Deception in Epistemic Causal Logic

Chiaki Sakama

Wakayama University
930 Sakaedani, Wakayama 640-8510 Japan
sakama@wakayama-u.ac.jp

Abstract. *Deception* is an act whereby one person causes another person to have a false belief. This paper formulates deception using causal relations between a speaker's utterance and a hearer's belief states in *epistemic causal logic*. Four different types of deception are considered: *deception by lying*, *deception by bluffing*, *deception by truthful telling*, and *deception by omission*, depending on whether a speaker believes what he/she says or not, and whether a speaker makes an utterance or not. Next several situations are considered where an act of deceiving happens. *Intentional deception* is accompanied by a speaker's intent to deceive. *Indirect deception* happens when false information is carried over from person to person. *Self-deception* is an act of deceiving the self. The current study formally characterizes various aspects of deception that have been informally argued in philosophical literature.

Keywords: deception · epistemic causal logic · lying

1 Introduction

Deception is a part of human nature and is a topic of interest in philosophy and elsewhere. Most philosophers agree that an act of deceiving implies a success of the act, while they disagree as to whether deceiving must be intentional or not [3, 18]. Deceiving is different from *lying*, in fact, there is deception without lying [1, 34]. There is no consensus as to stating conditions for describing someone as *self-deceived* [8]. In this way, deception has been subject to extensive studies on the one hand, but deception argued in philosophical literature is mostly conceptual, on the other hand.

Deception is also a topic of interest in computer science and AI. Recent development of machine learning involves various forms of deceptive activities on social media [6]. Historically, the question "*Can computers deceive humans?*" has been argued since Turing's imitation game [31]. Castelfranchi [4] argued the possibility of artificial agents that deceive humans in several ways, not only for malicious reasons but also for goodwill and in our interest. For instance, an intelligent personal assistant might deceive us to make a right decision. One could imagine a medical counseling system which does not always inform patients of the true state of affairs. Clark [7] develops a *lying machine* and provides empirical evidence that the machine reliably deceives ordinary humans. Isaac and Bridewell [16] argue that robots must possess a theory of mind in order to respond effectively to deceptive communication. Some studies show that robots can gain advantage over adversaries by deceptive behaviors [28, 35]. Deception is also of particular interest in a game-theoretical perspective [11, 15], and is adopted as a strategy

of intelligent agents in multiagent systems [27, 29, 36]. In spite of the broad interest in this topic, however, relatively little study has been devoted to developing a formal theory of deception. A formal account of deception helps us to better understand what is deception, and to design artificial agents that can detect deceptive acts in virtual societies. Deception is a perlocutionary act that produces an effect in the belief state of an addressee by communication. Formulation of deception then needs a logic that can express belief of agents, communication between agents and effects of communication. In this respect, a logical language that has causal relations as well as epistemic modality is useful for the purpose.

In this paper, we use the *causal logic* of [14] to represent causal relations between a deceiver's speech act and its effect on belief states of an addressee. We define formulas for representing utterance and belief of agents, and introduce a set of axioms in *epistemic causal logic*. Using the logic, we formulate four different types of deception, *deception by lying*, *deception by bluffing*, *deception by truthful telling*, and *deception by omission*, and distinguish them from *attempted deception* that may fail to deceive. We next discuss various aspects of deception such as *intended deception*, *indirect deception* and *self-deception*. We address formal properties of those different sorts of deception.

The rest of this paper is organized as follows. Section 2 introduces the logical framework used in this paper. Section 3 formulates different types of deception and investigates formal properties. Section 4 presents various aspects of deception. Section 5 addresses comparison with related studies. Section 6 concludes with remarks.

2 Epistemic Causal Logic

We first review the *causal logic* of [14] that is used in this paper. Let $\mathscr{L}$ be a language of propositional logic. An *atom* is a propositional variable p in $\mathscr{L}$. A *literal* is an atom p or its negation $\neg p$. *Formulas* (or *sentences*) in $\mathscr{L}$ are defined as follows: (i) an atom p is a formula. (ii) If φ and ψ are formulas, then $\neg\varphi$, $\varphi \wedge \psi$, $\varphi \vee \psi$, $\varphi \supset \psi$, and $\varphi \equiv \psi$ are all formulas. In particular, $\top$ and $\bot$ represent valid and contradictory formulas, respectively. We often use parentheses "()" in a formula as usual. Throughout the paper, Greek letters λ, φ, ψ represent formulas.

An *interpretation* I is a complete and consistent (finite) set of literals.[1] A literal ℓ is *true* in an interpretation I iff $\ell \in I$. The truth value of a formula φ in I is defined based on the usual truth tables of propositional connectives. An interpretation I *satisfies* a formula φ (written $I \models \varphi$) iff φ is true in I. Given formulas φ and ψ,

$$\varphi \Rightarrow \psi \tag{1}$$

is called a *causal rule*. φ is called a *cause* and ψ is called an *effect*. The rule (1) means that "ψ is caused if φ is true." In particular, the rule $(\top \Rightarrow \psi)$ is a *fact* representing that ψ is true, which is abbreviated as ψ.

A *(causal) theory* is a finite set of causal rules. A theory T is identified with the conjunction of all rules in T. Given a theory T and an interpretation I, define

$$T^I = \{\, \psi \mid (\varphi \Rightarrow \psi) \in T \text{ for some } \varphi \text{ and } I \models \varphi \,\}.$$

[1] That is, $\ell \in I$ iff $\neg\ell \notin I$ for any literal ℓ appearing in a theory.

Then I is a *model* of T if I is the unique model of T^I. If every model of T satisfies a formula F, then we say that T *entails* F (written $T \models F$). A theory T is *consistent* if it has a model; otherwise, T is *inconsistent* (written $T \models \bot$).

Example 1. Suppose the theory

$$T = \{\, p \Rightarrow q, \quad p \Rightarrow p, \quad \neg p \Rightarrow \neg p \,\}.$$

In T there is no cause for $\neg q$, then $\neg q$ is false or q is true. Since every true formula is caused, it must be the case that q is caused. This leads to the conclusion that p is true. As a result, T has the single model $\{p,q\}$. In fact, by putting $I = \{p,q\}$, it becomes $T^I = \{p,q\}$ and I is the unique model of T^I.

Note that $\Rightarrow$ is not identical to material implication in classical logic. In Example 1 if $\Rightarrow$ is replaced by material implication as: $T' = \{\, p \supset q, \quad p \supset p, \quad \neg p \supset \neg p \,\}$ then $p \supset p$ and $\neg p \supset \neg p$ are tautologies and can be removed. As a result, $\{\neg p, q\}$ and $\{\neg p, \neg q\}$ are also models of T', but these are not models of T. In fact, $p \Rightarrow p$ and $\neg p \Rightarrow \neg p$ are not tautologies in a causal theory. Formally, if a causal theory T contains $\varphi \Rightarrow \psi$ then T entails $\varphi \supset \psi$ but not vice versa.

Actions and their effects are represented by a causal theory. In this paper we consider an action as an utterance by an agent. Suppose that an agent a utters a sentence φ to an agent b at time t. The situation is represented by the atom $U^t_{ab}\varphi$. A belief state of an agent is represented as a fluent. When an agent a believes (resp. disbelieves) a sentence φ at time t, it is represented by the literal $B^t_a\varphi$ (resp. $\neg B^t_a\varphi$). Beliefs are possibly nested, for instance, the atom $B^{t+1}_b B^t_a \varphi$ represents that an agent b believes at $t+1$ that an agent a believes φ at t. By contrast, factual sentences are considered persistent and written as φ without time. Note that we handle $B^t_a\varphi$ or $B^{t+1}_b B^t_a\varphi$ as an *atom* in a theory, so that B^t_a is not an operator in modal epistemic logic.[2] This enables us to define the semantics without introducing a Kripke structure and to introduce necessary axioms depending on the objective. For the current use, the following axioms of utterance and belief are introduced. Let φ and ψ be sentences. Given agents a, b and time t,

(axioms of utterance): $U^t_{ab}\varphi \Rightarrow U^t_{ab}\varphi$ and $\neg U^t_{ab}\varphi \Rightarrow \neg U^t_{ab}\varphi$.

(axioms of belief): $B^t_a\varphi \Rightarrow B^t_a\varphi$ and $\neg B^t_a\varphi \Rightarrow \neg B^t_a\varphi$.
$B^t_a\varphi \equiv B^t_a\psi$ if $\varphi \equiv \psi$.
$B^t_a(\varphi \wedge \psi) \equiv B^t_a\varphi \wedge B^t_a\psi$.

(axioms of inertia): $B^t_a\varphi \wedge B^{t+1}_a\varphi \Rightarrow B^{t+1}_a\varphi$ and $\neg B^t_a\varphi \wedge \neg B^{t+1}_a\varphi \Rightarrow \neg B^{t+1}_a\varphi$.

(axiom of truth): $B^t_a\top$ for any t.

(axiom of rationality): $\neg B^t_a\bot$ for any t if an agent a is *rational*.

(axiom of credibility): $U^t_{ab}\varphi \Rightarrow B^{t+1}_b\varphi$ if an agent b is *credulous*.

(axiom of reflection): $U^t_{ab}\varphi \Rightarrow B^{t+1}_b B^t_a\varphi$ if an agent b is *reflective*.

The axioms of utterance represent that an utterance or non-utterance has a cause of this. The axioms of belief represent similar effects. The axioms of inertia represent that an agent retains beliefs unless there is a reason to abandon it. The axiom of rationality

[2] $B^t_a\varphi$ is considered an atom such as "$b\_a\_t\_\varphi$".

is assumed for agents having consistent beliefs. The axiom of credibility represents that if a speaker utters a sentence then a hearer believes it. The axiom of reflection represents that if a speaker utters a sentence then a hearer believes that the speaker believes the sentence. The first four axioms are always assumed, while the last three axioms are conditionally assumed. Note that the axiom of rationality is identified with $B^t_a\varphi \supset \neg B^t_a\neg\varphi$ as:

$$\begin{aligned}\neg B^t_a\bot &\equiv \neg B^t_a(\varphi\wedge\neg\varphi) \ \text{(axioms of belief)}\\ &\equiv \neg(B^t_a\varphi\wedge B^t_a\neg\varphi) \ \text{(axioms of belief)}\\ &\equiv \neg B^t_a\varphi\vee\neg B^t_a\neg\varphi \ \text{(De Morgan's law)}\end{aligned}$$

A causal logic with these axioms is called the *epistemic causal logic*.

3 Deception in Epistemic Causal Logic

3.1 Deception by Lying

Deception is different from lying. Carson [3] says:

> "unlike 'lying' the word 'deception' connotes success. An act must actually mislead someone (cause someone to have false beliefs) if it is to count as a case of deception. Many lies are not believed and do not succeed in deceiving anyone" [3, p. 55].

He then illustrates the relationship between lying, deception, and attempted deception as in Figure 1.

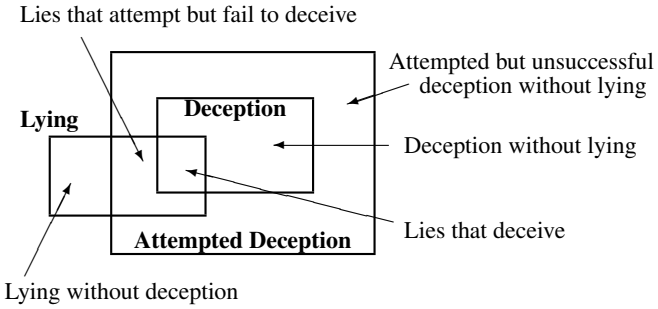

Fig. 1. Lying, deception and attempted deception [3]

Our primary interest in this section is to formulate "lies that deceive".[3] In this paper, we consider communication between two agents. Let a be an agent who utters a sentence (called a *speaker*), and b an agent who is an addressee (called a *hearer*). We first define the act of lying.

[3] Carson considers bluffing a type of lying and views deception by bluffing as lies that deceive. In this paper, we distinguish lying and bluffing, and view deception by bluffing as deception without lying.

Definition 1 (lying). Let a and b be two agents and φ a sentence. Then *lying* is defined as

$$\mathsf{LIE}^t_{ab}(\varphi) \stackrel{def}{=} B^t_a \neg\varphi \wedge U^t_{ab}\varphi \tag{2}$$

We say that *a lies* to *b* at *t* on the sentence φ.

By (2) a lies to b if a utters a believed-false sentence φ to b. Here we consider lying as a statement of a sentence, while it does not necessarily imply oral communication but it could be any type of communication using sentences. Note that a believes φ but the actual falsity of φ is not required in (2). Thus, if a speaker utters a believed-false sentence φ that is in fact true, then it is still lying. In this paper it does not matter whether lying involves *intention* to deceive.[4] Deception by lying is then defined as follows.

Definition 2 (deception by lying). Let a and b be two agents and φ a sentence. Then *deception by lying* (DBL) is defined as

$$\mathsf{DBL}^{t+1}_{ab}(\varphi) \stackrel{def}{=} \neg\varphi \wedge (\,\mathsf{LIE}^t_{ab}(\varphi) \Rightarrow B^{t+1}_b\varphi\,). \tag{3}$$

We say that *a deceives by lying* to *b* at $t+1$ on the sentence φ.

By (3) deception by lying is such that a lies to b at t on a false sentence φ, which causes b's believing φ at the next time step $t+1$. In contrast to lying, the actual falsity of φ is required. $\mathsf{DBL}^{t+1}_{ab}(\varphi)$ is often written as $\mathsf{DBL}_{ab}(\varphi)$ if time is unimportant in the context.

Example 2. Suppose a salesperson who lies that an investment is worth buying. The situation is represented by (3) with $a =$ salesperson, $b =$ customer, and $\varphi =$"an investment is worth buying" that is actually false. DBL then results in the customer's believing φ.

Note that (3) does not address whether a hearer believes φ *before* the act of lying. DBL happens whenever a hearer believes a false sentence φ as a result of lying by a speaker.

Example 3. Consider a theory

$$T = \mathsf{LIE}^t_{ab}(\varphi) \wedge \mathsf{DBL}^{t+1}_{ab}(\varphi).$$

Then T has two models:[5]

$$M_1 = \{\neg\varphi,\ B^t_a\neg\varphi,\ U^t_{ab}\varphi,\ B^{t+1}_a\neg\varphi,\ B^t_b\varphi,\ B^{t+1}_b\varphi\},$$
$$M_2 = \{\neg\varphi,\ B^t_a\neg\varphi,\ U^t_{ab}\varphi,\ B^{t+1}_a\neg\varphi,\ \neg B^t_b\varphi,\ B^{t+1}_b\varphi\}.$$

In both M_1 and M_2, a's belief in $\neg\varphi$ does not change from t to $t+1$ by the axioms of inertia. On the other hand, there are two different states of b's believing φ at t. M_1 represents that $\mathsf{DBL}^{t+1}_{ab}(\varphi)$ contributes causally toward b's continuing the false belief in φ. By contrast, M_2 represents that $\mathsf{DBL}^{t+1}_{ab}(\varphi)$ contributes causally toward b's acquiring the false belief in φ. Since $B^{t+1}_b\varphi$ is true in each model, it holds that

$$T \models B^{t+1}_b\varphi.$$

[4] We later consider intention in deceiving.

[5] Here, we omit $B^t_x\top$ and $B^{t+1}_x\top$ where $x = a, b$.

The situation of M_1 corresponds to *positive deception simpliciter* and M_2 corresponds to *positive deception secundum quid* in [5].

$\mathsf{DBL}_{ab}(\varphi)$ requires that φ is actually false. So if a speaker a utters a believed-false statement φ which is in fact true, then the speaker lies but DBL does not happen. The situation is formally stated as $\varphi \wedge \mathsf{DBL}^{t+1}_{ab}(\varphi) \models \bot$.

Example 4. A student, Bob, who believes that there will be no exam in tomorrow's math-class, tells his friend Mike that there will be an exam in tomorrow's math-class. Mike, who was absent from the math-class last week, believes Bob's information. The next day, it turns out that there is an exam in the class. In this case, Bob lies to Mike but Bob does not deceive Mike (and Mike does not believe that Bob lies to him).

Lying on a false sentence succeeds to deceive if a hearer is credulous.

Proposition 1. *Let a and b be two agents and φ a sentence. If b is credulous, then*

$$\neg\varphi \wedge \mathsf{LIE}^t_{ab}(\varphi) \models B^{t+1}_b \varphi\,.$$

Proof. Suppose an interpretation I such that $I \models \neg\varphi \wedge \mathsf{LIE}^t_{ab}(\varphi)$. Since b is credulous, $I \models U^t_{ab}\varphi$ implies $I \models B^{t+1}_b \varphi$ by the axiom of credibility. □

DBL on the valid sentence always fails. By contrast, DBL on the contradictory sentence fails if a hearer is rational.

Proposition 2. *Let a and b be two agents.*

- $\mathsf{DBL}^{t+1}_{ab}(\top) \models \bot$.
- $\mathsf{LIE}^t_{ab}(\bot) \wedge \mathsf{DBL}^{t+1}_{ab}(\bot) \models \bot$ *if b is rational.*

Proof. $\mathsf{DBL}^{t+1}_{ab}(\top) \equiv \bot$ by definition. $\mathsf{LIE}^t_{ab}(\bot) \wedge \mathsf{DBL}^{t+1}_{ab}(\bot)$ implies $B^{t+1}_b \bot$. If b is rational, this is impossible by the axiom of rationality. □

Suppose that a speaker a lies to b on the sentence φ but a hearer b already believes the contrary. If b is rational, then DBL on φ fails.

Proposition 3. *Let a and b be two agents and φ a sentence. If b is rational, then*

$$B^t_b\neg\varphi \wedge \mathsf{LIE}^t_{ab}(\varphi) \wedge \mathsf{DBL}^{t+1}_{ab}(\varphi) \models \bot\,.$$

Proof. $B^t_b\neg\varphi$ implies $B^{t+1}_b\neg\varphi$ by the axioms of inertia, and $B^{t+1}_b\neg\varphi$ implies $\neg B^{t+1}_b\varphi$ by the axiom of rationality. On the other hand, $\mathsf{LIE}^t_{ab}(\varphi) \wedge \mathsf{DBL}^{t+1}_{ab}(\varphi)$ implies $B^{t+1}_b\varphi$. Hence, $B^{t+1}_b\varphi \wedge \neg B^{t+1}_b\varphi \equiv \bot$. □

If a rational hearer is credulous, on the other hand, the hearer *revises* his/her belief and lying succeeds to deceive even if the hearer believes the contrary.

Proposition 4. *Let a and b be two agents and φ a sentence. If b is credulous and rational, then*

$$\neg\varphi \wedge B^t_b\neg\varphi \wedge \mathsf{LIE}^t_{ab}(\varphi) \models B^{t+1}_b\varphi\,.$$

Proof. For an interpretation $I \models \neg\varphi \wedge B_b^t \neg\varphi \wedge \mathsf{LIE}_{ab}^t(\varphi)$, $I \models U_{ab}^t\varphi$ implies $I \models B_b^{t+1}\varphi$ by the axiom of credibility, thereby $I \models \neg B_b^{t+1}\neg\varphi$ by the axiom of rationality. In this case, the axioms of inertia are not applied, and $I \models B_b^t\neg\varphi$ does not imply $I \models B_b^{t+1}\neg\varphi$. Then the result holds by Proposition 1. □

Suppose that a hearer is rational and reflective, and believes that a speaker is also rational. In this case, the hearer does not believe that a speaker is lying.

Proposition 5. *Let a and b be two agents and φ a sentence. If b is rational and reflective, and believes that a speaker a is rational, then*

$$\mathsf{LIE}_{ab}^t(\varphi) \wedge B_b^{t+1}(\mathsf{LIE}_{ab}^t(\varphi)) \models \bot.$$

Proof. $B_b^{t+1}(\mathsf{LIE}_{ab}^t(\varphi)) = B_b^{t+1}(B_a^t\neg\varphi \wedge U_{ab}^t\varphi)$ implies $B_b^{t+1}B_a^t\neg\varphi$ (axioms of belief). As b believes that a is rational, then $B_b^{t+1}(B_a^t\neg\varphi \supset \neg B_a^t\varphi)$ and $B_b^{t+1}B_a^t\neg\varphi$ imply $B_b^{t+1}\neg B_a^t\varphi$ (axioms of belief). Since b is reflective, $U_{ab}^t\varphi$ in $\mathsf{LIE}_{ab}^t(\varphi)$ implies $B_b^{t+1}B_a^t\varphi$ (axiom of reflection). Hence, $B_b^{t+1}\neg B_a^t\varphi \wedge B_b^{t+1}B_a^t\varphi \equiv B_b^{t+1}\bot$ (axioms of belief). This is impossible, since b is rational. □

Proposition 5 implies that if a rational and reflective hearer believes that a rational speaker is lying, there is no chance of DBL to succeed. Propositions 2, 3 and 5 characterize different situations where "lies that attempt but fail to deceive". In other words, they provide necessary conditions for DBL to succeed. If a rational hearer is not credulous, it is necessary that he/she does not believe to the contrary. If a rational hearer is reflective, it is necessary that he/she does not believe that a rational speaker is lying. Note that if an agent a successfully deceives another agent b by a lie φ, there is no guarantee that a can also deceive b using a stronger lie $\varphi \wedge \lambda$. A simple case is shown by putting $\lambda = \neg\varphi$, then $\mathsf{LIE}_{ab}^t(\varphi \wedge \neg\varphi) \wedge \mathsf{DBL}_{ab}^{t+1}(\varphi \wedge \neg\varphi)$ fails if b is rational.

3.2 Deception by Bluffing

We next provide an instance of "deception without lying" in Figure 1. When a speaker utters a sentence φ while he/she is uncertain about the truth of φ, we call it *bluffing*.

Definition 3 (bluffing). Let a and b be two agents and φ a sentence. Then *bluffing* is defined as

$$\mathsf{BLUF}_{ab}^t(\varphi) \stackrel{def}{=} \neg B_a^t\varphi \wedge \neg B_a^t\neg\varphi \wedge U_{ab}^t\varphi \tag{4}$$

We say that *a bluffs b* at t on the sentence φ.

By (4) a *bluffing agent* a believes neither φ nor $\neg\varphi$ when it utters φ. In case of lying (2), a speaker disbelieves φ but believes $\neg\varphi$. In bluffing a speaker also disbelieves $\neg\varphi$. The situation is also called *bullshit* in [13, 22, 24]. Deception by bluffing is then defined as follows.

Definition 4 (deception by bluffing). Let a and b be two agents and φ a sentence. Then *deception by bluffing* (DBB) is defined as

$$\mathsf{DBB}_{ab}^{t+1}(\varphi) \stackrel{def}{=} \neg\varphi \wedge (\mathsf{BLUF}_{ab}^t(\varphi) \Rightarrow B_b^{t+1}\varphi). \tag{5}$$

We say that *a deceives b* by bluffing at $t+1$ on the sentence φ.

$\mathsf{DBB}^{t+1}_{ab}(\varphi)$ is often written as $\mathsf{DBB}_{ab}(\varphi)$ if time is unimportant in the context.

Example 5. Bob, who does not know whether there will be an exam in tomorrow's math-class, tells his friend Mike that there will be no exam in tomorrow's math-class. Mike believes Bob's information. The next day, it turns out that there is an exam in the class. In this case, Bob deceives Mike by bluffing.

Like DBL, bluffing on a false sentence succeeds to deceive if a hearer is credulous.

Proposition 6. *Let a and b be two agents and φ a sentence. If b is credulous, then*

$$\neg\varphi \wedge \mathsf{BLUF}^t_{ab}(\varphi) \models B^{t+1}_b \varphi .$$

Proof. Suppose an interpretation I such that $I \models \neg\varphi \wedge \mathsf{BLUF}^t_{ab}(\varphi)$. Since b is credulous, $I \models U^t_{ab}\varphi$ implies $I \models B^{t+1}_b \varphi$ by the axiom of credibility. □

Both $\mathsf{DBB}_{ab}(\top)$ and $\mathsf{DBB}_{ab}(\bot)$ are inconsistent.

Proposition 7. *Let a and b be two agents. Then,*

$$\mathsf{DBB}^{t+1}_{ab}(\top) \vee \mathsf{DBB}^{t+1}_{ab}(\bot) \models \bot$$

Proof. Both $\mathsf{BLUF}^t_{ab}(\top)$ and $\mathsf{BLUF}^t_{ab}(\bot)$ imply $\neg B^t_a \top$ which violates the axiom of truth. The cause of (5) cannot be true, then $\mathsf{DBB}^{t+1}_{ab}(\top) \vee \mathsf{DBB}^{t+1}_{ab}(\bot)$ has no model. □

$\mathsf{DBB}_{ab}(\varphi)$ fails if a rational hearer believes $\neg\varphi$ at t.

Proposition 8. *Let a and b be two agents and φ a sentence. If b is rational, then*

$$B^t_b\neg\varphi \wedge \mathsf{BLUF}^t_{ab}(\varphi) \wedge \mathsf{DBB}^{t+1}_{ab}(\varphi) \models \bot .$$

Proof. $B^t_b\neg\varphi$ implies $B^{t+1}_b\neg\varphi$ (axioms of inertia), and $B^{t+1}_b\neg\varphi$ implies $\neg B^{t+1}_b\varphi$ (axiom of rationality). $\mathsf{BLUF}^t_{ab}(\varphi) \wedge \mathsf{DBB}^{t+1}_{ab}(\varphi)$ implies $B^{t+1}_b\varphi$. Hence, $B^{t+1}_b\varphi \wedge \neg B^{t+1}_b\varphi$. □

If a rational hearer is credulous, bluffing succeeds to deceive even if the hearer believes the contrary.

Proposition 9. *Let a and b be two agents and φ a sentence. If b is credulous and rational, then*

$$\neg\varphi \wedge B^t_b\neg\varphi \wedge \mathsf{BLUF}^t_{ab}(\varphi) \models B^{t+1}_b \varphi .$$

Proof. For an interpretation $I \models \neg\varphi \wedge B^t_b\neg\varphi \wedge \mathsf{BLUF}^t_{ab}(\varphi)$, $I \models U^t_{ab}\varphi$ implies $I \models B^{t+1}_b\varphi$ by the axiom of credibility, thereby $I \models \neg B^{t+1}_b\neg\varphi$ by the axiom of rationality. In this case, the axioms of inertia are not applied, and $I \models B^t_b\neg\varphi$ does not imply $I \models B^{t+1}_b\neg\varphi$. Then the result holds by Proposition 6. □

If a hearer is rational and reflective, he/she does not believe that a speaker is bluffing.

Proposition 10. *Let a and b be two agents and φ a sentence. If b is rational and reflective, then*

$$\mathsf{BLUF}^t_{ab}(\varphi) \wedge B^{t+1}_b(\mathsf{BLUF}^t_{ab}(\varphi)) \models \bot .$$

Proof. $B_b^{t+1}(\mathsf{BLUF}_{ab}^t(\varphi))$ implies $B_b^{t+1}(\neg B_a^t\varphi \wedge \neg B_a^t\neg\varphi)$, thereby $B_b^{t+1}\neg B_a^t\varphi$ (axioms of belief). As b is reflective, $U_{ab}^t\varphi$ in $\mathsf{BLUF}_{ab}^t(\varphi)$ implies $B_b^{t+1}B_a^t\varphi$ (axiom of reflection). Hence, $B_b^{t+1}\neg B_a^t\varphi \wedge B_b^{t+1}B_a^t\varphi \equiv B_b^{t+1}\bot$ (axioms of belief). This is impossible, since b is rational. □

Recall that attempted DBL fails if a rational and reflective hearer believes that a *rational* speaker is lying (Proposition 5). By contrast, attempted DBB fails if a rational and reflective hearer believes that a speaker is bluffing (Proposition 10). Note that in the latter case, it is not required that a hearer believes that a speaker is rational. The difference comes from $B_b^{t+1}(\mathsf{LIE}_{ab}^t(\varphi))$ and $B_b^{t+1}(\mathsf{BLUF}_{ab}^t(\varphi))$. $B_b^{t+1}(\mathsf{LIE}_{ab}^t(\varphi))$ implies $B_b^{t+1}B_a^t\neg\varphi$, while $B_b^{t+1}(\mathsf{BLUF}_{ab}^t(\varphi))$ implies $B_b^{t+1}\neg B_a^t\varphi$. To have $B_b^{t+1}\neg B_a^t\varphi$ in the former case, it is required that b believes that a is rational. This observation implies the next results.

Proposition 11. *Let a and b be two agents and φ a sentence. If b is rational and reflective, then*

- $\mathsf{LIE}_{ab}^t(\varphi) \wedge B_b^{t+1}(\mathsf{BLUF}_{ab}^t(\varphi)) \models \bot$.
- $\mathsf{BLUF}_{ab}^t(\varphi) \wedge B_b^{t+1}(\mathsf{LIE}_{ab}^t(\varphi)) \models \bot$ *if b believes that a is rational.*

By Propositions 10 and 11, attempted DBB fails if a rational and reflective hearer believes that (i) a speaker is bluffing, or (ii) a rational speaker is lying.

3.3 Deception by Truthful Telling

Truthful telling is the opposite of lying—a speaker utters a believed-true sentence.

Definition 5 (truthful telling). Let a and b be two agents and φ a sentence. Then *truthful telling* is defined as

$$\mathsf{TRT}_{ab}^t(\varphi) \stackrel{def}{=} B_a^t\varphi \wedge U_{ab}^t\varphi \tag{6}$$

We say that *a truthfully tells b* at t on the sentence φ.

The actual truth of φ is not the matter in (6). One may deceive others by honestly telling what he/she believes to be true.

Example 6. Bob, who believes that there will be no exam in tomorrow's math-class, tells his friend Mike that there will be no exam in tomorrow's math-class. Mike believes Bob's information. The next day, it turns out that there is an exam in the class. In this case, Bob deceives Mike by truthful telling.

The above example illustrates another instance of "deception without lying". We call this type "deception by truthful telling" that is formally defined as follows.

Definition 6 (deception by truthful telling). Let a and b be two agents and φ a sentence. Then *deception by truthful telling* (DBT) is defined as

$$\mathsf{DBT}_{ab}^{t+1}(\varphi) \stackrel{def}{=} \neg\varphi \wedge (\mathsf{TRT}_{ab}^t(\varphi) \Rightarrow B_b^{t+1}\varphi). \tag{7}$$

In (7) a's truthful utterance of φ makes a hearer b believe a false sentence φ. DBT is less malicious than DBL or DBB because a speaker truthfully tells a (misbelieved) fact. $\mathsf{DBT}^{t+1}_{ab}(\varphi)$ is often written as $\mathsf{DBT}_{ab}(\varphi)$ if time is unimportant in the context. By definition, DBL, DBB and DBT are mutually exclusive. Like DBL and DBB, truthful telling on a false sentence succeeds to deceive if a hearer is credulous.

Proposition 12. *Let a and b be two agents and φ a sentence. If b is credulous, then*

$$\neg\varphi \wedge \mathsf{TRT}^{t}_{ab}(\varphi) \models B^{t+1}_{b}\varphi\,.$$

Proof. Suppose an interpretation I such that $I \models \neg\varphi \wedge \mathsf{TRT}^{t}_{ab}(\varphi)$. Since b is credulous, $I \models U^{t}_{ab}\varphi$ implies $I \models B^{t+1}_{b}\varphi$ by the axiom of credibility. □

Proposition 13. *Let a and b be two agents. Then,*

$$\mathsf{DBT}^{t+1}_{ab}(\top) \models \bot\,.$$

Proof. When $\varphi \equiv \top$, (7) becomes $\bot$. □

By contrast, $\mathsf{DBT}^{t+1}_{ab}(\bot)$ may succeed if both a speaker and a hearer are irrational.

Proposition 14. *Let a and b be two agents. If both a and b are irrational, then*

$$\mathsf{TRT}^{t}_{ab}(\bot) \wedge \mathsf{DBT}^{t+1}_{ab}(\bot) \models B^{t+1}_{a}\bot \wedge B^{t+1}_{b}\bot.$$

Proof. The result holds by definition and the axioms of inertia. □

Proposition 14 characterizes a situation that an irrational speaker believes a false sentence, and he/she truthfully tells a hearer who is also irrational. As a result, the hearer believes the false sentence. As an example, a mathematician a claims that he/she finds a proof of squaring the circle, which is known to be impossible today. A hearer b, who is not well-informed in mathematics, believes it. This is a case of deception by truthful telling of $\mathsf{DBT}_{ab}(\bot)$. $\mathsf{DBT}^{t+1}_{ab}(\varphi)$ fails if a rational hearer believes $\neg\varphi$ at t.

Proposition 15. *Let a and b be two agents and φ a sentence. If b is rational, then*

$$B^{t}_{b}\neg\varphi \wedge \mathsf{TRT}^{t}_{ab}(\varphi) \wedge \mathsf{DBT}^{t+1}_{ab}(\varphi) \models \bot.$$

Proof. $B^{t}_{b}\neg\varphi$ implies $B^{t+1}_{b}\neg\varphi$ (axioms of inertia), and $B^{t+1}_{b}\neg\varphi$ implies $\neg B^{t+1}_{b}\varphi$ (axiom of rationality). $\mathsf{TRT}^{t}_{ab}(\varphi) \wedge \mathsf{DBT}^{t+1}_{ab}(\varphi)$ implies $B^{t+1}_{b}\varphi$. Hence, $B^{t+1}_{b}\varphi \wedge \neg B^{t+1}_{b}\varphi$. □

If a rational hearer is credulous, truthful telling succeeds to deceive even if the hearer believes the contrary.

Proposition 16. *Let a and b be two agents and φ a sentence. If b is credulous and rational, then*

$$\neg\varphi \wedge B^{t}_{b}\neg\varphi \wedge \mathsf{TRT}^{t}_{ab}(\varphi) \models B^{t+1}_{b}\varphi\,.$$

Proof. For an interpretation $I \models \neg\varphi \wedge B^{t}_{b}\neg\varphi \wedge \mathsf{TRT}^{t}_{ab}(\varphi)$, $I \models U^{t}_{ab}\varphi$ implies $I \models B^{t+1}_{b}\varphi$ by the axiom of credibility, thereby $I \models \neg B^{t+1}_{b}\neg\varphi$ by the axiom of rationality. In this case, the axioms of inertia are not applied, and $I \models B^{t}_{b}\neg\varphi$ does not imply $I \models B^{t+1}_{b}\neg\varphi$. Then the result holds by Proposition 12. □

DBT will not happen if a rational and reflective hearer believes that a (rational) speaker is lying/bluffing.

Proposition 17. *Let a and b be two agents and φ a sentence. Suppose that a hearer b is rational and reflective.*

- $\mathsf{TRT}^t_{ab}(\varphi) \wedge B^{t+1}_b(\mathsf{LIE}^t_{ab}(\varphi)) \models \bot$ *if b believes that a is rational.*
- $\mathsf{TRT}^t_{ab}(\varphi) \wedge B^{t+1}_b(\mathsf{BLUF}^t_{ab}(\varphi)) \models \bot$.

Proof. $U^t_{ab}\varphi$ in $\mathsf{TRT}^t_{ab}(\varphi)$ implies $B^{t+1}_b B^t_a \varphi$ by the axiom of reflection. $B^{t+1}_b(\mathsf{LIE}^t_{ab}(\varphi))$ implies $B^{t+1}_b B^t_a \neg\varphi$, and b's believing a's rationality implies $B^{t+1}_b \neg B^t_a \varphi$. Then, $B^{t+1}_b B^t_a \varphi \wedge B^{t+1}_b \neg B^t_a \varphi \equiv B^{t+1}_b \bot$ (axioms of belief). Also, $B^{t+1}_b(\mathsf{BLUF}^t_{ab}(\varphi))$ implies $B^{t+1}_b \neg B^t_a \varphi$. Then, $B^{t+1}_b B^t_a \varphi \wedge B^{t+1}_b \neg B^t_a \varphi \equiv B^{t+1}_b \bot$. This is impossible, since b is rational. □

3.4 Deception by Omission

Sometimes deception is done by *withholding information*. For instance, suppose a person who is selling a used car that has some problem in its engine. If he/she sells the car without informing a customer of the problem, it is deception by withholding information [3]. It is also called *deception by omission*, which is contrasted with *deception by commission* that involves an act of providing information [5].

Definition 7 (withholding information). Let a and b be two agents and φ a sentence. Then *withholding information* (WI) is defined as

$$\mathsf{WI}^t_{ab}(\varphi) \stackrel{def}{=} B^t_a \varphi \wedge \neg U^t_{ab}\varphi. \tag{8}$$

Deception by omission is then defined as follows.

Definition 8 (deception by omission). Let a and b be two agents and φ a sentence. Then *deception by omission* (DBO) is defined as

$$\mathsf{DBO}^{t+1}_{ab}(\varphi) \stackrel{def}{=} \varphi \wedge (\mathsf{WI}^t_{ab}(\varphi) \Rightarrow \neg B^{t+1}_b \varphi). \tag{9}$$

By (9) deception by omission happens when a speaker believes a true fact while provides no information of it. As a result, a hearer disbelieves the fact. $\mathsf{DBO}^{t+1}_{ab}(\varphi)$ is often written as $\mathsf{DBO}_{ab}(\varphi)$ if time is unimportant in the context.

Proposition 18. *Let a and b be two agents.*

- $\mathsf{DBO}^{t+1}_{ab}(\bot) \models \bot$.
- $\mathsf{WI}^t_{ab}(\top) \wedge \mathsf{DBO}^{t+1}_{ab}(\top) \models \bot$.

Proof. By definition, $\mathsf{DBO}^{t+1}_{ab}(\bot)$ has no model. When $\varphi \equiv \top$, the effect of (9) is $\neg B^{t+1}_b \top$. However, $\neg B^{t+1}_b \top$ does not happen by the axiom of truth. □

$\mathsf{DBO}^{t+1}_{ab}(\varphi)$ fails if a hearer is a rational agent who believes φ at t.

Proposition 19. *Let a and b be two agents and φ a sentence. If b is rational, then*

$$B_b^t\varphi \wedge \mathsf{WI}_{ab}^t(\varphi) \wedge \mathsf{DBO}_{ab}^{t+1}(\varphi) \models \bot.$$

Proof. $B_b^t\varphi$ implies $B_b^{t+1}\varphi$ by the axioms of inertia, while $\mathsf{WI}_{ab}^t(\varphi) \wedge \mathsf{DBO}_{ab}^{t+1}(\varphi)$ implies $\neg B_b^{t+1}\varphi$. Hence, $B_b^{t+1}\varphi \wedge \neg B_b^{t+1}\varphi$, and the result holds. □

One may argue that it would not be appropriate to regard (9) as deception. It happens that a person a does not tell his/her belief φ to another person b, and even if it results in b's ignorance of the true sentence φ it is not called deception. Deception by omission is usually accompanied with an intention of concealing. Then we argue deception accompanied with an intention in the next section.

4 Various Aspects of Deception

4.1 Intentional Deception

Deception is often distinguished between *intentional deception* and *unintentional* one [5].[6] DBL, DBB, DBT and DBO in Section 3 represent unintentional deception, that is, a speaker does not necessarily intend to deceive a hearer. In $\mathsf{DBL}_{ab}(\varphi)$, a speaker a lies a believed-false sentence φ to a hearer b, while the speaker may not believe that lying will result in the hearer's believing the false sentence φ. For instance, when a speaker says something manifestly false as a joke, he/she does not expect a hearer to believe it. In $\mathsf{DBT}_{ab}(\varphi)$, a speaker tells his/her belief to a hearer, so he/she would not expect it to result in deceiving. To formulate a speaker's intention to deceive, four types of deception are respectively modified as follows.

Definition 9 (intentional deception). Let a and b be two agents and φ, ψ sentences. Then, *intentional deception by lying* (I-DBL), *intentional deception by bluffing* (I-DBB), *intentional deception by truthful telling* (I-DBT), and *intentional deception by omission* (I-DBO) are respectively defined as follows.

$$\mathsf{I\text{-}DBL}_{ab}^{t+1}(\varphi) \stackrel{def}{=} \neg\varphi \wedge (\mathsf{LIE}_{ab}^t(\varphi) \wedge B_a^t B_b^{t+1}\varphi \Rightarrow B_b^{t+1}\varphi)$$
$$\mathsf{I\text{-}DBB}_{ab}^{t+1}(\varphi) \stackrel{def}{=} \neg\varphi \wedge (\mathsf{BLUF}_{ab}^t(\varphi) \wedge B_a^t B_b^{t+1}\varphi \Rightarrow B_b^{t+1}\varphi)$$
$$\mathsf{I\text{-}DBT}_{ab}^{t+1}(\varphi,\psi) \stackrel{def}{=} \neg\psi \wedge (\mathsf{TRT}_{ab}^t(\varphi) \wedge B_a^t(B_b^{t+1}\varphi \supset B_b^{t+1}\psi) \wedge B_a^t\neg\psi \Rightarrow B_b^{t+1}\psi)$$
$$\mathsf{I\text{-}DBO}_{ab}^{t+1}(\varphi) \stackrel{def}{=} \varphi \wedge (\mathsf{WI}_{ab}^t(\varphi) \wedge B_a^t\neg B_b^t\varphi \Rightarrow \neg B_b^{t+1}\varphi)$$

As before, $\mathsf{I\text{-}DBL}_{ab}^{t+1}(\varphi)$ (resp. $\mathsf{I\text{-}DBB}_{ab}^{t+1}(\varphi)$, $\mathsf{I\text{-}DBT}_{ab}^{t+1}(\varphi,\psi)$, and $\mathsf{I\text{-}DBO}_{ab}^{t+1}(\varphi)$) is often written as $\mathsf{I\text{-}DBL}_{ab}(\varphi)$ (resp. $\mathsf{I\text{-}DBB}_{ab}(\varphi)$, $\mathsf{I\text{-}DBT}_{ab}(\varphi,\psi)$, and $\mathsf{I\text{-}DBO}_{ab}(\varphi)$) if time is unimportant in the context.

[6] The meaning of the term "intentional deception" is different from "attempted deception" in Fig. 1. Intentional deception is a type of deception that involves the success of deceiving, while this is not always the case in attempted deception.

In I-DBL and I-DBB, the additional formula $B_a^t B_b^{t+1}\varphi$ in the cause says that a speaker a believes that a hearer b will believe the false sentence φ in the next time step. With this belief a utters φ to b at t, which we consider that a has an intention to deceive b. In I-DBT, on the other hand, a speaker a truthfully tells φ while a believes that a hearer b's believing φ leads to b's believing another sentence ψ in the next time step. Moreover, a believes the falsity of ψ and it is in fact false, which causes b's believing ψ. We call it intentional deception by truthful-telling because a expects that b's believing a believed-false sentence. In I-DBO, a speaker a withholds φ while believing b's ignorance of φ, which causes b's disbelieving φ (or prevents b from believing φ) in the next time step. In this case, we consider that a has an intention to conceal φ from b and call it intentional deception by omission.

Note that we do not introduce an additional predicate such as I_a to represent intention. Instead, we represent intention of a speaker by encoding a fact that a speaker recognizes the effect of his/her deceptive act on the hearer. Since intentional deception introduces additional causes to unintentional one, formal properties addressed in Section 3 hold for intentional deception as well (except DBT). When the distinction between intentional and unintentional DBL (resp. DBB, DBT, or DBO) is unimportant, we write as (I-)DBL (resp. (I-)DBB, (I-)DBT or (I-)DBO).

For a rational speaker DBT can be intentional only if a hearer comes to believe a false sentence that is *different* from the sentence of utterance.

Proposition 20. *Let a and b be two agents and φ a sentence. If a is rational, then*

$$\text{I-DBT}_{ab}(\varphi,\varphi) \models \bot.$$

Proof. $\text{TRT}^t_{ab}(\varphi)$ involves $B_a\varphi$ while $\text{I-DBT}_{ab}(\varphi,\varphi)$ contains $B_a\neg\varphi$ in its cause. Then $B_a\varphi \wedge B_a\neg\varphi \equiv B_a\bot$, which violates the axiom of rationality. □

Compared to others, I-DBT generally requires advanced techniques for a speaker because a deceiver is requested to select a sentence to be uttered that is different from the false fact which the deceiver wants a hearer to believe. The situation captures some feature of deception that "the deceiver takes a more circuitous route to his success, where lying is an easier and more certain way to mislead" [1, p.440]. According to studies in psychology, children lie by four years or earlier, mainly for avoiding punishment [10]. Very young children do not have advanced techniques of deception, then most deception by them is of the type (I-)$\text{DBL}_{ab}(\varphi)$ or (I-)$\text{DBB}_{ab}(\varphi)$ or (I-)$\text{DBO}_{ab}(\varphi)$ that is the most simple form of deception.

4.2 Indirect Deception

Suppose that an agent a lies to another agent b on a false sentence φ. Then b, who believes φ, truthfully tells φ to another agent c, which results in c's believing the false sentence φ. In this case, is a deceiving c as well as b?

Example 7. John, who visits a clinic for a medical check-up, is diagnosed as having a serious cancer. A doctor does not want to discourage him and lies to John that he is normal. John has no symptom giving him no reason to believe his cancer, and he told his wife that the result of a medical test is normal. In this scenario, a doctor intentionally deceives John by lying and John unintentionally deceives his wife by truthful telling.

The situation of Example 7 is represented in our formulation as: $\mathsf{I\text{-}DBL}^{t+1}_{ab}(\varphi) \wedge \mathsf{DBT}^{t+2}_{bc}(\varphi)$ where $a =$ doctor, $b =$ John, $c =$ wife, and $\varphi =$normal. In this case, we consider that a doctor indirectly deceives John's wife. Generally, acts of deceiving produce indirect deception as follows.

Definition 10 (indirect deception). Let a, b and c be three agents and φ, ψ sentences. Then *indirect deception by lying* (IN-DBL), *indirect deception by bluffing* (IN-DBB), *indirect deception by truthful telling* (IN-DBT), and *indirect deception by omission* (IN-DBO) are defined as follows:

$$\mathsf{IN\text{-}DBL}_{ac}(\varphi) \stackrel{def}{=} \mathsf{(I\text{-})DBL}^{t+1}_{ab}(\varphi) \wedge \mathsf{DBT}^{t+2}_{bc}(\varphi).$$
$$\mathsf{IN\text{-}DBB}_{ac}(\varphi) \stackrel{def}{=} \mathsf{(I\text{-})DBB}^{t+1}_{ab}(\varphi) \wedge \mathsf{DBT}^{t+2}_{bc}(\varphi).$$
$$\mathsf{IN\text{-}DBT}_{ac}(\varphi) \stackrel{def}{=} \mathsf{DBT}^{t+1}_{ab}(\varphi) \wedge \mathsf{DBT}^{t+2}_{bc}(\varphi).$$
$$\mathsf{IN\text{-}DBO}_{ac}(\varphi) \stackrel{def}{=} \mathsf{(I\text{-})DBO}^{t+1}_{ab}(\varphi) \wedge \neg U^{t+1}_{bc}\varphi \Rightarrow \neg B^{t+2}_{c}\varphi.$$
$$\mathsf{IN\text{-}I\text{-}DBT}_{ac}(\varphi, \psi) \stackrel{def}{=} \mathsf{I\text{-}DBT}^{t+1}_{ab}(\varphi, \psi) \wedge \mathsf{DBT}^{t+2}_{bc}(\psi).$$

In $\mathsf{IN\text{-}DBL}_{ac}(\varphi)$, a's lying on a sentence φ results in b's believing a false sentence φ, and then b's truthful telling on φ results in c's believing φ. $\mathsf{IN\text{-}DBB}_{ac}(\varphi)$ and $\mathsf{IN\text{-}DBT}_{ac}(\varphi)$ represent similar situations. In $\mathsf{IN\text{-}DBO}_{ac}(\varphi)$, a's withholding φ results in b's disbelieving a true sentence φ. Then b does not inform c of φ, which results in c's disbelieving φ. $\mathsf{IN\text{-}I\text{-}DBT}_{ac}(\varphi, \psi)$ represents indirect DBT that accompanies intention. By definition, indirect deception IN-DBL, IN-DBB, IN-DBT and IN-I-DBT succeed iff both a's deceiving b and b's deceiving c succeed. In contrast, IN-DBO succeeds if a's deceiving b succeeds.

In each definition, an agent a may have intention to deceive b, while an agent b does not have intention to deceive c. If an agent b also has intention to deceive c, then b is actively involved in the deceptive act. As a result, a is less responsible for c's being deceived, and we do not call it indirect deception. Note also that in each definition, an agent b makes a truthful statement (or no statement in case of IN-DBO). If this is not the case, suppose that

$$\mathsf{(I\text{-})DBL}^{t+1}_{ab}(\varphi) \wedge \mathsf{DBL}^{t+2}_{bc}(\neg\varphi).$$

$\mathsf{(I\text{-})DBL}^{t+1}_{ab}(\varphi)$ causes $B^{t+1}_{b}\varphi$, then b lies to c on the contrary $\neg\varphi$. In this case, $\mathsf{(I\text{-})DBL}^{t+1}_{ab}(\varphi)$ requires $\neg\varphi$ in the precondition, while $\mathsf{DBL}^{t+2}_{bc}(\neg\varphi)$ requires φ, which is impossible.

Generally, indirect deception could be chained like

$$\mathsf{(I\text{-})DBL}^{t+1}_{ab}(\varphi) \wedge \mathsf{DBT}^{t+2}_{bc}(\varphi) \wedge \mathsf{DBT}^{t+3}_{cd}(\varphi) \wedge \cdots$$

Such a situation happens when retweeting fake information on social media.

4.3 Self-Deception

Self-deception is an act of deceiving the self. Due to its paradoxical nature, self-deception has been controversial in philosophy and psychology [8, 9, 19, 30]. It is said that self-deception involves a person holding contradictory beliefs $B_a\bot$, or believing and disbelieving the same sentence at the same time $B_a\varphi \wedge \neg B_a\varphi$. In each case, it violates the

classical principle of consistency that rational agents are assumed to follow.[7] In this section, we characterize self-deception in our formulation.

Definition 11 (self-deception). Let a be an agent and φ, ψ sentences. Then, $(\mathsf{I}\text{-})\mathsf{DBL}_{aa}(\varphi)$, $(\mathsf{I}\text{-})\mathsf{DBB}_{aa}(\varphi)$, $\mathsf{DBT}_{aa}(\varphi)$, $\mathsf{I}\text{-}\mathsf{DBT}_{aa}(\varphi,\psi)$, and $(\mathsf{I}\text{-})\mathsf{DBO}_{aa}(\varphi)$ are called *self-deception*.

As such, a speaker and a hearer are identical in self-deception. Consequently, conflict may arise between belief as a speaker and belief as a hearer.

Proposition 21. *Let a be an agent and φ a sentence. Then,*

$$\mathsf{LIE}^t_{aa}(\varphi) \wedge \mathsf{DBL}^{t+1}_{aa}(\varphi) \models B^{t+1}_a \bot.$$

Proof. By definition, $\mathsf{LIE}^t_{aa}(\varphi)$ implies $B^t_a \neg\varphi$ which implies $B^{t+1}_a \neg\varphi$ by the axioms of inertia. On the other hand, $\mathsf{DBL}^{t+1}_{aa}(\varphi)$ implies $B^{t+1}_a \varphi$. Hence, $B^{t+1}_a \neg\varphi \wedge B^{t+1}_a \varphi \equiv B^{t+1}_a \bot$. □

Proposition 21 shows that $\mathsf{DBL}^{t+1}_{aa}(\varphi)$ involves a mental state of an agent who has contradictory belief wrt a false fact φ. This is possible only when the agent is irrational.[8] On the other hand, if a rational agent is credulous, self-deception does not involve contradictory belief.

Proposition 22. *Let a be an agent and φ a sentence. If a is credulous and rational, then*

$$\mathsf{LIE}^t_{aa}(\varphi) \wedge \mathsf{DBL}^{t+1}_{aa}(\varphi) \not\models B^{t+1}_a \bot.$$

Proof. If a is credulous, $U^t_{aa}\varphi$ implies $B^{t+1}_a \varphi$ by the axiom of credibility. As a is rational, $B^{t+1}_a \varphi$ implies $\neg B^{t+1}_a \neg\varphi$ by the axiom of rationality. Then the axioms of inertia do not produce $B^{t+1}_a \neg\varphi$ from $B^t_a \neg\varphi$. Hence, $B^{t+1}_a \bot$ is not entailed. □

Proposition 22 shows that a credulous agent revises its belief from $B^t_a \neg\varphi$ to $B^{t+1}_a \varphi$. As a result, contradictory belief is not produced. Propositions 21 and 22 are directly extended to $\mathsf{I}\text{-}\mathsf{DBL}^{t+1}_{aa}(\varphi)$. Next, suppose that an agent self-deceives by bluffing.

$$\mathsf{BLUF}^t_{aa}(\varphi) \wedge \mathsf{DBB}^{t+1}_{aa}(\varphi) \models \neg B^t_a \varphi \wedge B^{t+1}_a \varphi.$$

Thus, $\mathsf{BLUF}^t_{aa}(\varphi) \wedge \mathsf{DBB}^{t+1}_{aa}(\varphi)$ implies $B^{t+1}_a \varphi$ then the axioms of inertia do not imply $\neg B^{t+1}_a \varphi$ from $\neg B^t_a \varphi$. So a does not have contradictory belief as a result of $\mathsf{DBB}_{aa}(\varphi)$. In case of self-deception by omission, it holds that

$$\mathsf{WI}^t_{aa}(\varphi) \wedge \mathsf{DBO}^{t+1}_{aa}(\varphi) \models B^t_a \varphi \wedge \neg B^{t+1}_a \varphi.$$

An agent a believes a true sentence φ at t while he/she does not refer to it. Then a does not believe it in the next time step. The situation represents that a person, who believes something true but does not refer to it, will *forget* it. It is interesting to observe that the

[7] "In short, self-deception involves an inner conflict, perhaps the existence of contradiction" [9, p. 588].

[8] Jones [17] characterizes a group of "self-deception positions" consistently using KD as the logic of belief.

effects of $\mathsf{DBB}_{aa}(\varphi)$ and $\mathsf{DBO}_{aa}(\varphi)$ are symmetric. In case of $\mathsf{DBB}_{aa}(\varphi)$, a disbelieves φ at t but believes it at $t+1$; in case of $\mathsf{DBO}_{aa}(\varphi)$, on the other hand, a believes φ at t but disbelieves it at $t+1$.

$\mathsf{DBT}_{aa}(\varphi)$ does not involve inconsistency by definition, while inconsistency arises if it is accompanied by intention. This is because $\mathsf{I\text{-}DBT}^{t+1}_{aa}(\varphi,\psi)$ has $B^t_a\neg\psi$ in its cause and $B^{t+1}_a\psi$ in its effect. Since $B^t_a\neg\psi$ implies $B^{t+1}_a\neg\psi$ by the axioms of inertia, a will have the contradictory belief $B^{t+1}_a\bot$. In contrast, $\mathsf{I\text{-}DBB}_{aa}(\varphi)$ and $\mathsf{I\text{-}DBO}_{aa}(\varphi)$ do not involve inconsistency.

When self-deception implies $B^{t+1}_a\bot$, a rational agent cannot deceive oneself. On the other hand, contradiction does not arise if the axioms of inertia are not assumed. McLaughlin [20] argues that one can intentionally deceive oneself by losing relevant disbelief by the time one is to be taken in by the deceitful act. The following scenario is a modification of the "appointment example" of [20, p. 31].[9]

Example 8. There is a meeting three months ahead, say, on March 31. Mary is a member of the meeting but she is unwilling to attend it. She then deliberately recorded the wrong date, say, April 1st, for the meeting in her online calendar. Mary is very busy and has completely forgotten the actual date of the meeting. On April 1st, her online assistant informs her of the meeting, and she realizes that she missed the meeting.

Indirect self-deception is represented by putting $a=c$ in Definition 10. The situation of Example 8 is then represented by IN-DBL as

$$\mathsf{IN\text{-}DBL}_{aa}(\varphi)=\mathsf{I\text{-}DBL}^{t+1}_{ab}(\varphi)\wedge\mathsf{DBT}^{t+2}_{ba}(\varphi)$$

where $a=$ Mary, $b=$ online assistant, and $\varphi=$"Meeting on April 1st". It holds that

$$\mathsf{LIE}^t_{ab}(\varphi)\wedge B^t_aB^{t+1}_b\varphi\wedge\mathsf{I\text{-}DBL}^{t+1}_{ab}(\varphi)\wedge\mathsf{TRT}^{t+1}_{ba}(\varphi)\wedge\mathsf{DBT}^{t+2}_{ba}(\varphi)\models B^{t+2}_a\neg\varphi\wedge B^{t+2}_a\varphi.$$

In the absence of the axioms of inertia, however, $B^{t+2}_a\neg\varphi$ is not entailed, so $\mathsf{IN\text{-}DBL}_{aa}(\varphi)$ succeeds. As observed in this subsection, *self-deception does not always involve contradictory belief*. One can deceive the self using (I-)DBB, (I-)DBO or DBT. Moreover, if one does not retain his/her own belief over time, one can deceive himself/herself by IN-DBL without introducing contradictory belief. To the best of our knowledge, this is a new finding that has not yet been formally reported in the literature.

5 Related Work

There are some studies attempting to formulate deception using modal logic. Firozabadi *et al.* [12] formulate deception using a modal logic of action. According to their definition, an action of an agent is considered deceptive if he/she either does not have a belief about the truth value of some proposition but makes another agent believe that the proposition is true or false, or he/she believes that the proposition is true/false but makes another agent believe the opposite. These two cases are formally represented as: $\neg B_a\varphi\wedge E_aB_b\varphi$ or $B_a\neg\varphi\wedge E_aB_b\varphi$ where $E_a\psi$ means "an agent a brings about that ψ".

[9] McLaughlin calls it "self-induced deception".

Their formulation represents the result of deceptive action but does not represent which type of (speech) acts bring about false belief on a hearer. O'Neill [21] formulates deception using a modal logic of intentional communication. According to his definition, deception happens when a intends b to believe something that a believes to be false, and b believes it. The situation is formally represented as: $Dec_{ab}\,\varphi := I_a B_b \varphi \wedge B_a \neg\varphi \wedge B_b \varphi$. Attempted deception is defined by removing the conjunct $B_b\varphi$ in $Dec_{ab}\,\varphi$. $Dec_{ab}\,\varphi$ does not represent that b comes to have a false belief φ as a result of an action by a. Thus, a deceives b when b believes φ without any action of a. The problem comes from the fact that their logic has no mechanism of representing an action and its effect. Baltag and Smets [2] introduce a logic of conditional doxastic actions. $Lie_a(\varphi)$ represents an action in which an agent a publicly lies that she knows φ while in fact she does not know it. $True_a(\varphi)$ represents an action in which a makes a public truthful announcement that she knows φ. They have preconditions $\neg K_a\varphi$ and $K_a\varphi$, respectively. If a hearer already knows that φ is false, the action $Lie_a(\varphi)$ does not succeed. Hence, it does not allow a hearer's belief revision. The precondition $\neg K_a\varphi$ of $Lie_a(\varphi)$ represents the ignorance of φ which is not considered lying but bluffing in this paper. They argue deception by lying but do not distinguish it from deception without lying.

Jones [17] analyzes self-deception in the form of the Montaigne-family (e.g. $\neg B_a\varphi \wedge B_a B_a\varphi$) and concludes that self-deception cannot be represented in the logic of belief KD45 in a consistent manner. da Costa and French [8] formulates the inconsistent aspects of self-deception using *paraconsistent doxastic logic*. Those studies, as well as most philosophical studies, view self-deception as a state of mind having contradictory or inconsistent belief and argue how to resolve it. In contrast, we capture self-deception as an instance of deception in which a speaker and a hearer are identical. It is formulated not as a static belief state of an agent but as an effect of a cause in the belief state. In this setting we show that self-deception does not always involve contradiction.

Van Ditmarsch *et al*. [32, 33] study dynamic aspects of lying and bluffing using dynamic epistemic logic. It provides logics for different types of agents and investigates how the belief of an agent is affected by (un)truthful announcements. In their study, truthful announcements are not used for misleading hearers. Sarkadi *et al*. [22, 27] model deceptive agents using a BDI-like architecture and realize it in an agent-oriented programming language. The proposed model employs a *theory of mind* to analyze deceptive interactions among agents. The language has rich vocabularies to represent mental states of agents as well as reasoning mechanisms such as default reasoning and backward induction. The goal of their study is building a computational model for deceptive agents and implementing it in multi-agent environments.

Sakama *et al.* [24] formulate deception in which a speaker makes a truthful statement expecting that a hearer will misuse it to draw a wrong conclusion. It is similar to intentional deception by truthful telling in this paper, while it does not represent the *effect* of a deceptive act on a hearer's side. In this sense, deception formulated in [24] corresponds to attempted deception in this paper. Sakama and Caminada [25] provide a logical account of different categories of deception that were given by [5]. They use a modal logic of action and belief developed by [23], which is different from our current formulation. Moreover, the study does not distinguish deception by lying and deception without lying, as done in this paper. Sakama *et al.* [26] study logical account of

lies, bullshit, and withholding information, while their use in deception is not formally handled.

6 Concluding Remarks

This paper introduced a formal account of deception using an epistemic causal logic that can express both an act of deceiving and its effect on hearers' belief. We formulated different types of deception and argued their semantic properties. The current study focuses on the declarative aspect of deception. From the computational perspective, a causal rule of the form: $\ell_1 \wedge \cdots \wedge \ell_n \Rightarrow \ell_0$, where ℓ_i is a literal, is translated into a logic programming rule: $\ell_0 \leftarrow not\ \overline{\ell_1}, \ldots, not\ \overline{\ell_n}$ under the answer set semantics, where *not* is negation as failure and $\overline{\ell_i}$ is the literal complementary to ℓ_i [14]. Since the causal theory used in this paper consists of rules of this form, deception introduced in this paper could be implemented using logic programming.

The framework introduced in this paper is simple and has room for further extension. There is a situation in which a speaker simulates a conclusion that a hearer is likely to reach based on information the speaker provides and inference the hearer could execute. For instance, intentional deception by lying $\mathsf{I\text{-}DBL}^{t+1}_{ab}(\varphi)$ in Section 4.1 is extended to $\neg\psi \wedge (\mathsf{LIE}^t_{ab}(\varphi) \wedge B^t_a B^t_b(\varphi \supset \psi) \wedge B^t_a \neg\psi \Rightarrow B^{t+1}_b \psi)$, where a speaker a simulates modus ponens executed by a hearer b and lying on φ brings about hearer's believing the false fact ψ. As such, the current framework is extended to handle more complicated cases by taking *a theory of mind* into consideration. Those extensions are left for future study.

References

1. Adler, J. E.: Lying, deceiving, or falsely implicating. *Journal of Philosophy* 94:435–452 (1997)
2. Baltag, A., Smets, S.: The logic of conditional doxastic actions. In: Apt, K. R. and van Rooij, R. eds. *New Perspectives on Games and Interaction*, Texts in Logic and Games 4, 9–31, Amsterdam University Press (2008)
3. Carson, T. L.: *Lying and Deception: Theory and Practice*. Oxford University Press (2010)
4. Castelfranchi, C.: Artificial liars: why computers will (necessarily) deceive us and each other? *Ethics and Information Technology* 2:113–119 (2000)
5. Chisholm, R. M., Feehan, T. D.: The intent to deceive. *Journal of Philosophy* 74:143–159 (1977)
6. Chiluwa, I. E., Samoilenko, S. A. (Eds.).: *Handbook of research on deception, fake news, and misinformation online*. Information Science Reference/IGI Global. https://doi.org/10.4018/978-1-5225-8535-0 (2019)
7. Clark, M.: Mendacity and deception: uses and abuses of common ground. AAAI Fall Symposium, FS-11-02, AAAI Press (2011)
8. da Costa, N. C. A., French, S.: Belief, contradiction and the logic of self-deception. *American Philosophical Quarterly* 27:179–197 (1990)
9. Demos, R.: Lying to oneself. *Journal of Philosophy* 57:588–595 (1960)
10. Ekman, P.: *Why Kids Lie: How Parents Can Encourage Truthfulness*. Scribner (1989)
11. Ettinger, D., Jehiel, P.: A theory of deception. *American Economic Journal: Microeconomics* 2:1–20 (2010)

12. Firozabadi, B. S., Tan, Y. H., Lee, R. M.: Formal definitions of fraud. In: McNamara, P. and Prakken, H. eds. *Norms, Logics and Information Systems: New Studies in Deontic Logic and Computer Science*, 275–288. IOS Press (1999)
13. Frankfurt, H. G.: *On Bullshit*. Princeton University Press (2005)
14. Giunchiglia, E., Lee, J., Lifschitz, V., McCain, N., Turner, H.: Nonmonotonic causal theories. *Artificial Intelligence* 153: 49–104 (2004)
15. Hespanha, J. P., Ateskan, Y. S., Kizilocak, H.: Deception in non-cooperative games with partial information. In: *Proc. 2nd DARPA-JFACC Symp. Advances in Enterprise Control* (2000)
16. Isaac, A., Bridewell, W.: White lies on silver tongues: Why robots need to deceive (and how). In: *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*, Chapter 11, Oxford University Press (2017)
17. Jones, A. J. I.: On the logic of self-deception. *South American Journal of Logic* 1:387–400 (2015)
18. Mahon, J. E.: A definition of deceiving. *Journal of Applied Philosophy* 21:181–194 (2007)
19. McLaughlin B. P., Rorty, A. O. eds.: *Perspectives on Self-Deception*, University of California Press (1988)
20. McLaughlin, B. P.: Exploring the possibility of self-deception in belief. In: [19], 29–62 (1988)
21. O'Neill, B.: A formal system for understanding lies and deceit. Jerusalem Conference on Biblical Economics (2003)
22. Panisson, A. R., Sarkadi, S., Mcburney, P., Parsons, S., Bordini, R. H.: Lies, bullshit, and deception in agent-oriented programming languages In: *Proc. 20th Int'l Trust Workshop*, pp. 50–61 (2018)
23. Pörn, I.: On the nature of social order. In Fenstad, J. E. et al. eds. *Logic, Methodology, and Philosophy of Science*, VIII. Elsevier (1989)
24. Sakama, C., Caminada, M., Herzig, A.: A logical account of lying. In: *Proc. 12th European Conf. Logics in Artificial Intelligence, LNAI* 634:286–299, Springer (2010)
25. Sakama, C., Caminada, M.: The many faces of deception. *Thirty Years of Nonmonotonic Reasoning (NonMon@30)*, Lexington, KY, USA (2010)
26. Sakama, C., Caminada, M., Herzig, A.: A formal account of dishonesty. *Logic Journal of the IGPL* 23:259–294 (2015)
27. Sarkadi, S., Panisson, A. R., Bordini, R. H., Mcburney, P., Parsons, S., Chapman, M.: Modelling deception using theory of mind in multi-agent systems. *AI Communications* 32(3):1–16 (2019)
28. Shim, J., Arkin, R. C.: Biologically-inspired deceptive behavior for a robot. In: *Proc. 12th Int'l Conf. Simulation of Adaptive Behavior, LNCS* 7426, 401–411, Springer (2012)
29. Staab, E., Caminada, M.: On the profitability of incompetence. In: *Multi-Agent-Based Simulation XI, LNCS* 6532, 76–92, Springer (2011)
30. Trivers, R.: *The Folly of Fools: The Logic of Deceit and Self-Deception in Human Life*. Basic Books (2011)
31. Turing, A. M.: Computing machinery and intelligence. *Mind* 59:433–460 (1950)
32. van Ditmarsch, H., van Eijck, J., Sietsma, F., Wang, Y.: On the logic of lying. *Games, Actions and Social Software*, LNCS 7010, 41–72, Springer (2012)
33. van Ditmarsch, H.: Dynamics of lying. *Synthese* 191:745–777 (2014)
34. Vincent, J. M., Castelfranchi, C.: On the art of deception: how to lie while saying the truth. In: Parret, H., Sbisa, M., and Verschueren, J. eds. *Possibilities and Limitations of Pragmatics*, 749–777, J. Benjamins (1981)
35. Wagner, A. R., Arkin, R. C.: Acting deceptively: providing robots with the capacity for deception. *Journal of Social Robotics* 3:5–26 (2011)
36. Zlotkin, G., Rosenschein, J. S.: Incomplete information and deception in multi-agent negotiation. In: *Proc. 12th IJCAI*, pp. 225–231, Morgan Kaufmann (1991)