B_j K_k \neg\varphi$ [Aug. (20)]
(22) $\vdash (E_i^d B_j\varphi \Rightarrow E_i B_j\varphi \Rightarrow E_i \neg B_j K_k \neg\varphi) \Rightarrow ((E_i^d B_j\varphi \Rightarrow E_i B_j\varphi) \Rightarrow (E_i^d B_j\varphi \Rightarrow E_i \neg B_j K_k \neg\varphi))$ [Syll. (21)]
(23) $\vdash E_i^d B_j\varphi \Rightarrow E_i \neg B_j K_k \neg\varphi$ [Ax. ($E_i^d E_i$)]

Here is the proof of (2).

(1) $\vdash K_k\varphi \Rightarrow \varphi$ [Ax. (T)]
(2) $\vdash B_j(K_k\varphi \Rightarrow \varphi)$ [Nec (1)]
(3) $\vdash B_j(K_k\varphi \Rightarrow \varphi) \Rightarrow B_j K_k\varphi \Rightarrow B_j\varphi$ [Ax. ($K_{B_j}$)]
(4) $\vdash B_j K_k\varphi \Rightarrow B_j\varphi$ [MP (2) (3)]
(5) $\vdash (B_j K_k\varphi \Rightarrow B_j\varphi) \Rightarrow \neg B_j\varphi \Rightarrow \neg B_j K_k\varphi$ [Contrap. (4)]
(6) $\vdash \neg B_j\varphi \Rightarrow \neg B_j K_k\varphi$ [MP (4) (5)]
(7) $\vdash E_i(\neg B_j\varphi \Rightarrow \neg B_j K_k\varphi)$ [Nec (6)]
(8) $\vdash E_i(\neg B_j\varphi \Rightarrow \neg B_j K_k\varphi) \Rightarrow E_i \neg B_j\varphi \Rightarrow E_i \neg B_j K_k\varphi$ [Ax. ($K_{E_i}$)]
(9) $\vdash E_i \neg B_j\varphi \Rightarrow E_i \neg B_j K_k\varphi$ [MP (7) (8)]
(10) $\vdash E_i^d \neg B_j\varphi \Rightarrow E_i \neg B_j\varphi$ [Ax. ($E_i^d E_i$)]
(11) $\vdash E_i^d \neg B_j\varphi \Rightarrow E_i \neg B_j\varphi \Rightarrow E_i \neg B_j K_k\varphi$ [Aug. (9)]
(12) $\vdash (E_i^d \neg B_j\varphi \Rightarrow E_i \neg B_j\varphi \Rightarrow E_i \neg B_j K_k\varphi) \Rightarrow (E_i^d \neg B_j\varphi \Rightarrow E_i \neg B_j\varphi) \Rightarrow E_i^d \neg B_j\varphi \Rightarrow E_i \neg B_j K_k\varphi$ [Syll. (11)]
(13) $\vdash E_i^d \neg B_j\varphi \Rightarrow E_i \neg B_j K_k\varphi$ [MP (11) (10) (12)]

□

**Theorem 4.8.**

*(1)* $\vdash E_i^d B_j\varphi \Rightarrow \neg E_i^d B_j K_k \neg\varphi$
*(2)* $\vdash E_i^d \neg B_j\varphi \Rightarrow \neg E_i^d B_j K_k\varphi$

***Proof.***
Here is the proof of (1).

(1) $\vdash E_i \neg B_j K_k \neg\varphi \Rightarrow \neg E_i B_j K_k \neg\varphi$ [Ax. (D)]
(2) $\vdash \neg E_i B_j K_k \neg\varphi \Rightarrow \neg E_i^d B_j K_k \neg\varphi$ [Contrap. Ax. ($E_i E_i^d$)]

(3) $\vdash E_i^d B_j\varphi \Rightarrow \neg E_i B_j K_k \neg\varphi \Rightarrow \neg E_i^d B_j K_k \neg\varphi$ [Aug. (2)]
(4) $\vdash (E_i^d B_j\varphi \Rightarrow \neg E_i B_j K_k \neg\varphi \Rightarrow \neg E_i^d B_j K_k \neg\varphi) \Rightarrow (E_i^d B_j\varphi \Rightarrow \neg E_i B_j K_k \neg\varphi) \Rightarrow E_i^d B_j\varphi \Rightarrow \neg E_i^d B_j K_k \neg\varphi$ [Syll. (3)]
(5) $\vdash E_i^d B_j\varphi \Rightarrow \neg E_i^d B_j K_k \neg\varphi$ [MP (3) (Syll. ((Thm 4.7.1)-(1)))) (4)]

Here is the proof of (2).

(1) $\vdash E_i \neg B_j K_k\varphi \Rightarrow \neg E_i B_j K_k\varphi$ [Ax. $(D_{E_i})$]
(2) $\vdash \neg E_i B_j K_k\varphi \Rightarrow \neg E_i^d B_j K_k\varphi$ [Contrap. $(E_i^d E_i)$]
(3) $\vdash E_i^d \neg B_j\varphi \Rightarrow E_i \neg B_j K_k\varphi \Rightarrow \neg E_i^d B_j K_k\varphi$ [Aug. (2)]
(4) $\vdash (E_i^d \neg B_j\varphi \Rightarrow E_i \neg B_j K_k\varphi \Rightarrow \neg E_i^d B_j K_k\varphi) \Rightarrow (E_i^d \neg B_j\varphi \Rightarrow E_i \neg B_j K_k\varphi) \Rightarrow E_i^d \neg B_j\varphi \Rightarrow \neg E_i^d B_j K_k\varphi$ [Syll. (3)]
(5) $\vdash E_i^d \neg B_j\varphi \Rightarrow \neg E_i^d B_j K_k\varphi$ [MP (3) (Thm 4.7.2) (4)]

□

The next theorem is the *qui facit per alium facit per se* principle.

**Theorem 4.9** (Qui facit per alium facit per se)**.**

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

***Proof.***
**Left part:**

(1) $\vdash E_i^d E_j\varphi \Rightarrow E_i E_j\varphi$ [Ax. $(E_i^d E_i)$]
(2) $\vdash E_j\varphi \Rightarrow \varphi$ [Ax. (T)]
(3) $\vdash E_i(E_j\varphi \Rightarrow \varphi)$ [Nec $(E_i)$]
(4) $\vdash E_i(E_j\varphi \Rightarrow \varphi) \Rightarrow E_i E_j\varphi \Rightarrow E_i\varphi$ [Ax. (K)]
(5) $\vdash E_i E_j\varphi \Rightarrow E_i\varphi$ [MP (3) (4)]
(6) $\vdash (E_i^d E_j\varphi \Rightarrow E_i E_j\varphi \Rightarrow E_i\varphi) \Rightarrow (E_i^d E_j\varphi \Rightarrow E_i E_j\varphi) \Rightarrow E_i^d E_j\varphi \Rightarrow E_i\varphi$ [Syll. (5)]
(7) $\vdash E_i^d E_j\varphi \Rightarrow E_i\varphi$ [MP (4) (1) (6)]

**Right part:**

(1) $\vdash E_i^d E_j^d\varphi \Rightarrow E_i E_j^d\varphi$ [Ax. $(E_i^d E_i)$]
(2) $\vdash E_i(E_j^d\varphi \Rightarrow E_j\varphi)$ [Nec (Ax. $(E_i^d E_i)$)]
(3) $\vdash E_i(E_j^d\varphi \Rightarrow E_j\varphi) \Rightarrow E_i E_j^d\varphi \Rightarrow E_i E_j\varphi$ [Ax. (K)]
(4) $\vdash E_i E_j^d\varphi \Rightarrow E_i E_j\varphi$ [MP (2) (3)]
(5) $\vdash E_i^d E_j^d\varphi \Rightarrow E_i E_j^d\varphi \Rightarrow E_i E_j\varphi$ [Aug. (4)]
(6) $\vdash (E_i^d E_j^d\varphi \Rightarrow E_i E_j^d\varphi \Rightarrow E_i E_j\varphi) \Rightarrow (E_i^d E_j^d\varphi \Rightarrow E_i E_j^d\varphi) \Rightarrow E_i^d E_j^d\varphi \Rightarrow E_i E_j\varphi$ [Syll. (5)]
(7) $\vdash E_i^d E_j^d\varphi \Rightarrow E_i E_j\varphi$ [MP (5) (1) (6)]
(8) $\vdash E_i^d E_j^d\varphi \Rightarrow E_i E_j\varphi \Rightarrow E_i\varphi$ [Aug. (Ax. $(4_{E_i})$)]
(9) $\vdash (E_i^d E_j^d\varphi \Rightarrow E_i E_j\varphi \Rightarrow E_i\varphi) \Rightarrow (E_i^d E_j^d\varphi \Rightarrow E_i E_j\varphi) \Rightarrow E_i^d E_j^d\varphi \Rightarrow E_i\varphi$ [Syll. (8)]
(10) $\vdash E_i^d E_j^d\varphi \Rightarrow E_i\varphi$ [MP (8) (7) (9)]

**Conclusion:**

(1) $\vdash (E_i^d E_j\varphi \vee E_i^d E_j^d\varphi) \Rightarrow ((E_i^d E_j\varphi \Rightarrow E_i\varphi) \Rightarrow (E_i^d E_j^d\varphi \Rightarrow E_i\varphi) \Rightarrow E_i\varphi))$ [Elim. $(\vee)$]
(2) $\vdash ((E_i^d E_j\varphi \vee E_i^d E_j^d\varphi) \Rightarrow ((E_i^d E_j\varphi \Rightarrow E_i\varphi) \Rightarrow (E_i^d E_j^d\varphi \Rightarrow E_i\varphi) \Rightarrow E_i\varphi)) \Rightarrow ((E_i^d E_j\varphi \vee E_i^d E_j^d\varphi) \Rightarrow (E_i^d E_j\varphi \Rightarrow E_i\varphi)) \Rightarrow (E_i^d E_j\varphi \vee E_i^d E_j^d\varphi) \Rightarrow (E_i^d E_j^d\varphi \Rightarrow$

$E_i\varphi) \Rightarrow E_i\varphi$ [Syll. (1)]

(3) $\vdash (E_i^d E_j\varphi \vee E_i^d E_j^d\varphi) \Rightarrow (E_i^d E_j^d\varphi \Rightarrow E_i\varphi) \Rightarrow E_i\varphi$ [MP (1) (Aug. $(T_{E_i^d})$) (2)]

(4) $\vdash ((E_i^d E_j\varphi \vee E_i^d E_j^d\varphi) \Rightarrow (E_i^d E_j^d\varphi \Rightarrow E_i\varphi) \Rightarrow E_i\varphi) \Rightarrow ((E_i^d E_j\varphi \vee E_i^d E_j^d\varphi) \Rightarrow (E_i^d E_j^d\varphi \Rightarrow E_i\varphi)) \Rightarrow (E_i^d E_j\varphi \vee E_i^d E_j^d\varphi) \Rightarrow E_i\varphi$ [Syll. (3)]

(5) $(E_i^d E_j\varphi \vee E_i^d E_j^d\varphi) \Rightarrow E_i\varphi$ [MP (3) (Aug. $(\vdash E_i^d E_j^d\varphi \Rightarrow E_i\varphi)$) (4)]

□

### *A.6. Properties of manipulation, coercion, persuasion and deception*

In Sections 5.2 and 5.3, we provide theorems for manipulation and the related notions, i.e. coercion, perusasion and deception. In this section we give the proofs of these theorems. In some cases, the Hilbert proofs need several steps of obvious syllogisms and rewritings that make the proofs difficult to read. In those cases and for ease of reading, we allow ourselves to denote those steps by "...".

**Theorem 5.5.**

*(1)* $\vdash (MCE^d K_{i,j}^{\Sigma}(\varphi) \vee MCE^d B_{i,j}^{\Sigma}(\varphi)) \wedge K_j\neg\varphi \Rightarrow \bot$

*(2)* $\vdash (MCE^d K_{i,j}^{\Sigma}(\varphi) \vee MCE^d B_{i,j}^{\Sigma}(\varphi)) \wedge B_j\neg\varphi \Rightarrow \bot$

*(3)* $\vdash (MCE^d K_{i,j}^{\Sigma}(\varphi) \vee MCE^d B_{i,j}^{\Sigma}(\varphi)) \Rightarrow K_j\varphi$

***Proof.***

(1)

(a) $\vdash (MCE^d K_{i,j}^{\Sigma}(\varphi) \vee MCE^d B_{i,j}^{\Sigma}(\varphi)) \wedge K_j\neg\varphi \Rightarrow K_j\neg\varphi$ [Right elim. (∧)]

(b) $\vdash K_j\neg\varphi \Rightarrow \neg\varphi$ [Ax. (T)]

(c) $\vdash (MCE^d K_{i,j}^{\Sigma}(\varphi) \vee MCE^d B_{i,j}^{\Sigma}(\varphi)) \wedge K_j\neg\varphi \Rightarrow (K_j\neg\varphi \Rightarrow \neg\varphi)$ [Aug. (b)]

(d) $\vdash (MCE^d K_{i,j}^{\Sigma}(\varphi) \vee MCE^d B_{i,j}^{\Sigma}(\varphi)) \wedge K_j\neg\varphi \Rightarrow (K_j\neg\varphi \Rightarrow \neg\varphi)) \Rightarrow ((MCE^d K_{i,j}^{\Sigma}(\varphi) \vee MCE^d B_{i,j}^{\Sigma}(\varphi)) \wedge K_j\neg\varphi \Rightarrow K_j\neg\varphi) \Rightarrow (((MCE^d K_{i,j}^{\Sigma}(\varphi) \vee MCE^d B_{i,j}^{\Sigma}(\varphi)) \wedge K_j\neg\varphi \Rightarrow \neg\varphi)$ [Syll. (c)]

(e) $\vdash ((MCE^d K_{i,j}^{\Sigma}(\varphi) \vee MCE^d B_{i,j}^{\Sigma}(\varphi)) \wedge K_j\neg\varphi) \Rightarrow \neg\varphi$ [MP (c),(a),(d)]

(f) $\vdash ((MCE^d K_{i,j}^{\Sigma}(\varphi) \vee MCE^d B_{i,j}^{\Sigma}(\varphi)) \wedge K_j\neg\varphi) \Rightarrow (MCE^d K_{i,j}^{\Sigma}(\varphi) \vee MCE^d B_{i,j}^{\Sigma}(\varphi))$ [Left elim. (∧)]

(g) ...

(h) $\vdash ((MCE^d K_{i,j}^{\Sigma}(\varphi) \vee MCE^d B_{i,j}^{\Sigma}(\varphi))) \Rightarrow ((E_j^d\varphi \wedge \neg K_j E_i^d E_j^d\varphi) \vee (E_j^d\varphi \wedge \neg B_j E_i^d E_j^d\varphi))$ [CBR[13] Th. (T)]

(i) ...

(j) $\vdash ((E_j^d\varphi \wedge \neg K_j E_i^d E_j^d\varphi) \vee (E_j^d\varphi \wedge \neg B_j E_i^d E_j^d\varphi)) \Rightarrow E_j^d\varphi$ [Elim. (∧) (∨) ]

(k) $\vdash E_j^d\varphi \Rightarrow \varphi$ [Ax. (T)]

(l) $\vdash ((MCE^d K_{i,j}^{\Sigma}(\varphi) \vee MCE^d B_{i,j}^{\Sigma}(\varphi)) \wedge K_j\neg\varphi) \Rightarrow ((E_j^d\varphi \wedge \neg K_j E_i^d E_j^d\varphi) \vee (E_j^d\varphi \wedge \neg B_j E_i^d E_j^d\varphi)) \Rightarrow E_j^d\varphi \Rightarrow \varphi$ [Aug. (k)(×2)]

(m) $\vdash (((MCE^d K_{i,j}^{\Sigma}(\varphi) \vee MCE^d B_{i,j}^{\Sigma}(\varphi)) \wedge K_j\neg\varphi) \Rightarrow ((E_j^d\varphi \wedge \neg K_j E_i^d E_j^d\varphi) \vee (E_j^d\varphi \wedge \neg B_j E_i^d E_j^d\varphi)) \Rightarrow E_j^d\varphi \Rightarrow \varphi) \Rightarrow (((MCE^d K_{i,j}^{\Sigma}(\varphi) \vee MCE^d B_{i,j}^{\Sigma}(\varphi)) \wedge K_j\neg\varphi) \Rightarrow ((E_j^d\varphi \wedge \neg K_j E_i^d E_j^d\varphi) \vee (E_j^d\varphi \wedge \neg B_j E_i^d E_j^d\varphi))) \Rightarrow (((MCE^d K_{i,j}^{\Sigma}(\varphi) \vee MCE^d B_{i,j}^{\Sigma}(\varphi)) \wedge K_j\neg\varphi) \Rightarrow (E_j^d\varphi \Rightarrow \varphi))$ [Syll. (l)]

(n) $\vdash ((MCE^d K_{i,j}^{\Sigma}(\varphi) \vee MCE^d B_{i,j}^{\Sigma}(\varphi)) \wedge K_j\neg\varphi) \Rightarrow (E_j^d\varphi \Rightarrow \varphi)$ [MP

[13] We apply here the Case-Based Reasoning i.e. the elimination of the disjunction.

(l),(f),(m)]

(o) $\vdash (((MCE^d K^\Sigma_{i,j}(\varphi) \vee MCE^d B^\Sigma_{i,j}(\varphi)) \wedge K_j\neg\varphi) \Rightarrow (E^d_j\varphi \Rightarrow \varphi)) \Rightarrow ((((MCE^d K^\Sigma_{i,j}(\varphi) \vee MCE^d B^\Sigma_{i,j}(\varphi)) \wedge K_j\neg\varphi) \Rightarrow E^d_j\varphi) \Rightarrow (((MCE^d K^\Sigma_{i,j}(\varphi) \vee MCE^d B^\Sigma_{i,j}(\varphi)) \wedge K_j\neg\varphi) \Rightarrow \varphi))$ [Syll. (n)]

(p) $\vdash (((MCE^d K^\Sigma_{i,j}(\varphi) \vee MCE^d B^\Sigma_{i,j}(\varphi)) \wedge K_j\neg\varphi) \Rightarrow \varphi)$ [MP (l),(e),(m)]

(q) $\vdash (((MCE^d K^\Sigma_{i,j}(\varphi) \vee MCE^d B^\Sigma_{i,j}(\varphi)) \wedge K_j\neg\varphi) \Rightarrow \neg\varphi) \Rightarrow ((((MCE^d K^\Sigma_{i,j}(\varphi) \vee MCE^d B^\Sigma_{i,j}(\varphi)) \wedge K_j\neg\varphi) \Rightarrow \varphi)) \Rightarrow (\neg(MCE^d K^\Sigma_{i,j}(\varphi) \vee MCE^d B^\Sigma_{i,j}(\varphi)) \vee \neg K_j\neg\varphi))$ [*Reductio ad absurdum*]

(r) $\vdash (\neg(MCE^d K^\Sigma_{i,j}(\varphi) \vee MCE^d B^\Sigma_{i,j}(\varphi)) \vee \neg K_j\neg\varphi)$ [MP (e),(p),(q)]

(s) $\vdash (\neg(MCE^d K^\Sigma_{i,j}(\varphi) \vee MCE^d B^\Sigma_{i,j}(\varphi)) \vee \neg K_j\neg\varphi) \Leftrightarrow ((MCE^d K^\Sigma_{i,j}(\varphi) \vee MCE^d B^\Sigma_{i,j}(\varphi)) \wedge K_j\neg\varphi \Rightarrow \bot)$ [Th. ($\vdash \varphi \Leftrightarrow (\neg\varphi \Rightarrow \bot)$)]

(t) $\vdash (MCE^d K^\Sigma_{i,j}(\varphi) \vee MCE^d B^\Sigma_{i,j}(\varphi)) \wedge K_j\neg\varphi \Rightarrow \bot$ [MP (r),(s)]

(2)

(a) $\vdash (MCE^d K^\Sigma_{i,j}(\varphi) \vee MCE^d B^\Sigma_{i,j}(\varphi)) \wedge B_j\neg\varphi \Rightarrow B_j\neg\varphi$ [Right elim. ($\wedge$)]

(b) ...

(c) $\vdash ((E^d_j\varphi \wedge \neg K_j E^d_i E^d_j\varphi) \vee (E^d_j\varphi \wedge \neg B_j E^d_i E^d_j\varphi)) \wedge B_j\neg\varphi \Rightarrow E^d_j\varphi$ [Left elim. ($\wedge$) ]

(d) ...

(e) $\vdash (E^d_j\varphi \Rightarrow K_j E^d_j\varphi)$ [Ax. ($5_{K_j,E^d_j}$)]

(f) $\vdash ((E^d_j\varphi \wedge \neg K_j E^d_i E^d_j\varphi) \vee (E^d_j\varphi \wedge \neg B_j E^d_i E^d_j\varphi)) \Rightarrow (E^d_j\varphi \Rightarrow K_j E^d_j\varphi)$ [Aug. (e)]

(g) $\vdash K_j E^d_j\varphi \Rightarrow B_j E^d_j\varphi$ [Ax. ($K_jB_j$)]

(h) ...

(i) $\vdash B_j(E^d_j\varphi \Rightarrow \varphi)$ [Nec ($B_j$) on ($T_{E^d_j}$)]

(j) $\vdash B_j(E^d_j\varphi \Rightarrow \varphi) \Rightarrow (B_j E^d_j\varphi \Rightarrow B_j\varphi)$ [Ax. (K)]

(k) $\vdash B_j E^d_j\varphi \Rightarrow B_j\varphi$ [MP (i), (j)]

(l) ...

(m) $\vdash ((E^d_j\varphi \wedge \neg K_j E^d_i E^d_j\varphi) \vee (E^d_j\varphi \wedge \neg B_j E^d_i E^d_j\varphi)) \Rightarrow E^d_j\varphi \Rightarrow K_j E^d_j\varphi \Rightarrow B_j E^d_j\varphi \Rightarrow B_j\varphi$ [Aug. (k)]

(n) $\vdash (MCE^d K^\Sigma_{i,j}(\varphi) \vee MCE^d B^\Sigma_{i,j}(\varphi)) \wedge B_j\neg\varphi \Rightarrow B_j\varphi$ [MP (c) ... (m) (Syll. (m))]

(o) ...

(p) $\vdash B_j\varphi \Rightarrow (B_j\neg\varphi \Rightarrow (B_j\varphi \wedge B_j\neg\varphi))$ [Intro. ($\wedge$)]

(q) $\vdash (MCE^d K^\Sigma_{i,j}(\varphi) \vee MCE^d B^\Sigma_{i,j}(\varphi)) \wedge B_j\neg\varphi \Rightarrow (B_j\varphi \Rightarrow (B_j\neg\varphi \Rightarrow (B_j\varphi \wedge B_j\neg\varphi)))$ [Aug. (p)]

(r) $\vdash (MCE^d K^\Sigma_{i,j}(\varphi) \vee MCE^d B^\Sigma_{i,j}(\varphi)) \wedge B_j\neg\varphi \Rightarrow (B_j\varphi \wedge B_j\neg\varphi)$ [MP (n) (a) (Syll. (q)) ]

(s) $\vdash (MCE^d K^\Sigma_{i,j}(\varphi) \vee MCE^d B^\Sigma_{i,j}(\varphi)) \wedge B_j\neg\varphi \Rightarrow \bot$ [RE (r) ($\vdash B_j\varphi \wedge B_j\neg\varphi \Leftrightarrow \bot$)]

(3)

(a) $\vdash (E^d_i(E^d_j\varphi \wedge \neg K_j E^d_i E^d_j\varphi) \vee E^d_i(E^d_j\varphi \wedge \neg B_j E^d_i E^d_j\varphi)) \Rightarrow ((E^d_j\varphi \wedge \neg K_j E^d_i E^d_j\varphi) \vee (E^d_j\varphi \wedge \neg B_j E^d_i E^d_j\varphi))$ [Th. ($T_{E^d_i}$)]

(b) $\vdash (E^d_j\varphi \wedge \neg K_j E^d_i E^d_j\varphi) \vee (E^d_j\varphi \wedge \neg B_j E^d_i E^d_j\varphi)) \Rightarrow E^d_j\varphi$ [MP (Elim. ($\wedge$)) (Elim. ($\vee$))]

(c) $\vdash E^d_j\varphi \Rightarrow K_j E^d_j\varphi \Rightarrow K_j\varphi$ [Aug. (MP (Th. ($T_{E^d_i}$))(Nec $K_i$ + Ax. ($K_{K_i}$)))]

(d) ...

(e) $\vdash (MCE^d K^\Sigma_{i,j}(\varphi) \vee MCE^d B^\Sigma_{i,j}(\varphi)) \Rightarrow K_j\varphi$ [MP (...) Syll. (c)]

□

**Theorem 5.6.**

*(1)* $\vdash MCEK^{\Sigma}_{i,j}(E_j\varphi) \Leftrightarrow MCEK^{\Sigma}_{i,j}(\varphi)$
*(2)* $\vdash MCEK^{\Sigma}_{i,j}(\neg E_j\varphi) \Leftrightarrow MDEK^{\Sigma}_{i,j}(\varphi)$
*(3)* $\vdash MCEB^{\Sigma}_{i,j}(E_j\varphi) \Leftrightarrow MCEB^{\Sigma}_{i,j}(\varphi)$
*(4)* $\vdash MCEB^{\Sigma}_{i,j}(\neg E_j\varphi) \Leftrightarrow MDEB^{\Sigma}_{i,j}(\varphi)$
*(5)* $\vdash MDEK^{\Sigma}_{i,j}(E_j\varphi) \Leftrightarrow MDEK^{\Sigma}_{i,j}(\varphi)$
*(6)* $\vdash MDEK^{\Sigma}_{i,j}(\neg E_j\varphi) \Leftrightarrow MCEK^{\Sigma}_{i,j}(\varphi)$
*(7)* $\vdash MDEB^{\Sigma}_{i,j}(E_j\varphi) \Leftrightarrow MDEB^{\Sigma}_{i,j}(\varphi)$
*(8)* $\vdash MDEB^{\Sigma}_{i,j}(\neg E_j\varphi) \Leftrightarrow MCEB^{\Sigma}_{i,j}(\varphi)$

***Proof.*** All those theorems are obvious to show by using $(RE)$ and considering the following theorems $\vdash E_iE_i\varphi \Leftrightarrow E_i\varphi$ for (1) and (3), $\vdash E_j\neg E_j\varphi \Leftrightarrow \neg E_j\varphi$ for (2) and (4), $\vdash \neg E_iE_i\varphi \Leftrightarrow \neg E_i\varphi$ for (5) and (7), $\vdash \neg E_j\neg E_j\varphi \Leftrightarrow E_j\varphi$ for (6) and (8). □

**Theorem 5.7.**

*(1)* $\not\models \neg MCEK^{\Sigma}_{i,j}(\varphi) \wedge MCE^dK^{\Sigma}_{i,j}(\varphi) \Rightarrow \bot$
*(2)* $\not\models MCEK^{\Sigma}_{i,j}(\varphi) \wedge \neg MCE^dK^{\Sigma}_{i,j}(\varphi) \Rightarrow \bot$
*(3)* $\not\models MCEK^{\Sigma}_{i,j}(\varphi) \wedge MCE^dK^{\Sigma}_{i,j}(\varphi) \Rightarrow \bot$

***Proof.*** For both theorems, we will use the same model base, modulo a set of sets of possible worlds $X \subseteq 2^{\mathcal{W}}$. Let $\mathcal{M} = \langle \mathcal{W}, \{\mathcal{B}_i\}, \{\mathcal{K}_i\}, \{\mathcal{E}_i\}, \{\mathcal{E}^d_i\}, V\rangle$ be a KBE model, and $\Sigma$ be any finite and closed set of formulas that contains $\{\top, \bot\} \cup \{p\}$ and such that:

- $\mathcal{W} = \{w, x, y, z\}$
- $\mathcal{K}_i = \{(w,w),(x,x),(y,y),(z,z)\} = \mathcal{B}_i = \mathcal{E}_i = \mathcal{K}_j = \mathcal{B}_j = \mathcal{E}_j$
- $\mathcal{E}^d_i = \{(w,X),(x,\{\{x\}\}),(y,\{\{w,y\}\}),(z,\{\{z\}\})\}$
- $\mathcal{E}^d_j = \{(w,\{\{w,y,z\}\}),(x,\{\{x\}\}),(y,\{\{w,y,z\}\}),(z,\{\{z\}\})\}$
- $V(p) = \{w, y, z\}$

(0) if $X = \{\{w\}\}$, $X = \{\{w,z\}\}$ or $X = \{\{w\},\{w,z\}\}$, then we have:
  - $||E_jp|| = \{w,y,z\}$ and $||E^d_jp|| = \{w,y\}$
  - $||E^d_iE_jp|| = \{y\}$ and $||\neg E^d_iE_jp|| = \{w,x,z\}$
  - $||E^d_iE^d_jp|| = \{y\}$ and $||\neg E^d_iE^d_jp|| = \{w,x,z\}$
  - $||K_j\neg E^d_iE_jp|| = \{w,x,z\} \subseteq ||\neg K_jE^d_iE_jp|| = \{w,x,z\}$
  - $||K_j\neg E^d_iE^d_jp|| = \{w,x,z\} \subseteq ||\neg K_jE^d_iE^d_jp|| = \{w,x,z\}$
  - $||E^d_jp \wedge \neg K_jE^d_iE^d_jp|| = ||E^d_jp|| \cap ||\neg K_jE^d_iE^d_jp|| = \{w\}$,
  - $||E_jp \wedge \neg K_jE^d_iE_jp|| = ||E_jp|| \cap ||\neg K_jE^d_iE_jp|| = \{w,z\}$,

(1) if $X = \{\{w\}\}$, then $\mathcal{M}, w \models \neg MCEK^{\Sigma}_{i,j}(\varphi) \wedge MCE^dK^{\Sigma}_{i,j}(\varphi)$
since $\mathcal{M}, w \models \neg E^d_i(E_jp \wedge \neg K_jE^d_iE_jp) \wedge E^d_i(E^d_jp \wedge \neg K_jE^d_iE^d_jp)$

(2) if $X = \{\{w,z\}\}$, then $\mathcal{M}, w \models MCEK^{\Sigma}_{i,j}(\varphi) \wedge \neg MCE^dK^{\Sigma}_{i,j}(\varphi)$
since $\mathcal{M}, w \models E^d_i(E_jp \wedge \neg K_jE^d_iE_jp) \wedge \neg E^d_i(E^d_jp \wedge \neg K_jE^d_iE^d_jp)$

(3) if $X = \{\{w\},\{w,z\}\}$, then $\mathcal{M}, w \models MCEK^{\Sigma}_{i,j}(\varphi) \wedge MCE^dK^{\Sigma}_{i,j}(\varphi)$
since $\mathcal{M}, w \models E^d_i(E_jp \wedge \neg K_jE^d_iE_jp) \wedge E^d_i(E^d_jp \wedge \neg K_jE^d_iE^d_jp)$

Let us notice that it is possible to describe with this model the fact that the agent $i$

does not softly or strongly manipulate the agent $j$ if $X = \{\{w, x\}\}$. We would have $\mathcal{M}, w \models \neg MCEK^{\Sigma}_{i,j}(\varphi) \wedge \neg MCE^d K^{\Sigma}_{i,j}(\varphi)$ since $\mathcal{M}, w \models \neg E^d_i(E_j p \wedge \neg K_j E^d_i E_j p) \wedge \neg E^d_i(E^d_j p \wedge \neg K_j E^d_i E^d_j p)$. □

**Theorem 5.8.**

*(1)* $\vdash E^d_i E_i\varphi \wedge E^d_i \neg K_i E^d_i E_i\varphi \Rightarrow \bot$
*(2)* $\not\models MCEK^{\Sigma}_{i,i}(\varphi) \Rightarrow \bot$ *and* $\not\models MCEB^{\Sigma}_{i,i}(\varphi) \Rightarrow \bot$

***Proof.***

(1)
- (a) $\vdash E^d_i E_i\varphi \wedge E^d_i \neg K_i E^d_i E_i\varphi \Rightarrow K_i E^d_i E_i\varphi$ [Left Elim. ($\wedge$) + Ax. ($5_{K_i, E^d_i}$)]
- (b) $\vdash E^d_i E_i\varphi \wedge E^d_i \neg K_i E^d_i E_i\varphi \Rightarrow \neg K_i E^d_i E_i\varphi$ [Right Elim. ($\wedge$) + Th. ($T_{E^d_i}$)]
- (c) ...
- (d) $\vdash E^d_i E_i\varphi \wedge E^d_i \neg K_i E^d_i E_i\varphi \Rightarrow (K_i E^d_i E_i\varphi \wedge \neg K_i E^d_i E_i\varphi \Leftrightarrow \bot)$ [Conclusion]

(2) Let $M = \langle \mathcal{W}, \{\mathcal{B}_i\}, \{\mathcal{K}_i\}, \{\mathcal{E}_i\}, \{\mathcal{E}^d_i\}, V\rangle$ be a KBE model, and $\Sigma$ be any finite and closed set of formulas that contains $\{\top, \bot\} \cup \{p\}$ and such that:
- $\mathcal{W} = \{w, x, y\}$
- $\mathcal{K}_i = \{(w, w), (x, x), (y, y)\} = B_i = \mathcal{E}_i$
- $\mathcal{E}^d_i = \{(w, \{\{w\}\}), (x, \{\{x\}\}), (y, \{\{w, y\}\})\}$
- $V(p) = \{w, y\}$

We have:
- $||E_i p|| = \{w, y\}$
- $||\neg K_i E^d_i E_i p|| = ||\neg E^d_i E_i p|| = \{z \in \mathcal{W} : ||E_i p|| \notin \mathcal{E}^d_i(z)\} = \{w, x\}$
- $||E_i p \wedge \neg K_i E^d_i E_i p|| = ||E_i p|| \cap ||\neg K_i E^d_i E_i p|| = \{w\}$

Thus, since $||E_i p \wedge \neg K_i E^d_i E_i p|| \in \mathcal{E}^d_i(w)$, we have proved that there exists a KBE-model $\mathcal{M}$ and $w \in \mathcal{W}$ s.t. $\mathcal{M}, w \models E^d_i(E_i p \wedge \neg K_i E^d_i E_i p)$. Furthermore, since $\models E^d_i(E_i p \wedge \neg K_i E^d_i E_i p) \Leftrightarrow E^d_i(E_i p \wedge \neg K_i E^d_i E_i p \wedge \top)$ and by introduction of $\vee$ i.e. $\models \varphi \Rightarrow (\varphi \vee \psi)$, we prove the manipulation, i.e. $\mathcal{M}, w \models MCEK^{\Sigma}_{i,i}(\varphi)$. We notice that with this KBE model, we prove also $\not\models MCEB^{\Sigma}_{i,i}(\varphi) \Rightarrow \bot$.

□

**Theorem 5.9.**

*(1)* $\vdash CKCoe^{\Sigma}_{i,j}(\varphi) \Leftrightarrow \bigvee_{\psi \in \Sigma \,\oslash\, \{K_j E^d_i E^d_j \varphi\}} E^d_i(E^d_j \varphi \wedge \psi)$
*(2)* $\vdash CBCoe^{\Sigma}_{i,j}(\varphi) \Leftrightarrow \bigvee_{\psi \in \Sigma \,\oslash\, \{B_j E^d_i E^d_j \varphi\}} E^d_i(E^d_j \varphi \wedge \psi)$

***Proof.*** By rewriting $CKCoe^{\Sigma}_{i,j}(\varphi)$ and $\Sigma$, we have for (1):

$$\vdash CKCoe^{\Sigma}_{i,j}(\varphi) \Leftrightarrow \bigvee_{\psi \in \Sigma} E^d_i(E^d_j \varphi \wedge K_j E^d_i E^d_j \varphi \wedge \psi) \Leftrightarrow \bigvee_{\psi \in \Sigma \,\oslash\, \{K_j E^d_i E^d_j \varphi\}} E^d_i(E^d_j \varphi \wedge \psi)$$

The same methodology holds for (2). □

**Theorem 5.10.**

*(1)* $\vdash per^{\Sigma}_{i,j}(\neg E^d_j \varphi) \Leftrightarrow \bigvee_{\psi \in \Sigma} E^d_i(\neg E^d_j \varphi \wedge \psi)$

*(2)* $\vdash per^{\Sigma}_{i,j}(E^d_j\varphi) \Leftrightarrow \bigvee\limits_{\psi\in\Sigma} E^d_i(E^d_j\varphi \wedge \psi)$

***Proof.*** Let us recall Theorem 4.6 says $\vdash B_j\neg E^d_j\varphi \Leftrightarrow \neg E^d_j\varphi$ and $\vdash B_jE^d_j\varphi \Leftrightarrow E^d_j\varphi$. Thus, we have:

(1) $\vdash per^{\Sigma}_{i,j}(\neg E^d_j\varphi) \Leftrightarrow \bigvee_{\psi\in\Sigma} E^d_i(B_j\neg E^d_j\varphi \wedge \psi) \Leftrightarrow \bigvee\limits_{\psi\in\Sigma} E^d_i(\neg E^d_j\varphi \wedge \psi)$

(2) $\vdash per^{\Sigma}_{i,j}(E^d_j\varphi) \Leftrightarrow \bigvee_{\psi\in\Sigma} E^d_i(B_jE^d_j\varphi \wedge \psi) \Leftrightarrow \bigvee\limits_{\psi\in\Sigma} E^d_i(E^d_j\varphi \wedge \psi)$

□

**Theorem 5.11.**

*(1)* $\vdash CKCoe^{\Sigma}_{i,j}(\varphi) \Leftrightarrow per^{\Sigma \circledcirc \{K_jE^d_iE^d_j\varphi\}}_{i,j}(E^d_j\varphi)$

*(2)* $\vdash CBCoe^{\Sigma}_{i,j}(\varphi) \Leftrightarrow per^{\Sigma \circledcirc \{B_jE^d_iE^d_j\varphi\}}_{i,j}(E^d_j\varphi)$

*(3)* $\vdash DKCoe^{\Sigma}_{i,j}(\varphi) \Leftrightarrow per^{\Sigma \circledcirc \{K_jE^d_i\neg E^d_j\varphi\}}_{i,j}(\neg E^d_j\varphi)$

*(4)* $\vdash DBCoe^{\Sigma}_{i,j}(\varphi) \Leftrightarrow per^{\Sigma \circledcirc \{B_jE^d_i\neg E^d_j\varphi\}}_{i,j}(\neg E^d_j\varphi)$

***Proof.*** By rewriting $CKCoe^{\Sigma}_{i,j}(\varphi)$ and $\Sigma$, we have for (1):

$$\vdash CKCoe^{\Sigma}_{i,j}(\varphi) \Leftrightarrow \bigvee_{\psi\in\Sigma} E^d_i(E^d_j\varphi \wedge K_jE^d_iE^d_j\varphi \wedge \psi) \Leftrightarrow$$

$$\bigvee_{\psi\in\Sigma \circledcirc \{K_jE^d_iE^d_j\varphi\}} E^d_i(E^d_j\varphi \wedge \psi) \Leftrightarrow per^{\Sigma \circledcirc \{K_jE^d_iE^d_j\varphi\}}_{i,j}(E^d_j\varphi)$$

The same methodology holds for (2), (3) and (4). □

**Theorem 5.12.**

*(1)* $\vdash per^{\Sigma \circledcirc \{\neg K_jE^d_iE^d_j\varphi\}}_{i,j}(E^d_j\varphi) \Leftrightarrow MCE^dK^{\Sigma}_{i,j}(\varphi)$

*(2)* $\vdash per^{\Sigma \circledcirc \{\neg B_jE^d_iE^d_j\varphi\}}_{i,j}(E^d_j\varphi) \Leftrightarrow MCE^dB^{\Sigma}_{i,j}(\varphi)$

*(3)* $\vdash per^{\Sigma \circledcirc \{\neg K_jE^d_i\neg E^d_j\varphi\}}_{i,j}(\neg E^d_j\varphi) \Leftrightarrow MDE^dK^{\Sigma}_{i,j}(\varphi)$

*(4)* $\vdash per^{\Sigma \circledcirc \{\neg B_jE^d_i\neg E^d_j\varphi\}}_{i,j}(\neg E^d_j\varphi) \Leftrightarrow MDE^dB^{\Sigma}_{i,j}(\varphi)$

***Proof.*** By rewriting $MCE^dK^{\Sigma}_{i,j}(\varphi)$ and $\Sigma$, we have for (1):

$$\vdash per^{\Sigma \circledcirc \{\neg K_jE^d_iE^d_j\varphi\}}_{i,j}(E^d_j\varphi) \Leftrightarrow \bigvee_{\psi\in\Sigma \circledcirc \{\neg K_jE^d_iE^d_j\varphi\}} E^d_i(B_jE^d_j\varphi \wedge \psi) \Leftrightarrow$$

$$\bigvee_{\psi\in\Sigma \circledcirc \{\neg K_jE^d_iE^d_j\varphi\}} E^d_i(E^d_j\varphi \wedge \psi) \Leftrightarrow \bigvee_{\psi\in\Sigma} E^d_i(E^d_j\varphi \wedge \neg K_jE^d_iE^d_j\varphi \wedge \psi) \Leftrightarrow MCE^dK^{\Sigma}_{i,j}(\varphi)$$

The same methodology holds for (2), (3) and (4). □

**Theorem 5.13.**

*(1)* $\vdash KCrLie^{\Sigma}_{i,j}(E^d_j\varphi) \Rightarrow \bot$
*(2)* $\vdash BCrLie^{\Sigma}_{i,j}(E^d_j\varphi) \Rightarrow \bot$
*(3)* $\vdash KSoCo^{\Sigma}_{i,j}(E^d_j\varphi) \Leftrightarrow MCE^dK^{\Sigma}_{i,j}(\varphi)$
*(4)* $\vdash BSoCo^{\Sigma}_{i,j}(E^d_j\varphi) \Leftrightarrow MCE^dB^{\Sigma}_{i,j}(\varphi)$
*(5)* $\vdash KCrLie^{\Sigma}_{i,j}(\varphi) \Rightarrow Per^{\Sigma \oslash \{\neg K_jB_i\neg\varphi\}}_{i,j}(\varphi)$
*(6)* $\vdash BCrLie^{\Sigma}_{i,j}(\varphi) \Rightarrow Per^{\Sigma \oslash \{\neg B_jB_i\neg\varphi\}}_{i,j}(\varphi)$

***Proof.***

(1) (a) $\vdash B_i\neg E^d_j\varphi \wedge E^d_i(B_jE^d_j\varphi \wedge \neg K_jB_i\neg\varphi) \Rightarrow B_i\neg E^d_j\varphi \wedge E^d_i(E^d_j\varphi \wedge \neg K_jB_i\neg\varphi)$ [RE on Th. ($\vdash B_jE^d_j\varphi \Leftrightarrow E^d_j\varphi$)]
(b) $B_i\neg E^d_j\varphi \wedge E^d_i(E^d_j\varphi \wedge \neg K_jB_i\neg\varphi) \Leftrightarrow B_i\neg E^d_j\varphi \wedge B_iE^d_i(E^d_j\varphi \wedge \neg K_jB_i\neg\varphi)$ [RE on Th. ($\vdash B_iE^d_i\varphi \Leftrightarrow E^d_i\varphi$)]
(c) $\vdash B_i\neg E^d_j\varphi \wedge B_iE^d_i(E^d_j\varphi \wedge \neg K_jB_i\neg\varphi) \Rightarrow B_i\neg E^d_j\varphi \wedge B_iE^d_j\varphi$ [Right elim. ($\wedge$) + Th. ($T_{E^d_i}$) + Nec. $B_i$ + MP on (K)]
(d) $\vdash B_i\neg E^d_j\varphi \wedge B_iE^d_j\varphi \Leftrightarrow B_i\bot$
(e) $B_i\neg E^d_j\varphi \wedge E^d_i(B_jE^d_j\varphi \wedge \neg K_jB_i\neg\varphi) \Rightarrow \bot$

(2) The same methodology as (1) holds for this property.

(3) By rewriting $MCE^dK^{\Sigma}_{i,j}(\varphi)$ and $\Sigma$, we have:

$$\vdash KSoCo^{\Sigma}_{i,j}(E^d_j\varphi) \Leftrightarrow \bigvee_{\psi\in\Sigma} E^d_i(B_jE^d_j\varphi \wedge \neg K_jE^d_iB_jE^d_j\varphi \wedge \psi) \Leftrightarrow$$

$$\bigvee_{\psi\in\Sigma} E^d_i(E^d_j\varphi \wedge \neg K_jE^d_iE^d_j\varphi \wedge \psi) \Leftrightarrow MCE^dK^{\Sigma}_{i,j}(\varphi)$$

(4) The same methodology as (3) holds for this property.

(5) (a) $\vdash KCrLie^{\Sigma}_{i,j}(\varphi) \Rightarrow (\bigvee_{\psi\in\Sigma} E^d_i(B_j\varphi \wedge \neg K_jB_i\neg\varphi \wedge \psi))$ [Right Elim. ($\wedge$)]
(b) $\vdash (\bigvee_{\psi\in\Sigma} E^d_i(B_j\varphi \wedge \neg K_jB_i\neg\varphi \wedge \psi)) \Leftrightarrow Per^{\Sigma \oslash \{\neg K_jB_i\neg\varphi\}}_{i,j}(\varphi)$ [Def. $(Per^{\Sigma \oslash \{\neg K_jB_i\neg\varphi\}}_{i,j}(\varphi))$]
(c) $\vdash KCrLie^{\Sigma}_{i,j}(\varphi) \Rightarrow Per^{\Sigma \oslash \{\neg K_jB_i\neg\varphi\}}_{i,j}(\varphi)$ [RE (b) in (a)]

(6) The same methodology as (5) holds for (6).

□