A Formal Account of Deception

Chiaki Sakama
Department of Computer and Communication Sciences
Wakayama University, Sakaedani, Wakayama 640-8510, Japan
Email: sakama@sys.wakayama-u.ac.jp

Abstract
This paper studies a formal account of deception using a dynamic epistemic logic. Three different types of deception are considered: deception by lying, deception by bluffing and deception by truth-telling, depending on whether a speaker believes what he/she says 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.

1 Introduction
The question “Can computers deceive humans?” has been of interest since Turing’s imitation game. 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 (Mahon 2007; Carson 2010). Deceiving is different from lying, in fact, there is deception without lying (Adler 1997; Vincent and Castelfranchi 1981). There is no consensus as to stating conditions for describing someone as self-deceived (da Costa and French 1990). 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.

To better understand what is deception, we need a formal account of deception. Understanding deception will help us to know effective ways of using deception to achieve a particular goal, and to consider the best ways in which one could avoid being deceived. Such considerations are particularly of interest in a game-theoretical perspective (Hespanha, et al. 2000; Ettinger and Jehiel 2010) and designing intelligent agents in multagent systems (Zlotkin and Rosenschein 1991; Staab and Caminada 2010). Moreover, a formal account of deception would be useful for developing deceptive artificial agents (Castelfranchi 2000). For instance, an intelligent personal assistant might deceive us to influence 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 2011) develops a lying machine and provides empirical evidence that the machine reliably deceives ordinary humans. Recent studies argue the utility of deception for developing autonomous robots (Wagner and Arkin 2011; Shim and Arkin 2012). In spite of the broad interest in this topic, however, relatively little study has been devoted to developing a formal theory of deception. 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 paper, we study a logical account of deception. We use the agent announcement logic of (van Ditmarsch 2013), that is in a family of dynamic epistemic logics. In this logic, an agent can make three different types of announcement, truth-telling (agent believes the truth of a sentence it announces), lying (agent believes the falsity of a sentence it announces), and bluffing (agent is uncertain about the truth of a sentence it announces). These announcements are formulated as dynamic modal operators which transform epistemic states of addressees. Using the logic, we formulate three different types of deception, deception by lying, deception by bluffing and deception by truth-telling, and distinguish them from attempted deception that may fail to deceive. We next argue 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 agent announcement logic. 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 summarizes the paper.

2 Agent Announcement Logic
This section reviews the agent announcement logic of (van Ditmarsch 2013) that we use in this paper. Let $P$ be a set of propositional variables and $A$ a finite set of agents. Then a sentence $\varphi$ in the language is defined as follows:

$$\varphi ::= p \mid \neg \varphi \mid (\varphi \land \psi) \mid B_a \varphi \mid [!a \varphi] \psi \mid [|a \varphi] \psi \mid [!|a \varphi] \psi$$

where $p \in P$, $a \in A$ and $\psi$ is a sentence. The logical connectives $\top, \bot, \lor, \land$ and $\equiv$ are introduced as abbreviations as usual. The set of all sentences in the language is denoted by
Throughout the paper, lower case letters \(a, b, c, \ldots\) represent agents in \(A\) and Greek letters \(\lambda, \varphi, \psi\) represent sentences in \(\Phi\) unless otherwise stated. A sentence \(B_a \varphi\) is read as “an agent \(a\) believes \(\varphi\).” An agent \(a\) may address a sentence \(\varphi\) no matter what, whether she believes what she says, believes the opposite, or is uncertain. A truthful agent \(a\) believes that a sentence \(\varphi\) is true (\(B_a \varphi\)) when it announces \(\varphi\). A lying agent \(a\) believes that a sentence \(\varphi\) is false (\(B_a \neg \varphi\)) when it announces \(\varphi\). A blurring agent \(a\) believes neither \(\varphi\) nor \(\neg \varphi\) (\(B_a (\varphi \lor \neg \varphi)\)) when it announces \(\varphi\). We write \(\models \varphi\) iff a sentence \(\varphi\) is a theorem of the logic. We write \(\models \varphi\) iff a sentence \(\varphi\) is a theorem of the logic.

The semantics is given by the Kripke structure. An epistemic model is a triple \(M = (S, R, V)\) where \(S\) is a non-empty set of possible worlds, \(R : A \to \mathcal{P}(S \times S)\) is an accessibility function, \(V : A \to \mathcal{P}(S)\) is a valuation function (\(V_p\) restricts the set of worlds where \(p\) is true). An epistemic state \((M, s)\) for \(s \in S\) satisfies sentences as follows:

\[
M, s \models p \iff s \in V_p \\
M, s \models \neg \varphi \iff M, s \not\models \varphi \\
M, s \models \varphi \land \psi \iff M, s \models \varphi \text{ and } M, s \models \psi \\
M, s \models \exists t \in S : R_a(s, t) \text{ implies } M, t \models \varphi \\
M, s \models [!a] \varphi \iff M, s \models B_a \varphi \text{ implies } M'_{\varphi}, s \models \psi \\
M, s \models [\lambda a] \varphi \iff M, s \models \neg (B_a \varphi \lor B_a \neg \varphi) \text{ implies } M'_{\varphi}, s \models \psi
\]

where \(M'_{\varphi}\) is as \(M\) except for the accessibility relation \(R'\) that is defined over \(S\) such that \(R'_{ab} := R_a \cap (S \times [B_a [\lambda a] \varphi]_{b})\) for \(b \in A (b \neq a)\) where \([\varphi]_b := \{ s \in S : M, s \models \varphi \}\). When an agent \(a\) addresses \(\varphi\), an addressee \(b\) takes \(a\)'s belief in \(\varphi\) to be the truth no matter what. Thus, the accessibility relation for an agent \(a\) remains the same. On the other hand, states where \(\varphi\) is believed by \(a\) remain accessible to an addressee \(b\), while states where \(\varphi\) is not believed by \(a\) are no longer accessible to \(b\). A sentence \(\varphi\) is true in a model \(M\) (written \(M \models \varphi\)) iff \(M, s \models \varphi\) for any \(s \in S\). \(\varphi\) is valid iff \(M \models \varphi\) for any model \(M\).

Throughout the paper, agents are assumed to have the K45 axioms for the belief operator \(B\):

\[
B_a(\varphi \lor \psi) \supset (B_a \varphi \lor B_a \psi) \quad (\text{distribution}) \\
\quad B_a \varphi \supset B_a B_a \varphi \quad (\text{positive introspection}) \\
\quad \neg B_a \varphi \supset B_a \neg B_a \varphi \quad (\text{negative introspection})
\]

and inference rules:

\[
\frac{\varphi \quad \varphi \lor \psi}{\psi} \quad (\text{modus ponens}) \\
\frac{\varphi}{B_a \varphi} \quad (\text{necessitation})
\]

By the distribution axiom and the necessitation rule, each agent believes that other agents follow the same logic. Thus, \(B_a (B_b (\varphi \lor \psi)) \supset B_a (B_b \varphi \lor B_b \psi)\), for instance. An agent who has the additional D-axiom

\[
\neg B_a \bot
\]

is called a KD45 agent. A theorem is a formula that is obtained from axiom instances via the inference rules. We write \(\models \varphi\) iff a sentence \(\varphi\) is a theorem of the logic.

In this paper, we consider communication between two agents. Let \(a\) be an agent who makes an announcement (called a speaker), and \(b\) an agent who is an addressee (called a hearer).\(^1\) The axioms for the belief consequences of agent communication are given as follows:

\begin{align*}
\text{(A1)} & \quad [i_a \varphi]B_a \psi \quad \equiv \quad B_a \varphi \supset B_a [i_a \varphi] \psi \\
\text{(A2)} & \quad [i_a \varphi]B_a \psi \quad \equiv \quad B_a \varphi \supset B_a [i_a \varphi] \psi \\
\text{(A3)} & \quad [i_a \varphi]B_a \psi \quad \equiv \quad \neg (B_a \varphi \lor B_a \neg \varphi) \supset B_a [i_a \varphi] \psi \\
\text{(A4)} & \quad [i_a \varphi]B_a \psi \quad \equiv \quad B_a \varphi \supset B_a [i_a \varphi] \psi \\
\text{(A5)} & \quad [i_a \varphi]B_a \psi \quad \equiv \quad B_a \varphi \supset B_a [i_a \varphi] \psi \\
\text{(A6)} & \quad [i_a \varphi]B_a \psi \quad \equiv \quad \neg (B_a \varphi \lor B_a \neg \varphi) \supset B_a [i_a \varphi] \psi
\end{align*}

By \((A1)\)–\((A6)\), the speaker \(a\) recognizes his/her act of truth-telling, lying, or bluffing. By \((A4)\)–\((A6)\), on the other hand, the hearer \(b\) believes that the speaker always makes a truthful announcement. \((A1)\)–\((A6)\) hold by replacing \(B_a \psi\) (resp. \(B_a \psi\)) with \(\neg B_a \psi\) (resp. \(\neg B_a \psi\)) on the left of \(\equiv\) and replacing \(B_a \psi\) (resp. \(B_a \psi\)) with \(\neg B_a \psi\) (resp. \(\neg B_a \psi\)) on the right of \(\equiv\). This axiomatization characterizes a credulous addressee in the sense that a hearer always believes that a speaker is sincere (i.e., a hearer believes that a speaker believes the announcement). On the other hand, the agent announcement logic can also characterize skeptical agents who can distinguish believable announcement (\(\neg B_a \neg B_a \varphi\)) from unbelievable one (\(\neg B_a \neg B_a \varphi\)). In this case, a skeptical addressee only incorporates new information if the addressee considers it possible that a speaker believes that the announced sentence is true. The logic also has a variant for belief revising agents in which an agent believes everything that it is told by consistently revising its current possibly conflicting beliefs. Belief revising agents as well as skeptical agents have axiomatizations different from those for credulous addressees. In this paper, we consider credulous addressees unless stated otherwise. This is because deception usually happens when an addressee credulously believes that a speaker is truthful. The axiomatic system for credulous agents is sound and complete for K45, while it is incomplete for KD45. By contrast, the axiomatic system for skeptical agents is also complete for KD45. We provide some valid formulas that will be used in this paper.

**Proposition 2.1** (van Ditmarsch 2013) Let \(p, q \in P\) be propositional variables.

(i) \(B_a [i_a \varphi] q \equiv B_a [i_a \varphi] q \equiv B_a [i_a \varphi] q \equiv B_a q\).

(ii) \([i_a \varphi] B_a [i_a \varphi] p \equiv [i_a \varphi] B_a [i_a \varphi] p \equiv [i_a \varphi] B_a [i_a \varphi] p \equiv T\).

**Proposition 2.1** does not hold for arbitrary sentences in general.

\(^1\)Although we call ‘\(a\)’ a speaker and ‘\(b\)’ a hearer, we consider any kind of communication by which an agent \(a\) informs an agent \(b\) of a sentence, be it linguistically or not.
3 Deception in Agent Announcement Logic

3.1 Deception by Lying

Deception is different from lying. Carson (2010) 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” (Carson 2010, p. 55). He then illustrates the relationship between lying, deception, and attempted deception as in Figure 1.

Our primary interest in this section is to formulate “lies that deceive”.$^2$ Deception by lying is formulated in the logic of agent announcement as follows.

Definition 3.1 (deception by lying) Let $a$ and $b$ be two agents and $\varphi, \psi \in \Phi$. Then deception by lying (DBL) is defined as

$$\text{DBL}_{ab}(\varphi, \psi) \overset{\text{def}}{=} B_a \neg \varphi \land \neg \psi \land B_b (\varphi \supset \psi) \land (B_a \varphi) B_b \psi \lor (B_a \neg \varphi) B_b \neg \psi).$$

By definition, deception by lying is such that (i) a speaker $a$ believes the falsity of $\varphi$, (ii) a hearer $b$ believes the implication $\varphi \supset \psi$, and (iii) a hearer $b$ believes the true sentence $\psi$ after $a$’s lying announcement $\varphi$, or $b$ disbelieves the true sentence $\neg \psi$ after $a$’s lying announcement $\varphi$. In particular,

$$\text{DBL}_{ab}(\varphi, \psi) = B_a \neg \varphi \land \neg \psi \land (B_a \varphi) B_b \psi \lor (B_a \neg \varphi) B_b \neg \psi).$$

Example 3.1 Suppose a salesperson who lies that an investment promises a high return with little risk. The lie makes a customer believe that the investment is worth buying. The situation is represented by (1) with $a =$ salesperson, $b =$ customer, and $\varphi =$ “an investment promises high return” and $\psi =$ “the investment is worth buying.” A salesperson makes the announcement $\varphi$ which results in a customer’s believing $\psi$ (or disbelieving $\neg \psi$).

Note that (1) does not address whether a hearer believes $\psi$ or $\neg \psi$ before the announcement. DBL happens whenever a hearer believes a false sentence $\psi$ (or disbelieves a true sentence $\neg \psi$) as a result of the lying announcement by a speaker. Also note that a speaker believes the falsity of $\varphi$ but the actual falsity of $\varphi$ is not necessarily requested. So if a speaker $a$ makes a believed-false statement $\varphi$ which is in fact true, then the announcement $[i_a \varphi]$ is lying. By contrast, the sentence $\psi$ is actually false. So if a speaker’s lying announcement contributes to a hearer’s believing a true sentence (or disbelieving a false sentence), then it is not deception by lying.

Example 3.2 A student, Bob, who believes that there will be no exam in tomorrow’s math-class, says 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. Next day, it turns that there is an exam in the math-class. In this case, Bob lies to Mike but Bob does not deceive Mike (and Mike does not believe that Bob lies to him).

Note also that the implication $\varphi \supset \psi$ is believed by a hearer in (1), but is not necessarily believed by a speaker. So if a speaker lies $\varphi$ and it brings about a hearer’s believing a false sentence $\psi$, then it is considered deception by lying—indeed of the fact that the speaker expected the effect or not. For instance, if a political candidate makes a lying announcement and a person, Mary, who believes the announcement, considers him/her a good candidate and votes for him/her, then it is DBL even if the candidate does not expect Mary’s believing him/her a good candidate.

In the agent announcement logic, if a speaker $a$ makes an announcement $\varphi$ but a hearer $b$ already believes to the contrary, the hearer believes that the speaker is mistaken, namely $B_b (\neg \varphi \land \neg \varphi)$. If the hearer is a KD45 agent, neither $B_b \varphi$ nor $\neg B_b \neg \varphi$ holds after the lying announcement $[i_a \varphi]$, so that an attempted lie $\varphi$ fails to deceive.

Proposition 3.1 Let $b$ be a KD45 agent. For any $\varphi, \psi \in \Phi$,

$$\vdash B_b \neg \psi \land \text{DBL}_{ab}(\varphi, \psi) \supset \bot.$$  

Proof $B_b \neg \psi \land \text{DBL}_{ab}(\varphi, \psi)$ implies $B_b \neg \psi \land B_b \psi \lor B_b \neg \psi$ after the announcement $[i_a \varphi]$. Since $B_b \neg \psi \land B_b \psi \lor B_b \neg \psi \equiv B_b \bot$, the result holds. 

By Proposition 3.1, it holds that $B_b \neg \varphi \land \text{DBL}_{ab}(\varphi, \psi) \supset \bot$. Suppose that a hearer is a KD45 agent who believes that a speaker is also a KD45 agent. In this case, if a hearer believes that a speaker is lying, then DBL fails.

Proposition 3.2 Let $b$ be a KD45 agent who believes that another agent $a$ is KD45. For any $p \in P$ and $\psi \in \Phi$,

$$\vdash B_b B_a \neg p \land \text{DBL}_{ab}(p, \psi) \supset \bot.$$  

Proof It holds that $\vdash [i_a p] B_a B_b p \equiv \top$ (Proposition 2.1). Since $B_b B_a \neg p \land B_b B_b p \equiv B_b B_b \bot$ after the announcement $[i_a p]$, the result holds. 

DBL on the contradictory sentence fails.

Proposition 3.3 Let $b$ be a KD45 agent who believes that another agent $a$ is KD45. For any $\psi \in \Phi$,

$$\vdash \text{DBL}_{ab}(\bot, \psi) \supset \bot.$$  

Proof It holds by putting $p = \bot$ in Proposition 3.2. 

If a hearer is a KD45 agent who believes that a speaker is bluffing, then DBL fails.

**Proposition 3.4** Let $b$ be a KD45 agent. For any $p \in P$ and $\psi \in \Phi$, 

$$\vdash B_b(\neg B_a p \land \neg B_a \neg p) \land DBL_{ab}(p, \psi) \supset \bot.$$ 

**Proof.** Since $B_b(\neg B_a p \land \neg B_a \neg p) \land B_b p \equiv B_b \bot$ after the announcement $[!_a p]$, the result holds.

Propositions 3.1–3.4 characterize different situations where “lies that attempt but fail to deceive”. In each case, deception fails if a hearer has consistent belief. Finally, where “lies that attempt but fail to deceive”. In each case, 

$\lambda$ a 

**Proposition 3.5** Let $b$ be a KD45 agent. For any $\varphi , \psi \in \Phi$, 

$$\vdash B_b \neg \psi \land DBL_{ab}(\varphi, \psi) \supset \bot.$$ 

**Proof.** Similar to the proof of Proposition 3.1.

### 3.2 Deception by Blufﬁng

When a speaker announces a sentence $\varphi$ while he/she is uncertain about the truthfulness of $\varphi$ and it makes a hearer believe a false fact $\psi$, it is called deception by bluffing.3

**Example 3.3** Bob, who does not know whether there will be an exam in tomorrow’s math-class, says his friend Mike that there will be no exam in tomorrow’s math-class. Mike believes Bob’s information. Next day, it turns that there is an exam in the math-class. In this case, Bob deceives Mike by bluffing.

Deception by bluffing is formally deﬁned as follows.

**Deﬁnition 3.2 (deception by bluffing)** Let $a$ and $b$ be two agents and $\varphi , \psi \in \Phi$. Then deception by bluffing (DBB) is deﬁned as

$$DBB_{ab}(\varphi, \psi) \overset{def}{=} \neg (B_a \varphi \lor B_a \neg \varphi) \land \neg \psi \land B_b(\varphi \supset \psi) \land ([!_a \varphi]B_b \psi \lor [!_a \varphi]B_b \neg \psi). \quad (2)$$

DBB is different from DBL in two aspects. First, $[!_a \varphi]$ in (1) is replaced by $[!_a \varphi]$ in (2). Second, $B_a \neg \varphi$ in (1) is replaced by $\neg(B_a \varphi \lor B_a \neg \varphi)$ in (2). On the other hand, $\psi$ is false in both (1) and (2).

Like DBL, DBB fails if a hearer is a KD45 agent who believes $\neg \psi$.

**Proposition 3.6** Let $b$ be a KD45 agent who believes that another agent $a$ is KD45. For any $p \in P$ and $\psi \in \Phi$, 

$$\vdash B_a B_a \neg p \land DBB_{ab}(p, \psi) \supset \bot.$$ 

**Proof.** Similar to the proof of Proposition 3.2.

If a hearer is a KD45 agent who believes that a speaker is bluffing, then DBB fails.

**Proposition 3.7** Let $b$ be a KD45 agent. For any $p \in P$ and $\psi \in \Phi$, 

$$\vdash B_b(\neg B_a p \land \neg B_a \neg p) \land DBB_{ab}(p, \psi) \supset \bot.$$ 

**Proof.** It holds that $\vdash [!_a p]B_b B_a p \equiv \top$ (Proposition 2.1). Since $B_b(\neg B_a p \land \neg B_a \neg p) \land B_b p \equiv B_b \bot$ after the announcement $[!_a p]$, the result holds.

The difference between Propositions 3.6 and 3.7 is subtle. Attempted DBB fails if a KD45 hearer believes that a KD45 speaker is lying. By contrast, attempted DBB fails if a KD45 hearer believes that a speaker is bluffing. In the latter case, it is not required that a hearer believes that a speaker is KD45.

DBB on the contradictory sentence fails.

**Proposition 3.8** For any $\psi \in \Phi$, 

$$\vdash DBB_{ab}(\bot, \psi) \supset \bot.$$ 

**Proof.** $DBB_{ab}(\bot, \psi)$ implies $\neg B_a \top$, which implies $\bot$. 

It is impossible to make DBB on one’s own belief.

**Proposition 3.9** For any $\varphi , \psi \in \Phi$, 

$$\vdash DBB_{ab}(B_a \varphi, \psi) \lor DBB_{ab}(\neg B_a \varphi, \psi) \supset \bot.$$ 

**Proof.** Both $DBB_{ab}(B_a \varphi, \psi)$ and $DBB_{ab}(\neg B_a \varphi, \psi)$ imply $\neg B_a B_a \varphi \lor \neg B_a \neg B_a \varphi$. Since $\vdash \neg B_a B_a \varphi \supset \neg B_a \varphi$ by the positive introspection of K45, $\neg B_a B_a \varphi$ implies $\neg B_a \varphi$, which implies $\neg B_a \neg B_a \varphi$ by the negative introspection of K45. This contradicts $\neg B_a \neg B_a \varphi$.

### 3.3 Deception by Truth-Telling

One can deceive others by telling truthful sentences.

**Example 3.4** Suppose that John, who is interested in Mary, invites her to dinner on Christmas day. Mary, who has no interest in John, says that she has an appointment with another man. John then understands that Mary has a boyfriend, but Mary has an appointment with her father. In this scenario, Mary tells the truth, while John believes the false fact that she will have a Christmas dinner with a boyfriend.

The above example illustrates another instance of “deception without lying”. We call this type “deception by truth-telling” that is formally defined as follows.

**Deﬁnition 3.3 (deception by truth-telling)** Let $a$ and $b$ be two agents and $\varphi , \psi \in \Phi$. Then deception by truth-telling (DBT) is deﬁned as

$$DBT_{ab}(\varphi, \psi) \overset{def}{=} B_a \varphi \land \neg \psi \land B_b(\varphi \supset \psi) \land ([!_a \varphi]B_b \psi \lor [!_a \varphi]B_b \neg \psi). \quad (3)$$
Different from DBL, in DBT a speaker $a$’s truthful announcement $\varphi$ makes a hearer $b$ believe a false sentence $\psi$, or $a$’s truthful announcement $\varphi$ makes $b$ disbelieve a true sentence $\neg \psi$. In (3) a speaker $a$ believes that $\varphi$ is true, but the actual truth of $\varphi$ is not necessarily requested. In DBT, a speaker successfully deceives a hearer without having to resort to telling direct lies or bluffs. A hearer mistakenly concludes a false fact and gets duped by himself. This type of deception is also argued in (Vincent and Castelfranchi 1981; Adler 1997). By definition, DBL, DBB and DBT are mutually exclusive.

Like DBL and DBB, $\text{DBT}_{ab}(\varphi, \psi)$ fails if a hearer is a KD45 agent who believes $\neg \psi$.

**Proposition 3.10** Let $b$ be a KD45 agent. For any $\varphi, \psi \in \Phi$, 
\[ \vdash B_b \neg \psi \land \text{DBT}_{ab}(\varphi, \psi) \supset \bot. \]

**Proof.** Similar to the proof of Proposition 3.1. \( \square \)

Like DBL and DBB, if a KD45 hearer believes that a KD45 speaker is lying, then DBT fails.

**Proposition 3.11** Let $b$ be a KD45 agent who believes that another agent $a$ is KD45. For any $p \in P$ and $\psi \in \Phi$, 
\[ \vdash B_b B_a \neg p \land \text{DBT}_{ab}(p, \psi) \supset \bot. \]

**Proof.** Similar to the proof of Proposition 3.2. \( \square \)

In Example 3.4, if John believes that Mary has consistent belief and she is lying, then Mary’s DBT fails. DBT also fails if a KD45 hearer believes that a speaker is bluffing.

**Proposition 3.12** Let $b$ be a KD45 agent. For any $p \in P$ and $\psi \in \Phi$, 
\[ \vdash B_b (\neg B_a p \land \neg B_a \neg p) \land \text{DBT}_{ab}(p, \psi) \supset \bot. \]

**Proof.** Similar to the proof of Proposition 3.7. \( \square \)

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 (Carson 2010). It is also called deception by omission, which is contrasted with deception by commission that involves an act of providing information (Chisholm and Feehan 1977). We capture deception by omission as a result of no informative announcement, and define it as DBT with announcing a valid sentence.

**Definition 3.4 (deception by omission)** Let $a$ and $b$ be two agents and $\varphi, \psi \in \Phi$. Then deception by omission (DBO) is defined as
\[ \text{DBO}_{ab}(\psi) \overset{\text{def}}{=} \text{DBT}_{ab}(\top, \psi) \land B_a \neg \psi. \]  

By definition, deception by omission happens when a speaker believes a false fact while makes no announcement of it. As a result, a hearer believes the false fact.

**Proposition 3.13** For any $\psi \in \Phi$, 
\[ \vdash \text{DBO}_{ab}(\psi) \equiv \neg \psi \land B_b \psi \land B_a \neg \psi. \]

**Proof.** $\text{DBT}_{ab}(\top, \psi) \equiv \neg \psi \land B_b \psi \land (|a \top| B_b \psi \lor \neg |a \top| B_b \neg \psi)$. Then the result holds by the fact $|a \top| B_b \psi \equiv B_b \psi$. \( \square \)

Proposition 3.13 says that if deception by omission happens, then a hearer (initially) believes a false sentence $\psi$ while a speaker believes $\neg \psi$.

## 4 Various Aspects of Deception

### 4.1 Intentional Deception

Sometimes deception is distinguished between intentional deception and unintentional one (Chisholm and Feehan 1977). DBL, DBB and DBT in Section 3 represent unintentional deception, that is, a speaker does not necessarily intend to deceive a hearer. In $\text{DBL}_{ab}(\varphi, \psi)$, a speaker $a$ lies a believed-false sentence $\varphi$ to a hearer $b$, while the speaker does not necessarily believe that the announcement will result in the hearer’s believing another false sentence $\psi$. In $\text{DBT}_{ab}(\varphi, \psi)$, a speaker $a$ tells a believed-true sentence $\varphi$ to a hearer $b$, so that the speaker might not feel guilty even if the announcement leads a hearer to believe a false sentence $\psi$ as a result. To formulate a speaker’s intention to deceive, definitions of DBL, DBB and DBT are respectively modified as follows.

**Definition 4.1 (intentional deception)** Let $a$ and $b$ be two agents and $\varphi, \psi \in \Phi$. Then intentional deception by lying ($\text{I-DBL}$), intentional deception by bluffing ($\text{I-DBB}$) and intentional deception by truth-telling ($\text{I-DBT}$) are respectively defined as:
\[ \text{I-DBL}_{ab}(\varphi, \psi) \overset{\text{def}}{=} B_a \neg \psi \land B_b (\varphi \supset \psi) \land \text{DBL}_{ab}(\varphi, \psi). \]
\[ \text{I-DBB}_{ab}(\varphi, \psi) \overset{\text{def}}{=} B_a \neg \psi \land B_b (\varphi \supset \psi) \land \text{DBB}_{ab}(\varphi, \psi). \]
\[ \text{I-DBT}_{ab}(\varphi, \psi) \overset{\text{def}}{=} B_a \psi \land B_b (\varphi \supset \psi) \land \text{DBT}_{ab}(\varphi, \psi). \]

In particular, intentional deception by omission is defined as
\[ \text{I-DBO}_{ab}(\psi) \overset{\text{def}}{=} \text{I-DBT}_{ab}(\top, \psi) \land B_a \neg \psi \equiv \text{I-DBT}_{ab}(\top, \psi) \equiv \neg \psi \land B_b \psi \land B_a \neg \psi \land B_a B_b \psi. \]

Hence, we handle $\text{I-DBO}_{ab}(\psi)$ as a case of $\text{I-DBT}_{ab}(\varphi, \psi)$ with $\varphi = \top$.

In Definition 4.1, a speaker $a$ believes the falsity of the sentence $\psi$ and also believes that a hearer believes the implication $\varphi \supset \psi$. With this additional condition, if the speaker makes a lying announcement $[|a \top| \varphi]$ expecting that the announcement will cause the hearer’s believing the false sentence $\psi$, then it is intentional deception by lying. Similar accounts are made for definitions of I-DBB and I-DBT. Note that we do not introduce an additional modal operator 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 I-DBL (resp. I-DBB or I-DBT) implies DBL (resp. DBB or DBT), properties addressed in Section 3 hold for these intentional deception as well. In what follows, (I-)DBL (resp. (I-)DBB or (I-)DBT) means intentional or unintentional DBL (resp. DBB or DBT).

If a speaker deceives a hearer while believing his/her deceptive act, then it is intentional deception.

**Proposition 4.1** For any $\varphi, \psi \in \Phi$, 
\[ \vdash \text{DBX}_{ab}(\varphi, \psi) \land B_a (\text{DBX}_{ab}(\varphi, \psi)) \supset \text{I-DBX}_{ab}(\varphi, \psi) \]
where “X” means one of “L”, “B”, and “T”.
Proposition 4.2 For any \( \varphi \in \Phi \),
- \( \vdash \text{s-DBL}_{ab}(\varphi, \varphi) \equiv \text{DBL}_{ab}(\varphi, \varphi) \).
- \( \vdash \text{s-DBL}_{ab}(\varphi, \varphi) \supset \top \).
- \( \vdash \text{s-DBT}_{ab}(\varphi, \varphi) \supset \bot \) for any KD45 agent \( a \).

Proof. Since \( B_a(\text{DBX}_{ab}(\varphi, \psi)) \) implies \( B_a \neg \psi \land B_a B_\varphi(\varphi \supset \psi) \), the result holds by definition.

Definition 4.2 (indirect deception) Let \( a, b \) and \( c \) be three agents and \( \lambda, \varphi, \psi \in \Phi \). Then indirect deception by lying (IN-DBL), indirect deception by bluffing (IN-DBB), and indirect deception by truth-telling (IN-DBT) are respectively defined as:

\[
\text{IN-DBL}_{ac}(\varphi, \lambda) \equiv (\text{IN-DBL}_{ab}(\varphi, \psi) \land \text{DBT}_{bc}(\psi, \lambda)),
\]

\[
\text{IN-DBB}_{ac}(\varphi, \lambda) \equiv (\text{IN-DBB}_{ab}(\varphi, \psi) \land \text{DBT}_{bc}(\psi, \lambda)),
\]

\[
\text{IN-DBT}_{ac}(\varphi, \lambda) \equiv (\text{IN-DBT}_{ab}(\varphi, \psi) \land \text{DBT}_{bc}(\psi, \lambda)).
\]

In \( \text{IN-DBL}_{ac}(\varphi, \lambda) \), \( a \)'s lying announcement on a sentence \( \varphi \) results in \( b \)'s believing a false sentence \( \psi \), and \( b \)'s truthful announcement on a sentence \( \psi \) results in \( c \)'s believing a false sentence \( \lambda \). \( \text{IN-DBL}_{ac}(\varphi, \lambda) \) and \( \text{IN-DBT}_{ac}(\varphi, \lambda) \) represent similar situations. 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 announcement. If this is not the case, for instance,

\[
(\text{IN-DBL}_{ab}(\varphi, \psi) \land \text{DBL}_{bc}(\neg \psi, \lambda))
\]

then we do not consider that \( a \) indirectly deceives \( c \). In this case, \( a \)'s lying announcement contributes to \( b \)'s believing a false fact \( \psi \), but it would not contribute to \( c \)'s believing a false fact \( \lambda \) because \( b \) makes a lying announcement on the sentence \( \neg \psi \) (which is in fact true). In Example 4.1, if John attempts to surprise his wife and lies her that a medical test detects a brain tumor, then a doctor does not indirectly deceive John's wife. Indirect deception could be chained like

\[
(\text{IN-DBL}_{ab}(\varphi, \psi_1) \land \text{DBT}_{bc}(\psi_1, \psi_2) \land \text{DBT}_{cd}(\psi_2, \psi_3) \land \cdots)
\]

where "\( X \)" is one of L, B and T in general.

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 (Demos 1960; McLaughlin and Rorty 1988; da Costa and French 1990; Trivers 2011). Self-deception involves a person holding contradictory beliefs \( (B_a(\varphi \land \neg \varphi)) \), or believing and disbelieving the same sentence at the same time \( (B_a \varphi \land \neg B_a \varphi) \). In each case, it violates the classical principle of consistency that rational agents are assumed to follow. In this section, we characterize self-deception in our formulation.

Proposition 4.3 Let \( a \) be a KD45 agent. For any \( p \in P \),

\[
\vdash \text{DBL}_{aa}(p, p) \supset B_a \bot \land \neg p .
\]

Proof. By definition,

\[
\text{DBL}_{aa}(p, p) = B_a \neg p \land \neg p \land (\{i_a p\} B_a p \lor \{i_a p\} \neg B_a p).
\]
Using the axiom (A2), it becomes

\[ B_a \neg p \land \neg p \land ((B_a \neg p \supset B_{a[a]}p) \lor (B_a \neg p \supset \neg B_{a[a]}p)) \]
\[ \equiv B_a \neg p \land \neg p \land (B_{a[a]}p \lor \neg B_{a[a]}p) \]
\[ \equiv B_a \neg p \land \neg p \land (B_a \lor \neg B_a) \quad \text{(Proposition 2.1)} \]
\[ \equiv B_a \bot \land \neg p. \quad \Box \]

Proposition 4.3 shows that if an agent a is KD but not KD45, DBB_{aa}(p, p) involves a mental state of an agent who has contradictory belief wrt a false fact p. By contrast, DBL_{aa}(\varphi, \psi) implies

\[ B_a \neg \varphi \land \neg \psi \land B_a (\varphi \supset \psi) \land (B_a \psi \lor \neg B_a \neg \psi) \quad (6) \]

that is consistent for a sentence \( \psi \not\equiv \varphi \). (6) represents counterfactual inference involved in self-deception, that is, a speaker believes the falsity of \( \varphi \) while believes the effect \( \psi \) or its possibility that would be obtained if \( \varphi \) were the case. Such kind of reasoning is possible by KD45 agents.

Unlike DBL, self-deception by DBB or DBT is possible on an announced sentence even by a KD45 agent.

**Proposition 4.4** Let a be a KD45 agent. For any \( p \in P \),

- \( \vdash \text{DBB}_{aa}(p, p) \supset \neg (B_a p \lor B_a \neg p) \land \neg p. \)
- \( \vdash \text{DBT}_{aa}(p, p) \supset B_a p \land \neg p. \)

**Proof.**

- \( \text{DBB}_{aa}(p, p) = \neg (B_a p \lor B_a \neg p) \land \neg p \land (\neg [1_a]p) \lor B_{a[a]}p \lor \neg B_{a[a]}p) \]
  \[ \equiv \neg (B_a p \lor B_a \neg p) \land \neg p \land (B_a p \lor \neg B_a p) \quad \text{(by (A3) and Proposition 2.1)} \]
  \[ \equiv \neg (B_a p \lor B_a \neg p) \land \neg p. \]
- \( \text{DBT}_{aa}(p, p) = B_a p \land \neg p \land \neg [1_a]p \lor B_{a[a]}p \lor \neg B_{a[a]}p) \]
  \[ \equiv B_a p \land \neg p \land (B_a p \lor \neg B_a p) \quad \text{(by (A4) and Proposition 2.1)} \]
  \[ \equiv B_a p \land \neg p. \quad \Box \]

By Proposition 4.4, neither DBB_{aa}(p, p) nor DBT_{aa}(p, p) implies \( \bot \) nor \( B_a \bot \). Self-deception by DBB happens for an agent who has no belief on a false fact, while self-deception by DBT happens for an agent who has a false belief.

Next we consider self-deception accompanied by intention. Any KD45 agent cannot intentionally deceive oneself by DBL, DBB or DBT.

**Proposition 4.5** Let a be a KD45 agents. For any \( \varphi, \psi \in \Phi \),

- \( \vdash \text{I-DBL}_{aa}(\varphi, \psi) \supset \bot. \)
- \( \vdash \text{I-DBB}_{aa}(\varphi, \psi) \supset \bot. \)
- \( \vdash \text{I-DBT}_{aa}(\varphi, \psi) \supset \bot. \)

**Proof.** Intentional deception assumes \( B_a \neg \psi \), while \( B_a \psi \lor \neg B_a \neg \psi \) holds after the announcement. Hence, the result holds. \( \Box \)

Note that DBL_{aa}(\varphi, \psi), DBB_{aa}(\varphi, \psi), and DBT_{aa}(\varphi, \psi) are all consistent when deception is unintentional ((6) and Proposition 4.4). On the other hand, Proposition 4.5 states that all of them turn inconsistent if intention is involved in self-deception. This would explain that self-deception is possible by agents with consistent belief only when it is done unconsciously.\(^7\)

Finally, one can indirectly deceive oneself. The following scenario is a modification of the “appointment example” of (McLaughlin 1988, p. 31).\(^8\)

**Example 4.2** 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.

The above scenario is represented by IN-DBL as

\[ \text{IN-DBL}_{aa}(\varphi, \psi) = \text{I-DBL}_{aa}(\varphi, \psi) \land \text{DBT}_{ba}(\varphi, \psi) \]

where \( a = \text{Mary}, b = \text{online assistant}, \text{and } \varphi = \text{“Meeting on April 1st”}. \) As such, indirect self-deception is represented by putting \( a = c \) in Definition 4.2. Recall that self-deception on a lying sentence is impossible for KD45 agents (Proposition 4.3). Interestingly, however, KD45 agents can deceive oneself by using indirect DBL even on a lying sentence. Generally, indirect self-deception is represented by

\[ \text{IN-DBX}_{aa}(\varphi, \lambda) = (\text{I-DBX}_{ab}(\varphi, \psi) \land \text{DBT}_{ba}(\psi, \lambda) \]

where “X” is either L, B or T. A KD45 agent can act on indirect self-deception in general.

**Proposition 4.6** Let a be a KD45 agent. There are sentences \( \varphi, \lambda \in \Phi \) such that

\[ \not\exists \text{IN-DBX}_{aa}(\varphi, \lambda) \supset \bot. \]

Proposition 4.6 presents that self-deception does not always involve contradiction if it is done indirectly.

## 5 Related Work

Van Ditmarsch et al. (2012; 2013) 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. Our current study is intended to formulate different types of deception based on the logic and investigate their formal properties. In particular, truthful announcements are not used for misleading hearers in (van Ditmarsch, et al. 2012; van Ditmarsch 2013).

There are some studies attempting to formulate deception using modal logic. Firozabadi et al. (1999) formulate fraud and 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: $\neg B_a \varphi \land E_b \varphi$ or $B_a \neg \varphi \land E_b \varphi$, where $E_a \psi$ means “an agent $a$ brings about that $\psi$”. On the other hand, cases that an agent who does not succeed in his/her attempt to deceive another agent are formally represented as: $\neg B_a \varphi \land H_b \varphi$ or $B_a \neg \varphi \land H_b \varphi$, where $H_a \psi$ means “an agent $a$ attempts to bring about $\psi$, not necessarily successful”. Their formulation represents the result of deceptive action but does not represent which type of an announcement brings about false belief on a hearer.

O’Neill (2003) 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: $\text{Dec}_{ab} \varphi := I_a B_b \varphi \land B_a \neg \varphi \land B_b \varphi$. Attempted deception is defined by removing the conjunct $B_b \varphi$ in $\text{Dec}_{ab} \varphi$. $\text{Dec}_{ab} \varphi$ does not represent that $b$ comes to have a false belief $\varphi$ as a result of an action by $a$. Thus, $a$ deceives $b$ when $b$ believes $\varphi$ without any action of $a$. The problem comes from the fact that their logic does not have a mechanism of representing an action and its effect.

Baltag and Smets (2008) introduce a logic of conditional doxastic actions. According to their formulation, the action of public successful lying is characterized by an action plausibility model involving two actions $\text{Lie}_a(\varphi)$ and $\text{True}_a(\varphi)$. The former represents an action in which an agent $a$ publicly lies that she knows $\varphi$ while in fact she does not know it. The latter represents an action in which $a$ makes a public truthful announcement that she knows $\varphi$. They have preconditions $\neg K_a \varphi$ and $K_a \varphi$, respectively. If a hearer already knows that $\varphi$ is false, however, the action $\text{Lie}_a(\varphi)$ does not succeed. Such a condition is formulated as an action’s contextual appearance. Note that the precondition $\neg K_a \varphi$ of $\text{Lie}_a(\varphi)$ represents the ignorance of $\varphi$ and is different from the one used in the agent announcement logic. They argue deception by lying but do not consider deception that may happen without lying.

Jones (2013) analyzes self-deception in the form of the Montaigne-family (e.g. $\neg B_a \varphi \land B_a B_a \varphi$) and concludes that they cannot be represented in the logic of belief KD45 in a consistent manner. da Costa and French (1990) formulates the inconsistent aspects of self-deception using paraconsistent doxastic logic. Those studies, as well as most philosophical studies, view self-deception as having contradictory or inconsistent belief and argue how to resolve it. It captures an important aspect of self-deception, while we argue in Section 4.3 that some sorts of self-deception do not introduce contradiction (cf. (6), Propositions 4.4 and 4.6). In particular, we show that self-deception can be consistent and successful for a KD45 agent by means of indirect deception.

Sakama et al. (2010) formulate deception in which a speaker makes a truthful statement expecting that a hearer will misinterpret it to draw a wrong conclusion. It is similar to deception by truth-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 (Sakama, et al. 2010) corresponds to attempted deception in this paper. Sakama and Caminada (2010) provide a logical account of different categories of deception that were given by (Chisholm and Feehan 1977). They use a modal logic of action and belief developed by (Pörn 1989), 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. (2015) distinguish deception by lying, deception by bullshitting, deception by withholding information and deception by truth-telling using causal relation, while they do not investigate formal properties.

6 Conclusion

Deception is an important topic that has been studied in various disciplines and is now of interest in different areas of informatics. This paper introduced formal representation of deception using a dynamic epistemic logic that can express both an act of deceiving and its effect on hearers’ belief. The current study aims to turn conceptually defined notions in philosophy into a formally defined semantic framework in computational logic. Our formal account of deception tries to explain what is deception and what is not. It provides conditions under which deception is considered intentional and represents various aspects of self-deception. One of the interesting findings is that a KD45 agent is capable of self-deception via indirect deception, which, to the best of our knowledge, has never been argued in the literature. The abstract framework proposed in this paper is simple and would not capture all aspects of deception. Nevertheless, it can characterize various features of deception in human society, and serves as a preliminary investigation to a formal theory of deception.

Acknowledgments

We thank Hans van Ditmarsch for useful discussion.

References


Sakama, C. and Caminada, M. 2010. The many faces of deception. Thirty Years of Nonmonotonic Reasoning (NonMon@30), Lexington, KY, USA.


