Deceptive and Counter-Deceptive Machines
Papers from the AAAI 2015 Fall Symposium

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 study focuses on the question: "What are the computational formalisms at the heart of deceptive and counter-deceptive machines?" We formulate 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 we consider various situations 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. We investigate formal properties of different sorts of deception.

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 (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 multiagent 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. (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* φ in the language is defined as follows:

$$\varphi ::= p \mid \neg\varphi \mid (\varphi \wedge \psi) \mid B_a\varphi \mid [!_a\varphi]\psi \mid [\mathsf{i}_a\varphi]\psi \mid [!\mathsf{i}_a\varphi]\psi$$

where $p \in P$, $a \in A$ and ψ is a sentence. The logical connectives $\top$, $\bot$, $\vee$, $\supset$ and $\equiv$ are introduced as abbreviations as usual. The set of all sentences in the language is denoted by

Φ. Throughout the paper, lower case letters $a, b, c, \ldots$ represent agents in A and Greek letters λ, φ, ψ represent sentences in Φ unless otherwise stated. A sentence $B_a\varphi$ is read as "an agent a believes φ." The operators $[!_a\varphi]$, $[!_{¡a}\varphi]$ and $[!_{¡a}\varphi]$ respectively stand for "*a truthfully announces* φ", "a is *lying* that φ", and "a is *bluffing* that φ", where an agent a addresses a sentence φ to another agent. Using these operators, we read that

- $[!_a\varphi]\psi$: ψ is true after a's truthful announcement of φ
- $[!_{¡a}\varphi]\psi$: ψ is true after a's lying announcement of φ
- $[!_{¡a}\varphi]\psi$: ψ is true after a's bluffing announcement of φ.

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 (written R_a for $a \in A$), and $V : S \to \mathcal{P}(P)$ is a valuation function (V_p represents the set of worlds where p is true). An *epistemic state* (M, s) for $s \in S$ satisfies sentences as follows:

$$
\begin{array}{ll}
M, s \models p & \text{iff} \quad s \in V_p \\
M, s \models \neg\varphi & \text{iff} \quad M, s \not\models \varphi \\
M, s \models \varphi \wedge \psi & \text{iff} \quad M, s \models \varphi \text{ and } M, s \models \psi \\
M, s \models B_a\varphi & \text{iff} \quad \forall t \in S : R_a(s,t) \text{ implies } M, t \models \varphi \\
M, s \models [!_a\varphi]\psi & \text{iff} \quad M, s \models B_a\varphi \text{ implies } M^{\varphi}_a, s \models \psi \\
M, s \models [!_{¡a}\varphi]\psi & \text{iff} \quad M, s \models B_a\neg\varphi \text{ implies } M^{\varphi}_a, s \models \psi \\
M, s \models [!_{¡a}\varphi]\psi & \text{iff} \quad M, s \models \neg\left(B_a\varphi \vee B_a\neg\varphi\right) \\
 & \qquad\qquad \text{implies } M^{\varphi}_a, s \models \psi
\end{array}
$$

where M^{φ}_a is as M except for the accessibility relation R' that is defined over S such that $R'_a := R_a$ and $R'_b := R_b \cap (S \times [\![B_a\varphi]\!]_M)$ for $b \in A\,(a \neq b)$ where $[\![\varphi]\!]_M := \{s \in S \mid M, s \models \varphi\}$. A sentence φ is *true* in a model M (written $M \models \varphi$) iff $M, s \models \varphi$ for any $s \in S$. φ is *valid* iff $M \models \varphi$ for any model M.

An agent a may address a sentence φ no matter what, whether she believes what she says, believes the opposite, or is uncertain. The situation is represented as the *precondition* $B_a\varphi$ for a's truthful announcement, $B_a\neg\varphi$ for a's lying announcement, and $\neg\left(B_a\varphi \vee B_a\neg\varphi\right)$ for a's bluffing announcement. Throughout the paper, agents are assumed to have the K45 axioms for the belief operator B: (*distribution*) $B_a(\varphi \supset \psi) \supset (B_a\varphi \supset B_a\psi)$; (*positive introspection*) $B_a\varphi \supset B_aB_a\varphi$; (*negative introspection*) $\neg B_a\varphi \supset B_a\neg B_a\varphi$, and inference rules: (*modus ponens*) $\frac{\varphi \quad \varphi\supset\psi}{\psi}$ and (*necessitation*) $\frac{\varphi}{B_a\varphi}$. 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 $\vdash \varphi$ iff a sentence φ 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*). The axioms for the belief consequences of agent communication are given as follows.

$$
\begin{array}{lll}
\text{(A1)} & [!_a\varphi]B_a\psi \equiv B_a\varphi \supset B_a[!_a\varphi]\psi \\
\text{(A2)} & [!_{¡a}\varphi]B_a\psi \equiv B_a\neg\varphi \supset B_a[!_{¡a}\varphi]\psi \\
\text{(A3)} & [!_{¡a}\varphi]B_a\psi \equiv \neg(B_a\varphi \vee B_a\neg\varphi) \supset B_a[!_{¡a}\varphi]\psi \\
\text{(A4)} & [!_a\varphi]B_b\psi \equiv B_a\varphi \supset B_b[!_a\varphi]\psi \\
\text{(A5)} & [!_{¡a}\varphi]B_b\psi \equiv B_a\neg\varphi \supset B_b[!_a\varphi]\psi \\
\text{(A6)} & [!_{¡a}\varphi]B_b\psi \equiv \neg(B_a\varphi \vee B_a\neg\varphi) \supset B_b[!_a\varphi]\psi
\end{array}
$$

By (A1)–(A3), 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 make a truthful announcement. (A1)–(A6) hold by replacing $B_a\psi$ (resp. $B_b\psi$) with $\neg B_a\psi$ (resp. $\neg B_b\psi$) on the left of $\equiv$ and replacing $B_a[\ldots]\psi$ (resp. $B_b[\ldots]\psi$) with $\neg B_a[\ldots]\psi$ (resp. $\neg B_b[\ldots]\psi$) 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 ($\neg B_b\neg B_a\varphi$) from unbelievable one ($B_b\neg B_a\varphi$). 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) $B_a[!_ap]q \equiv B_a[!_{¡a}p]q \equiv B_a[!_{¡a}p]q \equiv B_aq$.
(ii) $[!_ap]B_bB_ap \equiv [!_{¡a}p]B_bB_ap \equiv [!_{¡a}p]B_bB_ap \equiv \top$.

Proposition 2.1 does not hold for arbitrary sentences in general.

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 relationship between lying, deception, and attempted deception as in Figure 1.

Our primary interest in this section is to formulate "lies that deceive". 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

$$
\begin{aligned}
\mathsf{DBL}_{ab}(\varphi,\psi) \stackrel{def}{=}\ & B_a\neg\varphi \wedge \neg\psi \wedge B_b(\varphi \supset \psi) \\
& \wedge([!_{¡a}\varphi]B_b\psi \vee [!_{¡a}\varphi]\neg B_b\neg\psi). \qquad (1)
\end{aligned}
$$

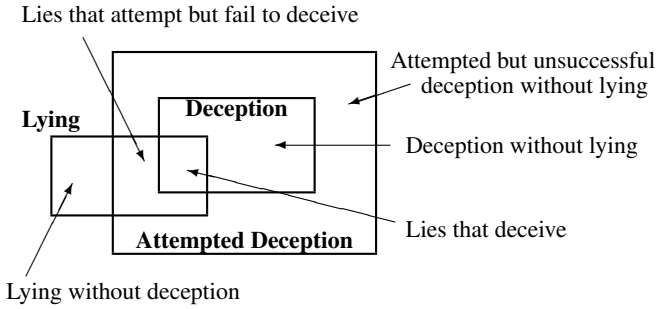

Figure 1: Lying, deception and attempted deception (Carson 2010)

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

$$\mathsf{DBL}_{ab}(\varphi, \varphi) = B_a\neg\varphi \wedge \neg\varphi \wedge ([!_a\varphi]B_b\varphi \vee [!_a\varphi]\neg B_b\neg\varphi).$$

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 φ ="an investment promises high return" and ψ ="the investment is worth buying." A salesperson makes the announcement φ which results in a customer's believing ψ (or disbelieving $\neg\psi$).

Note that (1) does not address whether a hearer believes ψ or $\neg\psi$ *before* the announcement. DBL happens whenever a hearer believes a false sentence ψ (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 φ but the actual falsity of φ is not necessarily requested. So if a speaker a makes a believed-false statement φ which is in fact true, then the announcement $[!_a\varphi]$ is lying. By contrast, the sentence ψ 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).

Example 3.2 shows a case of "lying without deception" in Figure 1. 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 φ and it brings about a hearer's believing a false sentence ψ, it is considered deception by lying, independent 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 φ but a hearer b already believes to the contrary, the hearer believes that the speaker is *mistaken*, namely $B_b(\neg\varphi \wedge B_a\varphi)$. If the hearer is a KD45 agent, neither $B_b\varphi$ nor $\neg B_b\neg\varphi$ holds after the lying announcement $[!_a\varphi]$, so that an attempted lie φ fails to deceive.

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

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

Proof. $B_b\neg\psi \wedge \mathsf{DBL}_{ab}(\varphi, \psi)$ implies $B_b\neg\psi \wedge (B_b\psi \vee \neg B_b\neg\psi)$ after the announcement $[!_a\varphi]$. Since $B_b\neg\psi \wedge (B_b\psi \vee \neg B_b\neg\psi) \equiv B_b\bot$, the result holds. □

By Proposition 3.1, $B_b\neg\varphi \wedge \mathsf{DBL}_{ab}(\varphi, \varphi) \supset \bot$. It is different from the case of $\varphi \equiv \neg\psi$ in (1), which becomes $\mathsf{DBL}_{ab}(\varphi, \neg\varphi) = B_a\neg\varphi \wedge \varphi \wedge B_b\neg\varphi \wedge ([!_a\varphi]B_b\neg\varphi \vee [!_a\varphi]\neg B_b\varphi)$. It represents a situation that both a speaker and a hearer believe a false fact $\neg\varphi$. In this case, a speaker's lying announcement $[!_a\varphi]$ sustains b's belief in $\neg\varphi$ or b's disbelief in φ. 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 $\psi \in \Phi$,*

$$\vdash B_bB_a\neg p \wedge \mathsf{DBL}_{ab}(p, \psi) \supset \bot.$$

Proof. It holds that $\vdash [!_a p]B_bB_ap \equiv \top$ (Proposition 2.1). Since $B_bB_a\neg p \wedge B_bB_ap \equiv B_bB_a\bot$ after the announcement $[!_a p]$, the result holds. □

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

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

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

Propositions 3.1–3.3 characterize different situations where "lies that attempt but fail to deceive" in Figure 1. In each case, deception fails if an addressee has consistent belief. Finally, $\mathsf{DBL}_{ab}(\varphi, \psi)$ does not imply $\mathsf{DBL}_{ab}(\varphi \wedge \lambda, \psi)$ for $\lambda \in \Phi$ in general. This means that even if an agent a successfully deceives another agent b by a lie φ, there is no guarantee that a can also deceives b using a stronger lie $\varphi \wedge \lambda$. A simple case is shown by putting $\lambda = \neg\varphi$, then $\mathsf{DBL}_{ab}(\varphi \wedge \neg\varphi, \psi)$ fails by Proposition 3.3.

3.2 Deception by Bluffing

We next provide an instance of "deception without lying" in Figure 1.

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

$$\mathsf{DBB}_{ab}(\varphi, \psi) \stackrel{def}{=} \neg(B_a\varphi \vee B_a\neg\varphi) \wedge \neg\psi \wedge B_b(\varphi \supset \psi) \wedge ([!_a\varphi]B_b\psi \vee [!_a\varphi]\neg B_b\neg\psi). \quad (2)$$

DBB is different from DBL in two aspects. First, $[!_{\neg a}\varphi]$ in (1) is replaced by $[!_{\neg a}\varphi]$ in (2). Second, $B_a\neg\varphi$ in (1) is replaced by $\neg(B_a\varphi \vee B_a\neg\varphi)$ in (2). On the other hand, ψ is false in both (1) and (2).

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

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

$$\vdash B_b\neg\psi \wedge \mathsf{DBB}_{ab}(\varphi,\psi) \supset \bot.$$

Proof. Similar to the proof of Proposition 3.1. □

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

Proposition 3.5 *Let b be a KD45 agent who believes that another agent a is KD45. For any $p\in P$ and $\psi\in\Phi$,*

$$\vdash B_bB_a\neg p \wedge \mathsf{DBB}_{ab}(p,\psi) \supset \bot.$$

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

$$\vdash B_b(\neg B_ap \wedge \neg B_a\neg p) \wedge \mathsf{DBB}_{ab}(p,\psi) \supset \bot.$$

Proof. It holds that $\vdash [!_{\neg a}p]B_bB_ap \equiv \top$ (Proposition 2.1). Since $B_b(\neg B_ap \wedge \neg B_a\neg p) \wedge B_bB_ap \equiv B_b\bot$ after the announcement $[!_{\neg a}p]$, the result holds. □

The difference between Propositions 3.5 and 3.6 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 3.7 *For any $\psi\in\Phi$,*

$$\vdash \mathsf{DBB}_{ab}(\bot,\psi) \supset \bot.$$

Proof. $\mathsf{DBB}_{ab}(\bot,\psi)$ implies $\neg B_a\top$, which implies $\bot$. □

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

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

$$\vdash \mathsf{DBB}_{ab}(B_a\varphi,\psi) \vee \mathsf{DBB}_{ab}(\neg B_a\varphi,\psi) \supset \bot.$$

Proof. Both $\mathsf{DBB}_{ab}(B_a\varphi,\psi)$ and $\mathsf{DBB}_{ab}(\neg B_a\varphi,\psi)$ imply $\neg B_aB_a\varphi \wedge \neg B_a\neg B_a\varphi$. Since $\vdash \neg B_aB_a\varphi \supset \neg B_a\varphi$ by the positive introspection of K45, $\neg B_aB_a\varphi$ implies $\neg B_a\varphi$, which implies $B_a\neg B_a\varphi$ 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 3.3 Suppose that John, who is interested in Mary, invites her to dinner on the 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 boy friend, 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 boy friend.

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

$$\mathsf{DBT}_{ab}(\varphi,\psi) \stackrel{def}{=} B_a\varphi \wedge \neg\psi \wedge B_b(\varphi\supset\psi) \wedge([!_a\varphi]B_b\psi \vee [!_a\varphi]\neg B_b\neg\psi). \quad (3)$$

Different from DBL, in DBT a speaker a's truthful announcement φ makes a hearer b believe a false sentence ψ, or a's truthful announcement φ makes b disbelieve a true sentence $\neg\psi$. In (3) a speaker a believes that φ is true, but the actual truth of φ 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 exclusive with each other.

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

Proposition 3.9 *Let b be a KD45 agent. For any $\varphi,\psi\in\Phi$,*

$$\vdash B_b\neg\psi \wedge \mathsf{DBT}_{ab}(\varphi,\psi) \supset \bot.$$

Proof. Similar to the proof of Proposition 3.1. □

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

Proposition 3.10 *Let b be a KD45 agent who believes that another agent a is KD45. For any $p\in P$ and $\psi\in\Phi$,*

$$\vdash B_bB_a\neg p \wedge \mathsf{DBT}_{ab}(p,\psi) \supset \bot.$$

Proof. Similar to the proof of Proposition 3.2. □

In Example 3.3, 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 3.11 *Let b be a KD45 agent. For any $p\in P$ and $\psi\in\Phi$,*

$$\vdash B_b(\neg B_ap \wedge \neg B_a\neg p) \wedge \mathsf{DBT}_{ab}(p,\psi) \supset \bot.$$

Proof. Similar to the proof of Proposition 3.6. □

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 characterize it as DBT with announcing a valid sentence.

Proposition 3.12 (deception by omission) *For any $\psi\in\Phi$,*

$$\vdash \mathsf{DBT}_{ab}(\top,\psi) \equiv \neg\psi \wedge B_b\psi.$$

Proof. $\mathsf{DBT}_{ab}(\top,\psi) \equiv \neg\psi \wedge B_b\psi \wedge ([!_a\top]B_b\psi \vee [!_a\top]\neg B_b\neg\psi)$. Then the result holds by the fact $[!_a\top]B_b\psi \equiv B_b\psi$. □

Proposition 3.12 says that if deception by omission happens, a hearer (initially) believes a false sentence ψ.

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 $\mathsf{DBL}_{ab}(\varphi, \psi)$, a speaker a lies a believed-false sentence φ to a hearer b, while the speaker does not necessarily believe that the announcement will result in the hearer's believing another false sentence ψ. In $\mathsf{DBT}_{ab}(\varphi, \psi)$, a speaker a tells a believed-true sentence φ to a hearer b, so that the speaker might not feel guilty even if the announcement leads a hearer to believe a false sentence ψ 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* (I-DBL), *intentional deception by bluffing* (I-DBB) and *intentional deception by truth-telling* (I-DBT) are respectively defined as:

$$\mathsf{I\text{-}DBL}_{ab}(\varphi, \psi) \stackrel{def}{=} B_a\neg\psi \wedge B_aB_b(\varphi \supset \psi) \wedge \mathsf{DBL}_{ab}(\varphi, \psi).$$
$$\mathsf{I\text{-}DBB}_{ab}(\varphi, \psi) \stackrel{def}{=} B_a\neg\psi \wedge B_aB_b(\varphi \supset \psi) \wedge \mathsf{DBB}_{ab}(\varphi, \psi).$$
$$\mathsf{I\text{-}DBT}_{ab}(\varphi, \psi) \stackrel{def}{=} B_a\neg\psi \wedge B_aB_b(\varphi \supset \psi) \wedge \mathsf{DBT}_{ab}(\varphi, \psi).$$

In particular, *intentional deception by omission* becomes

$$\mathsf{I\text{-}DBT}_{ab}(\top, \psi) = \neg\psi \wedge B_b\psi \wedge B_a\neg\psi \wedge B_aB_b\psi.$$

In Definition 4.1, a speaker a believes the falsity of the sentence ψ and also believes that a hearer believes the implication $\varphi \supset \psi$. With this additional condition, if the speaker makes a lying announcement $[!_a\varphi]$ expecting that the announcement will cause the hearer's believing the false sentence ψ, 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 $\varphi, \psi \in \Phi$,*

$$\vdash \mathsf{DBX}_{ab}(\varphi, \psi) \wedge B_a(\mathsf{DBX}_{ab}(\varphi, \psi)) \supset \mathsf{I\text{-}DBX}_{ab}(\varphi, \psi)$$

where "X" means one of "L", "B", and "T".

Proof. Since $B_a(\mathsf{DBX}_{ab}(\varphi, \psi))$ implies $B_a\neg\psi \wedge B_aB_b(\varphi \supset \psi)$, the result holds by definition. □

Proposition 4.2 *For any $\varphi \in \Phi$,*

- $\vdash \mathsf{I\text{-}DBL}_{ab}(\varphi, \varphi) \equiv \mathsf{DBL}_{ab}(\varphi, \varphi)$.
- $\vdash \mathsf{I\text{-}DBB}_{ab}(\varphi, \varphi) \supset \bot$.
- $\vdash \mathsf{I\text{-}DBT}_{ab}(\varphi, \varphi) \supset \bot$ *for any KD45 agent a.*

Proof. $\mathsf{I\text{-}DBL}_{ab}(\varphi, \varphi) \equiv \mathsf{DBL}_{ab}(\varphi, \varphi)$ holds by definition. $\mathsf{I\text{-}DBB}_{ab}(\varphi, \varphi)$ implies $B_a\neg\varphi \wedge \neg B_a\varphi \wedge \neg B_a\neg\varphi \equiv \bot$. $\mathsf{I\text{-}DBT}_{ab}(\varphi, \varphi)$ implies $B_a\neg\varphi \wedge B_a\varphi \equiv B_a\bot$. □

By Proposition 4.2, there is no distinction between DBL and I-DBL if a speaker lies a believed-false sentence φ that results in a hearer's believing φ or disbelieving $\neg\varphi$. In other words, a liar always intends to deceive a hearer wrt the sentence being announced in DBL. On the other hand, $\mathsf{I\text{-}DBB}_{ab}(\varphi, \varphi)$ or $\mathsf{I\text{-}DBT}_{ab}(\varphi, \varphi)$ is impossible for a speaker having consistent beliefs. This is because in DBB a speaker has no belief of φ which contradicts the additional condition $B_a\neg\varphi$. In DBT a speaker believes φ which also contradicts the additional condition $B_a\neg\varphi$. By this fact, DBB or DBT can be intentional only if a hearer comes to believe a false sentence that is *different* from the sentence announced by a speaker. The fact also implies that, compared to I-DBL, I-DBB or I-DBT generally requires advanced techniques for a speaker because a deceiver is requested to select an announcement that is different from the false fact which the deceiver wants a hearer to believe. The situation is explained in the literature that "the deceiver takes a more circuitous route to his success, where lying is an easier and more certain way to mislead" (Adler 1997, p.440). According to studies in psychology, children lie by four years or earlier, mainly for avoiding punishment (Ekman 1989). Very young children do not have advanced techniques of deception, then most deception by them is of the type $\mathsf{(I\text{-})DBL}_{ab}(\varphi, \varphi)$ 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, makes a truthful announcement 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 4.1 John, who visits a clinic for a medical checkup, is diagnosed as having a serious cancer. A doctor does not inform the patient of this fact in fear of discouraging him. John has no symptom giving him any reason to believe this fact, 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 truth-telling.

The situation of Example 4.1 is represented in our formulation as

$$\mathsf{DBL}_{ab}(\varphi, \varphi) \wedge \mathsf{DBT}_{bc}(\varphi, \varphi) \tag{4}$$

where $a =$ doctor, $b =$ John, $c =$ wife, and $\varphi =$"normal". In this case, a doctor indirectly deceives John's wife by lying. Indirect deception (4) may happen for intentional DBL, but does not hold for intentional DBT for KD45 agents (Proposition 4.2). Generally, acts of deceiving produce indirect deception as follows.

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:

$$\mathsf{IN\text{-}DBL}_{ac}(\varphi,\lambda) \stackrel{def}{=} \mathsf{(I\text{-})DBL}_{ab}(\varphi,\psi) \wedge \mathsf{DBT}_{bc}(\psi,\lambda).$$

$$\mathsf{IN\text{-}DBB}_{ac}(\varphi,\lambda) \stackrel{def}{=} \mathsf{(I\text{-})DBB}_{ab}(\varphi,\psi) \wedge \mathsf{DBT}_{bc}(\psi,\lambda).$$

$$\mathsf{IN\text{-}DBT}_{ac}(\varphi,\lambda) \stackrel{def}{=} \mathsf{(I\text{-})DBT}_{ab}(\varphi,\psi) \wedge \mathsf{DBT}_{bc}(\psi,\lambda).$$

In $\mathsf{IN\text{-}DBL}_{ac}(\varphi,\lambda)$, a's lying announcement on a sentence φ results in b's believing a false sentence ψ, and b's truthful announcement on a sentence ψ results in c's believing a false sentence λ. $\mathsf{IN\text{-}DBB}_{ac}(\varphi,\lambda)$ and $\mathsf{IN\text{-}DBT}_{ac}(\varphi,\lambda)$ 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,

$$\mathsf{(I\text{-})DBL}_{ab}(\varphi,\psi) \wedge \mathsf{DBL}_{bc}(\neg\psi,\lambda)$$

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 ψ, but it would not contributes to c's believing a false fact λ 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

$$\mathsf{(I\text{-})DBX}_{ab}(\varphi,\psi_1) \wedge \mathsf{DBT}_{bc}(\psi_1,\psi_2) \wedge \mathsf{DBT}_{cd}(\psi_2,\psi_3) \wedge \cdots$$

where "X" is one of L, B and T in general.[1]

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 or psychology (Demos 1960; McLaughlin and Rorty 1988; da Costa and French 1990; Trivers 2011). Self-deception involves a person holding contradictory beliefs ($B_a(\varphi \wedge \neg\varphi)$), 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.[2] In this section, we characterize self-deception in our formulation.

In DBL, a KD45 agent cannot deceive itself on a lying sentence.

Proposition 4.3 *Let a be a KD45 agent. For any $p \in P$,*

$$\vdash \mathsf{DBL}_{aa}(p,p) \supset \bot\,.$$

Proof. By definition,

$$\mathsf{DBL}_{aa}(p,p) = B_a\neg p \wedge \neg p \wedge ([!_{a}p]B_ap \vee [!_{a}p]\neg B_a\neg p).$$

Using the axiom (A2), it becomes

$$\begin{aligned}
&B_a\neg p \wedge \neg p \wedge ((B_a\neg p \supset B_a[!_{a}p]p) \vee (B_a\neg p \supset \neg B_a[!_{a}p]\neg p))\\
&\quad\equiv B_a\neg p \wedge \neg p \wedge (B_a[!_{a}p]p \vee \neg B_a[!_{a}p]\neg p)\\
&\quad\equiv B_a\neg p \wedge \neg p \wedge (B_ap \vee \neg B_a\neg p) \ \text{(Proposition 2.1)}\\
&\quad\equiv B_a(\neg p \wedge p) \wedge \neg p\\
&\quad\equiv B_a\bot \wedge \neg p.
\end{aligned}$$

Thus, $\vdash \mathsf{DBL}_{aa}(p,p) \equiv B_a\bot \wedge \neg p$. Since a is a KD45 agent, the result holds. □

Proposition 4.3 implies that if an agent a is KD but not KD45, $\mathsf{DBL}_{aa}(p,p)$ involves a mental state of an agent who has contradictory belief wrt a false fact p.[3]

By contrast, $\mathsf{DBL}_{aa}(\varphi,\psi)$ implies

$$B_a\neg\varphi \wedge \neg\psi \wedge B_a(\varphi \supset \psi) \wedge (B_a\psi \vee \neg B_a\neg\psi) \quad (5)$$

that is consistent for a sentence $\psi \not\equiv \varphi$. (5) represents *counterfactual inference* involved in self-deception, that is, a speaker believes the falsity of φ while believes the effect ψ or its possibility that would be obtained if φ 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 $p \in P$,*

- $\vdash \mathsf{DBB}_{aa}(p,p) \supset \neg\,(B_ap \vee B_a\neg p) \wedge \neg p$.
- $\vdash \mathsf{DBT}_{aa}(p,p) \supset B_ap \wedge \neg p$.

Proof.

- $$\begin{aligned}
\mathsf{DBB}_{aa}(p,p) &= \neg\,(B_ap \vee B_a\neg p) \wedge \neg p\\
&\qquad \wedge ([!_{a}p]B_ap \vee [!_{a}p]\neg B_a\neg p)\\
&\equiv \neg\,(B_ap \vee B_a\neg p) \wedge \neg p \wedge (B_ap \vee \neg B_a\neg p)\\
&\qquad \text{(by (A3) and Proposition 2.1)}\\
&\equiv \neg\,(B_ap \vee B_a\neg p) \wedge \neg p.
\end{aligned}$$

- $$\begin{aligned}
\mathsf{DBT}_{aa}(p,p) &= B_ap \wedge \neg p \wedge ([!_{a}p]B_ap \vee [!_{a}p]\neg B_a\neg p)\\
&\equiv B_ap \wedge \neg p \wedge (B_ap \vee \neg B_a\neg p)\\
&\qquad \text{(by (A4) and Proposition 2.1)}\\
&\equiv B_ap \wedge \neg p.
\end{aligned}$$ □

Proposition 4.4 presents that neither $\mathsf{DBB}_{aa}(p,p)$ nor $\mathsf{DBT}_{aa}(p,p)$ implies contradiction. 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 $\varphi,\psi \in \Phi$,*

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

[1] Consider rumors that get distorted and exaggerated.

[2] "In short, self-deception involves an inner conflict, perhaps the existence of contradiction" (Demos 1960, p. 588).

[3] (Jones 2013) characterizes a group of "self-deception positions" consistently using KD as the logic of belief.

Proof. Intentional deception assumes $B_a\neg\psi$, while $B_a\psi \vee \neg B_a\neg\psi$ holds after the announcement. Hence, the result holds. □

Note that $\mathsf{DBL}_{aa}(\varphi,\psi)$, $\mathsf{DBB}_{aa}(\varphi,\psi)$, and $\mathsf{DBT}_{aa}(\varphi,\psi)$ are all consistent when deception is unintentional ((5) 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.[4]

Finally, one can indirectly deceive oneself. The following scenario is a modification of the "appointment example" of (McLaughlin 1988, p. 31).[5]

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

$$\mathsf{IN\text{-}DBL}_{aa}(\varphi,\varphi) = \mathsf{I\text{-}DBL}_{ab}(\varphi,\varphi) \wedge \mathsf{DBT}_{ba}(\varphi,\varphi)$$

where $a =$ Mary, $b =$ online assistant, and $\varphi =$"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

$$\mathsf{IN\text{-}DBX}_{aa}(\varphi,\lambda) = \mathsf{(I\text{-})DBX}_{ab}(\varphi,\psi) \wedge \mathsf{DBT}_{ba}(\psi,\lambda)$$

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*

$$\nvdash \mathsf{IN\text{-}DBX}_{aa}(\varphi,\lambda) \supset \bot.$$

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. 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: $\neg B_a\varphi \wedge E_aB_b\varphi$ or $B_a\neg\varphi \wedge E_aB_b\varphi$ where $E_a\psi$ means "an agent a brings about that ψ". On the other hand, cases that an agent who does not succeed in his/her attempt to deceive another agent are formally represented as: $\neg B_a\varphi \wedge H_aB_b\varphi$ or $B_a\neg\varphi \wedge H_aB_b\varphi$, where $H_a\psi$ means that "an agent a attempts to bring about ψ, 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: $Dec_{ab}\,\varphi := I_aB_b\varphi \wedge B_a\neg\varphi \wedge B_b\varphi$. 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 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 $Lie_a(\varphi)$ and $True_a(\varphi)$. The former represents an action in which an agent a publicly lies that she knows φ while in fact she does not know it. The latter 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, however, the action $Lie_a(\varphi)$ does not succeed. Such a condition is formulated as an action's *contextual appearance*. Note that the precondition $\neg K_a\varphi$ of $Lie_a(\varphi)$ represents the ignorance of φ and is different from the one used in the agent announcement logic. They argue deception accompanied 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. $\neg B_a\varphi \wedge B_aB_a\varphi$) 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. (5), Propositions 4.4 and 4.6). Sakama *et al.* (2010) 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 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 logical account of different categories of de-

[4] "... self-deception occurs when the conscious mind is kept in dark" (Trivers 2011, p. 9).

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

ception 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

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 explains what is deception and what is not. It provides conditions under which deception is considered intentional and represents various aspects of self-deception. A dynamic epistemic logic can express both an act of deceiving and its effect on addressees' belief. 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

Adler, J. E. 1997. Lying, deceiving, or falsely implicating. *Journal of Philosophy* 94:435–452.

Baltag, A. and Smets, S. 2008. 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.

Carson, T. L. 2010. *Lying and Deception: Theory and Practice*. Oxford University Press.

Castelfranchi, C. 2000. Artificial liars: why computers will (necessarily) deceive us and each other? *Ethics and Information Technology* 2:113–119.

Chisholm, R. M. and Feehan, T. D. 1977. The intent to deceive. *Journal of Philosophy* 74:143–159.

Clark, M. 2011. Mendacity and deception: uses and abuses of common ground. AAAI Fall Symposium, FS-11-02, AAAI Press.

da Costa, N. C. A. and French, S. 1990. Belief, contradiction and the logic of self-deception. *American Philosophical Quarterly* 27:179–197.

Demos, R. 1960. Lying to oneself. *Journal of Philosophy* 57:588–595.

Ekman, P. 1989. *Why Kids Lie: How Parents Can Encourage Truthfulness*. Scribner.

Ettinger, D. and Jehiel, P. 2010. A theory of deception. *American Economic Journal: Microeconomics* 2:1–20.

Firozabadi, B. S.; Tan, Y. H.; and Lee, R. M. 1999. 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.

Hespanha, J. P.; Ateskan, Y. S.; and Kizilocak, H. 2000. Deception in non-cooperative games with partial information. In: Proc. 2nd DARPA-JFACC Symposium on Advances in Enterprise Control.

Jones, A. J. I. 2013. Self-deception and the logic of belief. Invited talk at: 11th European Workshop on Multi-Agent Systems (EUMAS).

Mahon, J. E. 2007. A definition of deceiving. *Journal of Applied Philosophy* 21:181–194.

McLaughlin B. P. and Rorty, A. O. eds. 1988. *Perspectives on Self-Deception*, University of California Press.

McLaughlin, B. P. 1988. Exploring the possibility of self-deception in belief. In: (McLaughlin and Rorty 1988), 29–62.

O'Neill, B. 2003. A formal system for understanding lies and deceit. Jerusalem Conference on Biblical Economics.

Pörn, I. 1989. On the nature of social order. In Fenstad, J. E. et al. eds. *Logic, Methodology, and Philosophy of Science*, VIII. Elsevier.

Sakama, C.; Caminada, M.; and Herzig, A. 2010. A logical account of lying. In: *Proc. 12th European Conference on Logics in Artificial Intelligence, Lecture Notes in Artificial Intelligence* 634, 286–299, Springer.

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

Sakama, C.; Caminada, M.; and Herzig, A. 2015. A formal account of dishonesty. *Logic Journal of the IGPL* 23:259-294.

Shim, J. and Arkin, R. C. 2012. Biologically-inspired deceptive behavior for a robot. In: *Proc. 12th Int'l Conf. Simulation of Adaptive Behavior, LNCS* 7426, 401–411, Springer.

Staab, E. and Caminada, M. 2011. On the profitability of incompetence. In: *Multi-Agent-Based Simulation XI, Lecture Notes in Computer Science* 6532, 76–92, Springer.

Trivers, R. 2011. *The Folly of Fools: The Logic of Deceit and Self-Deception in Human Life*. Basic Books.

van Ditmarsch, H.; van Eijck, J.; Sietsma, F.; and Wang, Y. 2012. On the logic of lying. *Games, Actions and Social Software*, LNCS 7010, 41–72, Springer.

van Ditmarsch, H. 2014. Dynamics of lying. *Synthese* 191:745–777.

Vincent, J. M. and Castelfranchi, C. 1981. 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.

Wagner, A. R. and Arkin, R. C. 2011. Acting deceptively: providing robots with the capacity for deception. *Journal of Social Robotics* 3:5–26.

Zlotkin, G. and Rosenschein, J. S. 1991. Incomplete information and deception in multi-agent negotiation. In: *Proc. 12th Int'l Joint Conference on Artificial Intelligence*, 225–231, Morgan Kaufmann.