The Many Faces of Deception

Chiaki Sakama[1] and Martin Caminada[2]

[1] Department of Computer and Communication Sciences
Wakayama University, Japan
sakama@sys.wakayama-u.ac.jp
[2] Faculty of Science, Technology and Communication
University of Luxembourg, Luxembourg
martin.caminada@uni.lu

Abstract. This paper formalises various forms of *deception*, that were originally described in an informal way by Chisholm and Feehan [5]. To this end, we use a modal logic of action and belief developed by Pörn *et al.* We provide logical definitions of *deception by commission* and *deception by omission*, and investigate their formal properties. We also study *intended deception* and discuss its relationship to *lying* and *withholding information*. According to our formulation, deception is captured as a form of reasoning about causation that is inherently nonmonotonic in its nature.

1 Introduction

Deception is an act whereby one person causes another person to have a false belief. In spite of its commonality in human society, little study has been devoted to developing a formal account of deception. Once one has a good understanding of what deception is, one can consider the best ways of using deception to achieve a particular goal, as well as the best ways to avoid being deceived. Such considerations are particularly of interest in a game-theoretical perspective [7] and designing intelligent agents in multiagent systems [20]. The definition of deception has been studied by a number of philosophers (for instance, [1, 5, 12] and the references therein). Among others, Chisholm and Feehan [5] provide eight basic ways in which a person may deceive another person with respect to a proposition. Their classification is based on three distinctions: *deception by commission* versus *deception by omission* (which is related to the attitude of the deceiver), *positive deception* versus *negative deception* (which is related to the belief state of the hearer), and *deception simpliciter* versus *deception secundum quid* (which is related to whether one changes a belief state or merely sustains it). These three types of distinctions yield eight different categories in which an agent deceives another agent. These categories can then further be divided depending on whether the deception is intended or not.

Deception involves a success or an achievement of the act. Mahon [12] mentions that "Whether or not an act of deceiving has occurred depends on whether or not a particular effect – normally, the having of a false belief – has been produced in another; if no such effect has been produced in another, then no deceiving has occurred." In this sense, deception is considered a *perlocutionary act* [19]. This implies that, to provide

a formal account of deception, we need a logic that can express belief, action, causation, and intention of an agent. To this end, we use a modal logic of belief and action, introduced by Pörn *et al.* [15, 16, 18], and then extend the logic to represent intention of an agent. We formulate eight different categories of deception in the logic, and investigate their formal properties. We also distinguish intended and unintended deception, and compare intended deception with lying and withholding information.

The rest of this paper is organized as follows. Section 2 introduces a modal language for belief and action. Section 3 formulates different categories of deception and investigates formal properties. Section 4 argues intended deception and its relationship to lying and withholding information. Section 5 discusses nonmonotonicity of deception and related studies. Finally, Section 6 rounds off the paper.

2 A modal logic for belief and action

In this paper, we consider a propositional modal logic for a causal theory of action, based on a theory developed by Pörn and Sandu. The logic introduced in this section is due to [16, 18][3], except for the introduction of the unary predicate $utter_{xy}$ to the language.

Let P be a set of propositional constants and A a finite set of agents. Then, a *sentence* in the language $\mathcal{L}_0$ is defined as follows.

(i) If $p \in P$, then p is a sentence.
(ii) If ϕ and ψ are sentences, then $\neg\phi$, $\phi \vee \psi$, $\phi \wedge \psi$, $\phi \supset \psi$, and $\phi \equiv \psi$ are all sentences. In particular, $\top$ and $\bot$ represent valid and contradictory sentences, respectively.
(iii) If ϕ is a sentence and $a, b \in A$, then $utter_{ab}(\phi)$ is a sentence.
(iv) If ϕ is a sentence and $a \in A$, then $B_a\phi$, $D_a\phi$, and $D'_a\phi$ are all sentences.

The set of all sentences in $\mathcal{L}_0$ is denoted by Φ. Here, $utter_{ab}$ is a unary predicate that represents a speech act of an agent. A sentence $utter_{ab}(\phi)$ means that an agent a expresses a sentence ϕ to another agent b. On the other hand, B_a, D_a, and D'_a are modal operators which respectively mean that

$B_a\phi$: "a believes ϕ",
$D_a\phi$: "it is necessary for something which a does that ϕ",
$D'_a\phi$: "but for a's action it would be the case that ϕ".

The Kripkean semantics of the operators $D_a\phi$ and $D'_a\phi$ are given as follows. Let (W, R, V) be a structure of the language, where W is a non-empty set of possible worlds, R a binary accessibility relation in W, and V a valuation function which assigns a set $U \subseteq W$ to each atomic sentence. The D_a*-models* are structures of the kind (W, R_{D_a}, V) in which R_{D_a} is reflexive and $(w, w') \in R_{D_a}$ if and only if everything which a brings about it in w is the case in w'. By contrast, the D'_a*-models* are structures of the kind $(W, R_{D'_a}, V)$ in which $R_{D'_a}$ is serial[4] and $(w, w') \in R_{D'_a}$ if and only if not everything which a brings about it in w is the case in w'.

[3] Pörn [15] also provides a logic which is slightly different from [16].
[4] R is serial if $\forall u \in W\, \exists v \in W$ s.t. $(u, v) \in R$.

The axiom system for the B-operator is KD45, while the axiom systems for the D-operator and the D'-operator are KT and KD, respectively. A logic BDD', that is defined over $\mathcal{L}_0$, has the following axioms and inference rules. For $\phi, \psi \in \Phi$,

$(\mathbf{P})$ All propositional tautologies.
$(\mathbf{U_C})$ $utter_{ab}(\phi \wedge \psi) \equiv utter_{ab}(\phi) \wedge utter_{ab}(\psi)$.
$(\mathbf{K_B})$ $B_a\phi \wedge B_a(\phi \supset \psi) \supset B_a\psi$. $(\mathbf{K_D})$ $D_a\phi \wedge D_a(\phi \supset \psi) \supset D_a\psi$.
$(\mathbf{K_{D'}})$ $D'_a\phi \wedge D'_a(\phi \supset \psi) \supset D'_a\psi$. $(\mathbf{T_D})$ $D_a\phi \supset \phi$.
$(\mathbf{D_B})$ $B_a\phi \supset \neg B_a\neg\phi$. $(\mathbf{D_{D'}})$ $D'_a\phi \supset \neg D'_a\neg\phi$.
$(\mathbf{4_B})$ $B_a\phi \supset B_aB_a\phi$. $(\mathbf{5_B})$ $\neg B_a\phi \supset B_a\neg B_a\phi$.

$$(\mathbf{MP})\ \frac{\phi \quad \phi \supset \psi}{\psi} \qquad (\mathbf{N_B})\ \frac{\phi}{B_a\phi} \qquad (\mathbf{N_D})\ \frac{\phi}{D_a\phi} \qquad (\mathbf{N_{D'}})\ \frac{\phi}{D'_a\phi}$$

Using these modalities, new operators E and F are defined as

$$E_a\phi \stackrel{def}{=} D_a\phi \wedge \neg D'_a\phi : \text{“}a \text{ brings it about that } \phi\text{”},$$
$$F_a\phi \stackrel{def}{=} \neg D_a\neg\phi \wedge \neg D'_a\phi : \text{“}a \text{ let it be the case that } \phi\text{”}.$$

The sentence $E_a\phi$ consists of two parts: $D_a\phi$ represents that a's action is sufficient for a state of affairs ϕ to occur, while $\neg D'_a\phi$ represents the necessary condition for agency, that is, without a's action ϕ might not have occurred. $F_a\phi$ is obtained by replacing $D_a\phi$ in $E_a\phi$ with its dual $\neg D_a\neg\phi$ meaning that it is compatible with everything which a does that ϕ. The operators E and F have the following properties [16]:

$(\mathbf{T_E})$ $E_a\phi \supset \phi$.
$(\mathbf{D_E})$ $E_a\phi \supset \neg E_a\neg\phi$.
$(\mathbf{K_E})$ $E_a(\phi \supset \psi) \supset (E_a\phi \supset E_a\psi)$.
$(\mathbf{C_E})$ $(E_a\phi \wedge E_a\psi) \supset E_a(\phi \wedge \psi)$.
$(\mathbf{No})$ $\neg E_a\top \wedge \neg F_a\top$.
$(\mathbf{EF})$ $E_a\phi \supset F_a\phi$.

$(\mathbf{T_E})$ represents the success of actions. $(\mathbf{No})$ represents that no agent can bring about what is logically true, that is, the logical truth is independent of agents and their actions.[5]

The system BDD' is further extended to represent a *causal relation* between an action and an effect. A language $\mathcal{L}_1$ is constructed from $\mathcal{L}_0$ by introducing a binary relation $\phi \Rightarrow \psi$ which is read as “ψ is a consequence of ϕ”. The relation $\Rightarrow$ satisfies the

[5] $\neg F_a\top$ is not in [16] but directly follows by the definition of F_a and $(\mathbf{N_{D'}})$.

following axioms and inference rules [18]. For $\phi, \psi, \chi \in \Phi$,

$(\mathbf{C_1})$ $\neg(\phi \Rightarrow \bot)$.

$(\mathbf{C_2})$ $(\phi \Rightarrow \psi \wedge \psi \Rightarrow \chi) \supset \phi \Rightarrow \chi$.

$(\mathbf{C_3})$ $(\phi \Rightarrow \psi) \supset \neg(\psi \Rightarrow \phi)$.

$(\mathbf{C_4})$ $\neg(\phi \Rightarrow \psi \wedge \neg\phi \Rightarrow \psi)$.

$(\mathbf{C_5})$ $(\phi \Rightarrow \psi) \supset (\phi \wedge \psi)$.

$(\mathbf{C_6})$ $(\phi \Rightarrow \psi \wedge \phi \Rightarrow \chi) \supset (\phi \Rightarrow \psi \wedge \chi)$.

$(\mathbf{C_7})$ $\dfrac{\phi \equiv \psi}{(\phi \Rightarrow \chi) \equiv (\psi \Rightarrow \chi)}$ $(\mathbf{C_8})$ $\dfrac{\phi \equiv \psi}{(\chi \Rightarrow \phi) \equiv (\chi \Rightarrow \psi)}$

We remark some useful properties of $\Rightarrow$. Putting $\phi \equiv \psi$ in $(\mathbf{C_3})$, it holds that

$$\neg(\phi \Rightarrow \phi).$$

In contrast to $(\mathbf{C_6})$, the following implication does *not* hold in general.

$$(\phi \Rightarrow \chi \wedge \psi \Rightarrow \chi) \supset (\phi \wedge \psi \Rightarrow \chi).$$

The system $BDD'C$, defined over $\mathcal{L}_1$, is an extension of BDD' with $(\mathbf{C_1})$–$(\mathbf{C_8})$. In $BDD'C$ the operator B_a is also applied to sentences in $\mathcal{L}_1$. If a sentence ϕ is a theorem of $BDD'C$, we write $\vdash \phi$. By $\mathbf{N_B}$ each agent believes that other agents follow the same logic $BDD'C$. Thus, $B_a B_b \phi \supset B_a \neg B_b \neg \phi$ and $B_a(E_b\phi \wedge E_b(\phi \supset \psi)) \supset B_a E_b \psi$, for instance.

Action logic is *nonmonotonic* in the sense that an agent does not necessarily bring about all the consequences of his/her actions [18]. So the following inference rules do *not* hold in general:

$$(\mathbf{Mon}) \quad \frac{\phi \supset \psi}{E_a\phi \supset E_a\psi} \qquad \frac{\phi \supset \psi}{F_a\phi \supset F_a\psi}$$

Thus, one cannot conclude $E_a\psi$ from $E_a\phi$ and $\phi \supset \psi$. Otherwise, $E_a\top$ is derived from $E_a\phi$ and $\phi \supset \top$, which contradicts $\neg E_a\top$ by (**No**). The nonmonotonicity of F_a is shown in a similar way. It is worth noting that the logic does not handle temporal aspects of actions, and it just describes worlds where an action has been taken place.

3 Deception

Chisholm and Feehan [5] consider eight basic ways in which a person may deceive another person with respect to a proposition. They are classified into two groups: *deception by commission* and *deception by omission*.

3.1 Deception by Commission

Let a and b be two agents and p a false proposition. Then, *deception by commission* is the following four cases:

(DC1) a contributes causally toward b's acquiring the belief in p.
(DC2) a contributes causally toward b's continuing in the belief in p.
(DC3) a contributes causally toward b's ceasing to believe in $\neg p$.
(DC4) a contributes causally toward preventing b from acquiring the belief in $\neg p$.

(DC1) is called *positive deception simpliciter* and (DC2) is called *positive deception secundum quid*. (DC3) is called *negative deception simpliciter* and (DC4) is called *negative deception secundum quid*. Here, (DC1) and (DC2) are called "positive" because deception causes b to believe p, while (DC3) and (DC4) are called "negative" because deception causes b to disbelieve $\neg p$. On the other hand, (DC1) and (DC3) are called "simpliciter" because a simply causes b to change his/her belief with respect to p, while (DC2) and (DC4) are called "secundum quid" because a qualifies b to hold his/her belief with respect to p. (DC1)–(DC4) do not restrict deception to be a speech act, but here we consider deception as a statement of a sentence. The expression "contributes causally toward" means that an agent contributes causally to making a certain things happen, but it does not necessarily imply that the agent is in any sense a "total cause" of the things he/she makes happen [4]. The above four categories of deception are then formulated in the language $\mathcal{L}_1$ as follows.

Definition 3.1 (positive deception by commission) Let a and b be two agents and $\sigma \in \Phi$. Then, *positive deception simpliciter by commission* (shortly, $PDSC$) is defined as

$$PDSC_{ab}(\sigma) \stackrel{def}{=} B_a \neg\sigma \wedge (utter_{ab}(\sigma) \Rightarrow E_a B_b \sigma)\,. \tag{1}$$

By contrast, *positive deception secundum quid by commission* (shortly, $PDSQC$) is defined as

$$PDSQC_{ab}(\sigma) \stackrel{def}{=} B_a \neg\sigma \wedge (utter_{ab}(\sigma) \Rightarrow F_a B_b \sigma)\,. \tag{2}$$

In this case, we say that an agent a *deceives* another agent b on the sentence σ by PDSC (or PDSQC).

(1) represents that a deceives b on σ by PDSC if a utters a believed-false sentence σ and the utterance causally brings it about that b believes σ. (2) represents that a deceives b on σ by PDSQC if a utters a believed-false sentence σ and the utterance causally lets it be the case that b believes σ. By Definition 3.1 and ($\mathbf{C_5}$), $PDSC_{ab}(\sigma)$ implies $utter_{ab}(\sigma) \wedge E_a B_b \sigma$, and $PDSQC_{ab}(\sigma)$ implies $utter_{ab}(\sigma) \wedge F_a B_b \sigma$. In $PDSC_{ab}(\sigma)$, $E_a B_b \sigma$ implies $D_a B_b \sigma \wedge \neg D'_a B_b \sigma$. Of which, $D_a B_b \sigma$ means that a's utterance is sufficient for b's believing σ, and $\neg D'_a B_b \sigma$ means that without a's utterance, it might not happen that b's believing σ. In other words, b believes σ after the act $utter_{ab}(\sigma)$, and the utterance causally contributes to b's believing σ. In $PDSQC_{ab}(\sigma)$, on the other hand, $F_a B_b \sigma$ implies $\neg D_a \neg B_b \sigma \wedge \neg D'_a B_b \sigma$. Of which, $\neg D_a \neg B_b \sigma$ means that a's utterance is not a sufficient condition for b's disbelieving σ. In other words, b would not disbelieve σ after the act $utter_{ab}(\sigma)$, and the utterance causally contributes to b's believing σ.

Example 3.1 Suppose a salesperson a who believes that a product has no value ($B_a \neg\sigma$ where $\sigma = valuable$). If the salesperson utters to a customer b that it is valuable

($utter_{ab}(\sigma)$) and the speech act leads the customer, who disbelieves that the product is valuable, to believing it valuable ($B_b\sigma$), then it is $PDSC_{ab}(\sigma)$. On the other hand, if the speech act could let the customer, who believes that the product is valuable, continue to have the (wrong) belief, then it is $PDSQC_{ab}(\sigma)$.

Definition 3.2 (negative deception by commission) Let a and b be two agents and $\sigma \in \Phi$. Then, *negative deception simpliciter by commission* (shortly, $NDSC$) is defined as

$$NDSC_{ab}(\sigma) \stackrel{def}{=} B_a\neg\sigma \wedge (utter_{ab}(\sigma) \Rightarrow E_a\neg B_b\neg\sigma)\,. \tag{3}$$

By contrast, *negative deception secundum quid by commission* (shortly, $NDSQC$) is defined as

$$NDSQC_{ab}(\sigma) \stackrel{def}{=} B_a\neg\sigma \wedge (utter_{ab}(\sigma) \Rightarrow F_a\neg B_b\neg\sigma)\,. \tag{4}$$

In this case, we say that an agent a *deceives* another agent b on the sentence σ by NDSC (or NDSQC).

(3) represents that a deceives b on σ by NDSC if a utters a believed-false sentence σ and the utterance causally brings it about that b disbelieves $\neg\sigma$. (4) represents that a deceives b on σ by NDSQC if a utters a believed-false sentence σ and the utterance causally lets it be the case that b disbelieves $\neg\sigma$ (thus preventing b from acquiring the belief in $\neg\sigma$). By Definition 3.2 and ($\mathbf{C_5}$), $NDSC_{ab}(\sigma)$ implies $utter_{ab}(\sigma) \wedge E_a\neg B_b\neg\sigma$, and $NDSQC_{ab}(\sigma)$ implies $utter_{ab}(\sigma) \wedge F_a\neg B_b\neg\sigma$. In $NDSC_{ab}(\sigma)$, $E_a\neg B_b\neg\sigma$ implies $D_a\neg B_b\neg\sigma \wedge \neg D'_a\neg B_b\neg\sigma$. Of which, $D_a\neg B_b\neg\sigma$ means that a's utterance is sufficient for b's disbelieving $\neg\sigma$, and $\neg D'_a\neg B_b\neg\sigma$ means that without a's utterance, it might not happen that b's disbelieving $\neg\sigma$. In other words, b disbelieves $\neg\sigma$ after the act $utter_{ab}(\sigma)$, and the utterance causally contributes to b's disbelieving $\neg\sigma$. In $NDSQC_{ab}(\sigma)$, on the other hand, $F_a\neg B_b\neg\sigma$ implies $\neg D_a B_b\neg\sigma \wedge \neg D'_a\neg B_b\neg\sigma$. Of which, $\neg D_a B_b\neg\sigma$ means that a's utterance is not a sufficient condition for b's believing $\neg\sigma$. In other words, b would not believe $\neg\sigma$ after the act $utter_{ab}(\sigma)$, and the utterance causally contributes to b's disbelieving $\neg\sigma$.

Example 3.2 Consider again a salesperson who believes that a product has no value but utters to a customer that it is valuable. If the speech act makes the customer believe the possibility of the value ($\neg B_b\neg\sigma$), then it is $NDSC_{ab}(\sigma)$. On the other hand, if by the speech act the customer continues to believe the possibility of the value, then it is $NDSQC_{ab}(\sigma)$.

We will often use the notation DC_{ab} (or DC when subscripts are unimportant), in referring to either $PDSC_{ab}$, $PDSQC_{ab}$, $NDSC_{ab}$, or $NDSQC_{ab}$.

3.2 Deception by Omission

Let a and b be two agents and p a false proposition. Then, *deception by omission* is the following four cases:

(DO1) a allows b to acquire the belief in p.

(DO2) a allows b to continue in the belief in p.
(DO3) a allows b to cease to have the belief in $\neg p$.
(DO4) a allows b to continue without the belief in $\neg p$.

(DO1) is called *positive deception simpliciter* and (DO2) is called *positive deception secundum quid*. (DO3) is called *negative deception simpliciter* and (DO4) is called *negative deception secundum quid*. Here, an agent *allows* a certain state of affairs to occur provided only (i) he/she could prevent that state of affairs from occurring and (ii) he/she does not thus prevent it from occurring [5]. We capture the act that "a allows b" as "a makes no utterance to b on a believed-true sentence". That is, a proactively do nothing to affect b's state of mind of believing what is true. The above four categories of deception are then formulated in the language $\mathcal{L}_1$ as follows.

Definition 3.3 (positive deception by omission) Let a and b be two agents and $\sigma \in \Phi$. Then, *positive deception simpliciter by omission* (shortly, $PDSO$) is defined as

$$PDSO_{ab}(\sigma) \stackrel{def}{=} B_a\neg\sigma \wedge (\neg utter_{ab}(\neg\sigma) \Rightarrow E_a B_b\sigma)\,. \tag{5}$$

By contrast, *positive deception secundum quid by omission* (shortly, $PDSQO$) is defined as

$$PDSQO_{ab}(\sigma) \stackrel{def}{=} B_a\neg\sigma \wedge (\neg utter_{ab}(\neg\sigma) \Rightarrow F_a B_b\sigma)\,. \tag{6}$$

In this case, we say that an agent a *deceives* another agent b on the sentence σ by PDSO (or PDSQO).

(5) represents that a deceives b on σ by PDSO if a does not utter a believed-true sentence $\neg\sigma$ and the non-utterance causally brings it about that b believes σ. (6) represents that a deceives b on σ by PDSQO if a does not utter a believed-true sentence $\neg\sigma$ and the non-utterance causally lets it be the case that b believes σ. By Definition 3.3 and ($\mathbf{C_5}$), both $PDSO_{ab}(\sigma)$ and $PDSQO_{ab}(\sigma)$ imply $\neg utter_{ab}(\neg\sigma)$. So the act of deception implies no utterance of the sentence $\neg\sigma$. This is the only difference from positive deception by commission.

Example 3.3 Suppose a child a who believes that he/she failed to get a passing grade ($B_a\neg\sigma$ where $\sigma = passing\_grade$) but does not utter the fact to his/her parent b ($\neg utter_{ab}(\neg\sigma)$). If the non-utterance leads the parent to believing that the child gets a passing grade, then it is $PDSO_{ab}(\sigma)$. On the other hand, if the non-utterance leads the parent to retaining the wrong belief ($B_b\sigma$), then it is $PDSQO_{ab}(\sigma)$.

Definition 3.4 (negative deception by omission) Let a and b be two agents and $\sigma \in \Phi$. Then, *negative deception simpliciter by omission* (shortly, $NDSO$) is defined as

$$NDSO_{ab}(\sigma) \stackrel{def}{=} B_a\neg\sigma \wedge (\neg utter_{ab}(\neg\sigma) \Rightarrow E_a\neg B_b\neg\sigma)\,. \tag{7}$$

By contrast, *negative deception secundum quid by omission* (shortly, $NDSQO$) is defined as

$$NDSQO_{ab}(\sigma) \stackrel{def}{=} B_a\neg\sigma \wedge (\neg utter_{ab}(\neg\sigma) \Rightarrow F_a\neg B_b\neg\sigma)\,. \tag{8}$$

In this case, we say that an agent a *deceives* another agent b on the sentence σ by NDSO (or NDSQO).

(7) represents that a deceives b on σ by NDSO if a does not utter a believed-true sentence $\neg\sigma$ and the non-utterance causally brings it about that b disbelieves $\neg\sigma$. (8) represents that a deceives b on σ by NDSQO if a does not utter a believed-true sentence $\neg\sigma$ and the non-utterance causally lets it be the case that b disbelieves $\neg\sigma$. $NDSO_{ab}(\sigma)$ and $NDSQO_{ab}(\sigma)$ imply $\neg utter_{ab}(\neg\sigma)$, so the act of deception implies no utterance of the sentence $\neg\sigma$. This is the only difference from negative deception by commission.

Example 3.4 Consider again a child who believes that he/she failed to get a passing grade. If the non-utterance makes the parent believe the possibility of the child's getting a passing grade ($\neg B_b \neg\sigma$), then it is $NDSO_{ab}(\sigma)$. On the other hand, if the non-utterance leads the parent to retaining the belief on the possibility of the child's getting a passing grade, then it is $NDSQO_{ab}(\sigma)$.

We will often use the notation DO_{ab} (or DO when subscripts are unimportant), in referring to either $PDSO_{ab}$, $PDSQO_{ab}$, $NDSO_{ab}$, or $NDSQO_{ab}$.

Comparing deception by commission and deception by omission, Chisholm and Feehan argue that the former is considered to be more sinful than the latter [5]. This is because in DC a deceiver behaves more proactively than in DO. Thus, $PDSC$ is worse (or more sinful) than $PDSO$, $PDSQC$ is worse than $PDSQO$, $NDSC$ is worse than $NDSO$, and $NDSQC$ is worse than $NDSQO$. These relationships imply a guideline for a speech-act of deception that should ideally be satisfied by the agents, for moral as well as for self-interested reasons (lower punishments if caught). We provide them as postulates for DC and DO as follows.

Postulates for DC and DO

$$(\textbf{PDSC-PDSO}):\ B_a(PDSC_{ab}(\sigma)) \wedge B_a(PDSO_{ab}(\sigma)) \supset \neg PDSC_{ab}(\sigma).$$
$$(\textbf{PDSQC-PDSQO}):\ B_a(PDSQC_{ab}(\sigma)) \wedge B_a(PDSQO_{ab}(\sigma)) \supset \neg PDSQC_{ab}(\sigma).$$
$$(\textbf{NDSC-NDSO}):\ B_a(NDSC_{ab}(\sigma)) \wedge B_a(NDSO_{ab}(\sigma)) \supset \neg NDSC_{ab}(\sigma).$$
$$(\textbf{NDSQC-NDSQO}):\ B_a(NDSQC_{ab}(\sigma)) \wedge B_a(NDSQO_{ab}(\sigma)) \supset \neg NDSQC_{ab}(\sigma).$$

(**PDSC-PDSO**) says that if a believes that both $PDSC_{ab}(\sigma)$ and $PDSO_{ab}(\sigma)$ are effective, then a does not opt for $PDSC_{ab}(\sigma)$. In other words, if one achieves deception both by uttering a believed-false sentence and by not uttering a believed-true sentence, then the later one is preferred from moral and self-interested reasons. The rest of postulates have similar meanings.

3.3 Formal Properties

Let Dec be one of the eight categories of deception, that is, it is either $PDSC$, $PDSQC$, $NDSC$, $NDSQC$, $PDSO$, $PDSQO$, $NDSO$, or $NDSQO$. We first show that it is inconsistent to deceive on valid or contradictory sentences.

Proposition 3.1 *Let a and b be two agents. Then,*

(i) $\vdash Dec_{ab}(\top) \supset \bot$.
(ii) $\vdash Dec_{ab}(\bot) \supset \bot$.

Proof. (i) $DC_{ab}(\top)$ or $DO_{ab}(\top)$ implies $B_a\bot$ that implies $\neg B_a\top$ ($\mathbf{D_B}$), while $\top$ implies $B_a\top$ ($\mathbf{N_B}$). Contradiction.

(ii) $PDSC_{ab}(\bot)$ or $PDSO_{ab}(\bot)$ implies $E_aB_b\bot$ that implies $B_b\bot$ ($\mathbf{T_E}$). That leads to contradiction as above. $PDSQC_{ab}(\bot)$ or $PDSQO_{ab}(\bot)$ implies $F_aB_b\bot$ that implies $\neg D_a\neg B_b\bot \wedge \neg D'_aB_b\bot$. However, from $B_b\top$ ($\mathbf{N_B}$) it follows that $\neg B_b\bot$ ($\mathbf{D_B}$), so $D_a\neg B_b\bot$ ($\mathbf{N_D}$). Contradiction. Next, $NDSC_{ab}(\bot)$ or $NDSO_{ab}(\bot)$ implies $E_a\neg B_b\top$ that implies $\neg B_b\top$ ($\mathbf{T_E}$). Contradiction. $NDSQC_{ab}(\bot)$ or $NDSQO_{ab}(\bot)$ implies $F_a\neg B_b\top$ that implies $\neg D_aB_b\top \wedge \neg D'_a\neg B_b\top$. However, from $B_b\top$ ($\mathbf{N_B}$) it follows that $D_aB_b\top$ ($\mathbf{N_D}$). Contradiction. □

Self-deception leads to contradiction.

Proposition 3.2 *Let a be an agent and $\sigma \in \Phi$. Then, $Dec_{aa}(\sigma) \supset \bot$.*

Proof. We show the results for deception by commission. $PDSC_{aa}(\sigma)$ implies $B_a\neg\sigma \wedge E_aB_a\sigma$. $B_a\neg\sigma$ implies $\neg B_a\sigma$ ($\mathbf{D_B}$), while $E_aB_a\sigma$ implies $B_a\sigma$ ($\mathbf{T_E}$). Contradiction. $NDSC_{aa}(\sigma)$ implies $B_a\neg\sigma \wedge E_a\neg B_a\neg\sigma$. $E_a\neg B_a\neg\sigma$ implies $\neg B_a\neg\sigma$ which contradicts $B_a\neg\sigma$. Next, $PDSQC_{aa}(\sigma)$ implies $B_a\neg\sigma \wedge F_aB_a\sigma$. $B_a\neg\sigma$ implies $\neg B_a\sigma$ ($\mathbf{D_B}$), which implies $D_a\neg B_a\sigma$ ($\mathbf{N_D}$). On the other hand, $F_aB_a\sigma$ implies $\neg D_a\neg B_a\sigma$. Contradiction. $NDSQC_{aa}(\sigma)$ implies $B_a\neg\sigma \wedge F_a\neg B_a\neg\sigma$. $B_a\neg\sigma$ implies $D_aB_a\neg\sigma$ ($\mathbf{N_D}$), while $F_a\neg B_a\neg\sigma$ implies $\neg D_aB_a\neg\sigma$. Contradiction. The results for deception by omission are shown in a similar manner. □

Generally, $\phi \Rightarrow \chi_1$ and $\psi \Rightarrow \chi_2$ do not imply $\phi \wedge \psi \Rightarrow \chi_1 \wedge \chi_2$. If this is the case, the implication $(\phi \Rightarrow \chi \wedge \psi \Rightarrow \chi) \supset (\phi \wedge \psi \Rightarrow \chi)$ holds when $\chi_1 \equiv \chi_2$, but this implication does not hold as remarked in Section 2. This implies that two deceptions do not work conjunctively in general.

Proposition 3.3 *Let a and b be two agents and $\sigma, \lambda \in \Phi$. Then,*

$$\not\forall\ Dec_{ab}(\sigma) \wedge Dec_{ab}(\lambda) \supset Dec_{ab}(\sigma \wedge \lambda)\,.$$

Proposition 3.3 indicates that even if an agent a successfully deceives another agent b on sentences σ and λ individually, there is no guarantee that a can also deceives b on the sentence $\sigma \wedge \lambda$ using the axiom ($\mathbf{U_C}$). In other words, *a success of a small deception does not imply a success of a big deception in general.*

Deception simpliciter does not imply deception secundum quid because $E_a\phi \Rightarrow E_a\psi$ and $E_a\psi \supset F_a\psi$ do not imply $E_a\phi \Rightarrow F_a\psi$. By the nonmonotonicity of E_a and F_a, positive deception does not imply negative deception.

Proposition 3.4 *Let a and b be two agents and $\sigma \in \Phi$. Then,*

(i) $\not\forall\ PDSC_{ab}(\sigma) \supset NDSC_{ab}(\sigma)$.
(ii) $\not\forall\ PDSQC_{ab}(\sigma) \supset NDSQC_{ab}(\sigma)$.
(iii) $\not\forall\ PDSO_{ab}(\sigma) \supset NDSO_{ab}(\sigma)$.
(iv) $\not\forall\ PDSQO_{ab}(\sigma) \supset NDSQO_{ab}(\sigma)$.

Proof. (i) $PDSC_{ab}(\sigma)$ implies $E_aB_b\sigma$, but $E_aB_b\sigma$ and $B_a\sigma \supset \neg B_a\neg\sigma$ do not imply $E_a\neg B_b\neg\sigma$ by the nonmonotonicity of E_a, hence the result holds. Likewise, (ii)–(iv) hold by the nonmonotonicity of E_a and F_a. □

These facts imply that there is no inclusion relation between eight categories of deception (Figure 1).

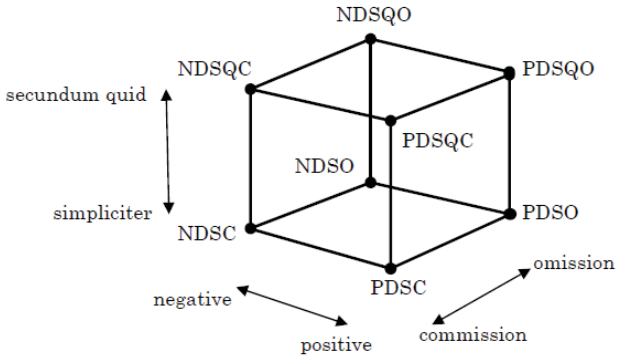

Fig. 1. Eight categories of deception

4 Intended Deception

4.1 Intention to Deceive

One usually has motives for making deception. For instance, a salesperson deceives a customer on the value of a product to sell it. In this situation, the salesperson has an intention to convince the customer of the value of the product. In this section, we consider deception accompanied with such *intention*. To this end, we expand our logic by introducing the modal operator representing an intention of an agent.[6]

The propositional modal language $\mathcal{L}_2$ is an extension of the language $\mathcal{L}_0$, which is obtained from $\mathcal{L}_0$ by introducing the modal operator I_a. Given a sentence ϕ in $\mathcal{L}_2$, $I_a\phi$ represents that "an agent a intends ϕ". Semantically, $I_a\phi$ holds if and only if ϕ is true in all states of affairs compatible with a's current intentions.[7] The I-operator also follows the KD45 system and has the following axioms and inference rules.

$(\mathbf{P})$ All propositional tautologies.

$(\mathbf{K_I})$ $I_a\phi \wedge I_a(\phi \supset \psi) \supset I_a\psi$ $\qquad$ $(\mathbf{D_I})$ $I_a\phi \supset \neg I_a\neg\phi$

$(\mathbf{4_{IB}})$ $I_a\phi \supset B_a I_a\phi$ $\qquad$ $(\mathbf{5_{IB}})$ $\neg I_a\phi \supset B_a\neg I_a\phi$

$(\mathbf{MP})$ $\dfrac{\phi \quad \phi \supset \psi}{\psi}$ $\qquad$ $(\mathbf{N_I})$ $\dfrac{\phi}{I_a\phi}$

The axiomatic system for B and I is also introduced in [6] as a logic for intentional communication. Remark that $(\mathbf{N_I})$ says that all theorems hold at all state of affairs that an agent might intend to realize [6].

[6] Pörn [15] captures intention as holding a belief concerning one's own action. He then represents intentional action as $B_a(p \supset Shall\ E_a q)$, meaning that "$a$ intends to bring it about that q if p", where $Shall$ is a modality representing "it shall be the case that". Instead of introducing $Shall$ and employing a circuitous definition, we introduce the modal operator I_a representing a's intention.

[7] With such an interpretation, it becomes acceptable to apply an intention operator to arbitrary sentences.

The set of all sentences in $\mathcal{L}_2$ is denoted by Ψ. A language $\mathcal{L}_3$ is constructed from $\mathcal{L}_2$ by introducing the causal relation $\phi \Rightarrow \psi$ for $\phi, \psi \in \Psi$.

The system $BDD'CI$, defined over $\mathcal{L}_3$, is obtained by introducing $(\mathbf{K_I})$, $(\mathbf{D_I})$, $(\mathbf{4_{IB}})$, $(\mathbf{5_{IB}})$, and $(\mathbf{N_I})$ to $BDD'C$.

Definition 4.1 (intended deception by commission) Let a and b be two agents and $\sigma, \phi \in \Psi$. Then, the following four categories of deception are called *intended deception by commission*.

$$IPDSC_{ab}(\sigma,\phi) \stackrel{def}{=} B_a\neg\phi \wedge I_a B_b\phi \wedge B_a B_b(\sigma \supset \phi) \wedge (utter_{ab}(\sigma) \Rightarrow E_a B_b\phi)\,. \quad (9)$$

$$IPDSQC_{ab}(\sigma,\phi) \stackrel{def}{=} B_a\neg\phi \wedge I_a B_b\phi \wedge B_a B_b(\sigma \supset \phi) \wedge (utter_{ab}(\sigma) \Rightarrow F_a B_b\phi)\,. \quad (10)$$

$$INDSC_{ab}(\sigma,\phi) \stackrel{def}{=} B_a\neg\phi \wedge I_a\neg B_b\neg\phi \wedge B_a B_b(\sigma \supset \phi) \wedge (utter_{ab}(\sigma) \Rightarrow E_a\neg B_b\neg\phi)\,. \quad (11)$$

$$INDSQC_{ab}(\sigma,\phi) \stackrel{def}{=} B_a\neg\phi \wedge I_a\neg B_b\neg\phi \wedge B_a B_b(\sigma \supset \phi) \wedge (utter_{ab}(\sigma) \Rightarrow F_a\neg B_b\neg\phi)\,. \quad (12)$$

We will use the notation IDC_{ab} (or IDC when subscripts are unimportant), in referring to either $IPDSC_{ab}$, $IPDSQC_{ab}$, $INDSC_{ab}$, or $INDSQC_{ab}$.

In (9), a has an intention to make b believe the believed-false sentence ϕ. a also believes that a sentence σ leads b to believing ϕ, and the utterance of a sentence σ brings it about that b believes ϕ. In this case, we say that an agent a *deceives* another agent b on the sentence σ for ϕ by IPDSC. IPDSQC, INDSC and INDSQC are explained in similar ways.

In contrast to $DC_{ab}(\sigma)$, in $IDC_{ab}(\sigma,\phi)$ a speaker utters a sentence σ to make b believe ϕ. Note that a believes the falsity of ϕ, but does not necessarily believe the falsity of σ.

Example 4.1 Bob and Mary are working at the same office and Bob is playing a computer game. They know that their boss usually arrives at the office at ten o'clock. Mary knows that the boss will not arrive at ten today for some reasons ($B_a\neg\phi$ where $\phi = boss\_arrives\_at\_ten$), but intends Bob to believe he'll arrive, in order for Bob to stop playing the game. When the clock strikes ten, she utters "Now it's ten o'clock". From this utterance ($utter_{ab}(\sigma)$ where $\sigma = it's\_ten$), it follows that Bob realizes the boss is coming ($B_b\phi$). This is an example of $IPDSQC_{ab}(\sigma,\phi)$.

When $\phi \equiv \sigma$, $IDC_{ab}(\sigma,\sigma)$ is simplified as follows.[8]

[8] Chisholm and Feehan [5] also consider intended deception in the sense of Proposition 4.1.

Proposition 4.1 *Let a and b be two agents and $\sigma \in \Psi$.*

$$IPDSC_{ab}(\sigma,\sigma) \equiv PDSC_{ab}(\sigma) \wedge I_a B_b \sigma.$$
$$IPDSQC_{ab}(\sigma,\sigma) \equiv PDSQC_{ab}(\sigma) \wedge I_a B_b \sigma.$$
$$INDSC_{ab}(\sigma,\sigma) \equiv NDSC_{ab}(\sigma) \wedge I_a \neg B_b \neg \sigma.$$
$$INDSQC_{ab}(\sigma,\sigma) \equiv NDSQC_{ab}(\sigma) \wedge I_a \neg B_b \neg \sigma\,.$$

Intended deception by omission is defined in a similar way.

Definition 4.2 (intended deception by omission) Let a and b be two agents and $\sigma, \phi \in \Psi$. Then, the following four categories of deception are called *intended deception by omission*.

$$IPDSO_{ab}(\sigma,\phi) \stackrel{def}{=} B_a \neg\phi \wedge I_a B_b \phi \wedge B_a B_b(\neg\sigma \supset \neg\phi) \wedge (\neg utter_{ab}(\neg\sigma) \Rightarrow E_a B_b \phi)\,. \quad (13)$$

$$IPDSQO_{ab}(\sigma,\phi) \stackrel{def}{=} B_a \neg\phi \wedge I_a B_b \phi \wedge B_a B_b(\neg\sigma \supset \neg\phi) \wedge (\neg utter_{ab}(\neg\sigma) \Rightarrow F_a B_b \phi)\,. \quad (14)$$

$$INDSO_{ab}(\sigma,\phi) \stackrel{def}{=} B_a \neg\phi \wedge I_a \neg B_b \neg\phi \wedge B_a B_b(\neg\sigma \supset \neg\phi) \wedge (\neg utter_{ab}(\neg\sigma) \Rightarrow E_a \neg B_b \neg\phi)\,. \quad (15)$$

$$INDSQO_{ab}(\sigma,\phi) \stackrel{def}{=} B_a \neg\phi \wedge I_a \neg B_b \neg\phi \wedge B_a B_b(\neg\sigma \supset \neg\phi) \wedge (\neg utter_{ab}(\neg\sigma) \Rightarrow F_a \neg B_b \neg\phi)\,. \quad (16)$$

We will use the notation IDO_{ab} (or IDO when subscripts are unimportant), in referring to either $IPDSO_{ab}$, $IPDSQO_{ab}$, $INDSO_{ab}$, or $INDSQO_{ab}$.

When $\phi \equiv \sigma$, $IDO_{ab}(\sigma,\sigma)$ is simplified as follows.

Proposition 4.2 *Let a and b be two agents and $\sigma \in \Psi$.*

$$IPDSO_{ab}(\sigma,\sigma) \equiv PDSO_{ab}(\sigma) \wedge I_a B_b \sigma.$$
$$IPDSQO_{ab}(\sigma,\sigma) \equiv PDSQO_{ab}(\sigma) \wedge I_a B_b \sigma.$$
$$INDSO_{ab}(\sigma,\sigma) \equiv NDSO_{ab}(\sigma) \wedge I_a \neg B_b \neg \sigma.$$
$$INDSQO_{ab}(\sigma,\sigma) \equiv NDSQO_{ab}(\sigma) \wedge I_a \neg B_b \neg \sigma\,.$$

Properties similar to Propositions 3.1– 3.4 hold for IDC and IDO. In contrast to intended deception, $DC_{ab}(\sigma)$ and $DO_{ab}(\sigma)$ are also called *unintended deception*.

Like unintended deception, IDC is considered more sinful than IDO. Then, postulates for intended deception are given as follows.

Postulates for IDC and IDO

$$(\textbf{IPDSC-IPDSO}): \; B_a(IPDSC_{ab}(\sigma,\phi)) \wedge B_a(IPDSO_{ab}(\sigma,\phi)) \supset \neg IPDSC_{ab}(\sigma,\phi).$$

$$(\textbf{IPDSQC-IPDSQO}): \; B_a(IPDSQC_{ab}(\sigma,\phi)) \wedge B_a(IPDSQO_{ab}(\sigma,\phi)) \supset \neg IPDSQC_{ab}(\sigma,\phi).$$

$$(\textbf{INDSC-INDSO}): \; B_a(INDSC_{ab}(\sigma,\phi)) \wedge B_a(INDSO_{ab}(\sigma,\phi)) \supset \neg INDSC_{ab}(\sigma,\phi).$$

$$(\textbf{INDSQC-INDSQO}): \; B_a(INDSQC_{ab}(\sigma,\phi)) \wedge B_a(INDSQO_{ab}(\sigma,\phi)) \supset \neg INDSQC_{ab}(\sigma,\phi).$$

4.2 Relationship to Lying

Next we consider a relationship between intended deception and *lying*. As addressed in the introduction, the verb *deceive* is a success or an achievement verb, so that deception accompanies an achievement of the act. In this respect, "deceiving differs from lying" [12]. That is, "*lie* is not a success or an achievement verb, and an act of lying is not a perlocutionary act. Whether or not an act of lying has occurred does not depend on whether or not a particular effect – for example, the belief that what the liar says is true – has been produced in another; if no effect has been produced in another, then an act of lying may have occurred nevertheless" (ibid). With this respect, a formal definition of lying is given in [17] as follows.

$$LIE_{ab}(\sigma) \stackrel{def}{=} utter_{ab}(\sigma) \wedge B_a\neg\sigma \wedge I_a B_b\sigma. \tag{17}$$

The definition says that a lies to b on a sentence σ if a utters the believed-false sentence σ to b with the intention that σ is believed by b. Comparing (17) with

$$IPDSC_{ab}(\sigma,\sigma) \equiv B_a\neg\sigma \wedge I_a B_b\sigma \wedge (utter_{ab}(\sigma) \Rightarrow E_a B_b\sigma),$$

we can observe the difference between lying and deception. That is, in lying a speaker a utters a sentence σ, but lying in general does not concern about whether a hearer b believes σ or not.[9] On the other hand, it is vital in IPDSC that the utterance brings it about that $B_b\sigma$, and it is represented by $utter_{ab}(\sigma) \Rightarrow E_a B_b\sigma$. Formally, the following results hold.

Proposition 4.3 *Let a and b be two agents and $\sigma \in \Psi$. Then,*

(i) $IPDSC_{ab}(\sigma,\sigma) \supset LIE_{ab}(\sigma)$.
(ii) $IPDSQC_{ab}(\sigma,\sigma) \supset LIE_{ab}(\sigma)$.

Proof. Both $IPDSC_{ab}(\sigma,\sigma)$ and $IPDSQC_{ab}(\sigma,\sigma)$ imply $utter_{ab}(\sigma)$. Hence, the results hold. □

The converse implications of the above results do not hold in general. There is no implication relationship between lying and intended negative deception by commission.

[9] Chisholm and Feehan [5] employ a stronger definition of lying and argue that lying always involves the intent of positive deception simpliciter.

4.3 Relationship to Withholding Information

Sometimes the act of simply remaining silent with a deceptive intention is called "lie of omission" [13] or "withholding information" [3]. Carson remarks that there is a clear distinction between withholding information and deception. According to [3], "to withhold information is to fail to offer information that would help someone acquire true beliefs and/or correct false beliefs. Not all cases of withholding information constitute deception ... withholding information can constitute deception if there is a clear expectation, promise, and/or professional obligation that such information will be provided".

Here we capture withholding information such that an agent a makes no utterance to b on a believed-true sentence $\neg\sigma$ with the intention that σ is believed by b. Formally, it is stated as follows.

$$WI_{ab}(\sigma) \stackrel{def}{=} \neg utter_{ab}(\neg\sigma) \wedge B_a\neg\sigma \wedge I_a B_b\sigma. \tag{18}$$

Comparing (18) with

$$IPDSO_{ab}(\sigma,\sigma) \equiv B_a\neg\sigma \wedge\ I_a B_b\sigma \wedge\ (\neg utter_{ab}(\neg\sigma) \Rightarrow E_a B_b\sigma),$$

we can observe the difference between withholding information and deception that is similar to the difference between lying and deception. Formally, the following results hold.

Proposition 4.4 *Let a and b be two agents and $\sigma \in \Psi$. Then,*

(i) $IPDSO_{ab}(\sigma,\sigma) \supset WI_{ab}(\sigma)$.
(ii) $IPDSQO_{ab}(\sigma,\sigma) \supset WI_{ab}(\sigma)$.

Proof. Both $IPDSO_{ab}(\sigma,\sigma)$ and $IPDSQO_{ab}(\sigma,\sigma)$ imply $\neg utter_{ab}(\neg\sigma)$. Hence, the results hold. □

The converse implications of the above results do not hold in general. There is no implication relationship between withholding information and intended negative deception by omission.

5 Discussion

5.1 Nonmonotonicity in Deception

In many cases, deception involves nonmonotonicity. For instance, in intended deception $IPDSC_{ab}(\sigma,\phi)$, the speaker a utters a sentence σ with the intention to make b believe ϕ. This is done by the belief of a that b believes the sentence $\sigma \supset \phi$. Such beliefs are defeasible, however, and $B_a B_b(\sigma \supset \phi) \wedge utter_{ab}(\sigma)$ does not necessarily have the consequence $E_a B_b\phi$. Moreover, as we mentioned in Section 2, the action operators E_a and F_a are nonmonotonic, so that deception that is defined using those operators inherits nonmonotonicity (cf. Proposition 3.4).

There are different definitions of deception in the literature of philosophy [12]. Caminada [2] investigates another type of deception introduced by Adler [1]. Different

from deception considered in this paper, a deceiver asserts what he/she believes true, while, at the same time, he/she conceals something of the truth hoping that a hearer will make an incorrect inference based on incomplete beliefs. This type of deception is formally defined in [17] as[10]

$$DEC_{ab}(\sigma,\delta) \stackrel{def}{=} utter_{ab}(\sigma) \wedge B_a\sigma \wedge I_aB_b\sigma \wedge\ B_aB_b((\sigma \wedge \neg B_b\neg\delta) \supset \delta) \\ \wedge B_a\neg B_b\neg\delta \wedge \neg utter_{ab}(\neg\delta) \wedge B_a\neg\delta \wedge I_aB_b\delta.$$

The definition says that the speaker a utters a believed-true sentence σ with the intention of making a hearer b believe it. a believes that b uses σ to reach a default conclusion δ. a also believes that b does not believe the falsity of δ while a believes it. a does not utter the believed-true sentence $\neg\delta$ to b. And believing δ by the hearer b is what the speaker a intends to achieve. By the definition, we can see that $DEC_{ab}(\sigma,\delta)$ implies $WI_{ab}(\delta)$. That is, deception of this type uses withholding information.

In essence, Caminada *et al.*'s notion of deception [2, 17] relies on the *nonmonotonic inference* capabilities of the hearer to reach a wrong conclusion. For instance, if the speaker tells the believed-true statement that Tweety is a bird, while withholding the believed-true statement that Tweety is a penguin, then the hearer might make a default inference yielding the conclusion that Tweety flies, which the speaker believes to be false. Hence, the speaker is able to implant a believed-false statement into the belief base of the hearer, without having to resort to telling direct lies (that is, without having to tell things the speaker believes to be false). Such nonmonotonicity is realized in $DEC_{ab}(\sigma,\delta)$ by the formula $B_b((\sigma \wedge \neg B_b\neg\delta) \supset \delta)$ in a way similar to *autoepistemic logic*. The role of nonmonotonic reasoning in Caminada *et al.*'s treatment of deception is that of a facilitator of purposely wrong inferences, by providing believed-true information in a strategic way.

Another important difference between Caminada *et al.*'s notion of deception and the one considered in this paper is that the former does not necessarily imply the success of the act. In fact, $DEC_{ab}(\sigma,\delta)$ does not describe the effect of deception. In this sense, deception formulated in [2, 17] is considered *attempted deception* [3]. The relationship between intended deception, attempted deception, lying, and withholding information is illustrated in Figure 2.[11] The six categories of intended deception considered in Section 4.1 are put in the figure. IPDSC and IPDSQC are lies (Section 4.2) and IPDSO and IPDSQO are withholding information (Section 4.3). INDSC, INDSQC, INDSO, and INDSQO are intended deception, but they are neither lying nor withholding information. As argued above, DEC is attempted deception that uses withholding information.

In practice, detecting attempts to deceive is important to prevent success of deception. To detect deception and its attempt, however, information is required about other agent's belief and intention. The problems, as well as the issue of how to detect the opponent's beliefs in order to generate the most effective forms of deception, are still open research issues.

[10] We slightly modified the definition of [17] by explicitly including the condition $\neg utter_{ab}(\neg\delta)$.
[11] A similar figure between lying, deception, and attempted deception is in [3].

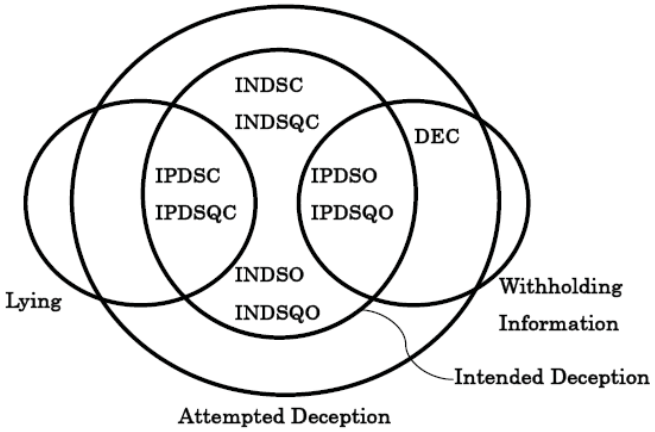

Fig. 2. Intended Deception, Attempted Deception, Lying and Withholding Information

5.2 Related Work

Formal studies of dishonesty have been relatively rare until now. A semi-formal treatment of dishonesty is provided by Caminada [2]. In this work, three basic categories of dishonesty are discussed: *lies* (where the speaker makes a statement which he/she believes to be false) [11, 13], *bullshit* (where the speaker makes a statement he/she does not believe to be true and does not believe to be false) [10], and *deception* (where the speaker is making a statement he/she believes to be true, but hopes that the hearer will use it to draw a conclusion that is believed to be false) [1]. This distinction is worked out in more details in [17], where a full formal account is given. The notion of deception employed in [2] and [17] is based on the one by Adler [1]. So it is different from that of Chisholm and Feehan [5] and is therefore also different from the notion of deception in the current paper.

Some studies use action logic to characterize dishonesty of agents. Firozabadi *et al.* [8] formulate *fraud* using modal operators for obligation, action and belief. According to their definition, fraud is a situation such that an agent violates one of its obligations and also deceives another agent that the obligation is fulfilled. 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 to believe the opposite. These two cases are formally defined in [8] as: $\neg B_a\phi \wedge E_a B_b\phi$ and $B_a\neg\phi \wedge E_a B_b\phi$. These definitions do not mention any mean that an agent uses to bring about a result. The authors also consider an agent who does not succeed in his/her attempt to deceive another agent is still a deceptive agent. Such cases, that we call attempted deception and distinguish them from deception, are formally represented by $\neg B_a\phi \wedge H_a B_b\phi$ and $B_a\neg\phi \wedge H_a B_b\phi$, where $H_a\psi$ means that "an agent a attempts to bring about ψ, not necessarily successful". Their objective is to characterize specific types of fraud situations that may occur in organized interactions like trade procedures.

Firozabadi and Jones [9] define lying in terms action logic. Suppose that an agent a, by saying something or by sending a particular document, gets another agent b to believe that σ. Let Δ denote "the document or information is delivered to an agent b". Then, they define lying as $\neg B_a\, \sigma \wedge E_a(\Delta, B_b\sigma)$ which represents that a disbelieves σ, and that a makes b believe that σ by bringing it about that Δ. Their definition of lying is close to our definition of PDSC, but there are some important differences. First, in their definition the agent a does not believe $\neg\sigma$, while in our definition of PDSC a does $(B_a \neg \sigma)$. Second, the definition does not explicitly specify any connection between Δ and σ, while in PDSC the utterance of σ $(utter_{ab}(\sigma))$ brings it about that b's believing σ. Moreover, they do not formulate different categories of deception as done in this paper. In fact, their primary interest is to formulate *trust* of an agent, that is, "the fact that it is a who has sent the document counts, for an agent b, as grounds for accepting the truth of σ" [9]. To formulate this, they use the notion $E_a\Delta \Rightarrow_b \sigma$ meaning that "b trusts that a's bringing about Δ indicates the truth of σ". With this notion, they define a deceitful communicative action as $\neg B_a\, \sigma \wedge E_a B_b E_a \Delta \wedge B_a(E_a\Delta \Rightarrow_b \sigma)$, which does not guarantee successful lying on the part of a. In this paper, we do not consider such a trust relation between agents.

O'Neill [14] provides logical definitions of lies and deception based on the propositional modal logic of [6]. He defines intended deception as $Dec_{ab}\, \sigma = I_a B_b \sigma \wedge B_a \neg\sigma \wedge B_b\sigma$. It means that deception happens when a intends to make b believe a believed-false sentence σ and b believes it. However, the definition does not represent that b comes to have a false belief σ as a result of some behavior of a. According to his definition, a deceives b when b believes σ regardless of any behavior of a, which is absurd. The problem comes from the fact that the logic [6] does not have a mechanism of specifying action and causality.

6 Conclusion

In this paper, we have provided a formal account of Chisholm and Feehan's different notions of deception. We have done so using Pörn's *et al.*'s action logic. The logic turns out to be expressive enough in order to formally describe the various notions of deception of Chisholm and Feehan. We have also identified some formal properties and were able to formulate some postulates that an agent should try to satisfy, in its own interest as well as for moral reasons.

The notion of deception has been studied by a number of philosophers, while the topic has received relatively little attention in AI. The present study as well as our previous work [17] attempts to analyze the issue using a relatively simple logical formalization. Our aim is to turn conceptually defined notions in philosophy into a formally defined semantic problem in computational logic. In future work, we investigate computational methods of dishonest attitudes of agents and examine their applications. Detecting dishonest agents and preventing a success of deception are also topics for future research.

References

1. Adler, J. E.: Lying, deceiving, or falsely implicating. *J. Philosophy* 94(9), 435–452 (1997).
2. Caminada, M.: Truth, lies and bullshit, distinguishing classes of dishonesty. In: *Proc. IJCAI Workshop on Social Simulation* (2009).
3. Carson, T. L.: *Lying and deception: theory and practice*. Oxford University Press (2010).
4. Chisholm, R. M.: On the logic of intentional action. In: *Agent, Action and Reason*, R. Binkley *et al*. (eds), University of Toronto Press and Blackwell, pp. 38–69 (1971).
5. Chisholm, R. M. and Feehan, T. D.: The intent to deceive. *Journal of Philosophy* 74(3), 143–159 (1977).
6. Colombetti, M.: A modal logic of intentional communication. *Mathematical Social Sciences* 38, 171–196 (1999).
7. Ettinger, D. and Jehiel, P.: A theory of deception. *American Economic Journal: Microeconomics* 2(1): 1–20 (2010).
8. Firozabadi, B. S., Tan, Y. H. and Lee, R. M.: Formal definitions of fraud. In: P. McNamara and H. Prakken (eds). *Norms, Logics and Information Systems – New Studies in Deontic Logic and Computer Science*, pp. 275–288. IOS Press (1999).
9. Firozabadi, B. S. and Jones, A. J. I.: On the characterisation of a trusting agent – aspects of a formal approach. In: C. Castelfranchi and Y. H. Tan (eds.), *Trust and Deception in Virtual Societies*, Kluwer Academic Publishers, pp. 157–168 (2001).
10. Frankfurt, H. G.: *On Bullshit*. Princeton Univ. Press (2005).
11. Kupfer, J.: The moral presumption against lying. *Review of Metaphysics* 36, 103–126 (1982).
12. Mahon, J. E.: A definition of deceiving. *J. Applied Philosophy* 21(2), 181–194 (2007).
13. Mahon, J. E.: Two definitions of lying. *J. Applied Philosophy* 22(2), 211–230 (2008).
14. O'Neill, B.: A formal system for understanding lies and deceit. *Jerusalem Conference on Biblical Economics* (2003).
15. Pörn, I.: *Action Theory and Social Science*. Reidel, Dordrecht (1977).
16. Pörn, I.: On the nature of social order. In J. E. Fenstad et al. (eds.), *Logic, Methodology, and Philosophy of Science*, VIII. Elsevier (1989).
17. Sakama, C., Caminada, M. and Herzig, A.: A logical account of lying. In: *Proc. 12th European Conference on Logics in Artificial Intelligence, Lecture Notes in Artificial Intelligence* 6341, pp. 286–299 (2010).
18. Sandu, G.: Formal Logic of Action. Licentiate Thesis, University of Helsinki (1986).
19. Searle, J. R.: *Speech Acts*. Cambridge University Press (1969).
20. Staab, E. and Caminada, M.: On the profitability of incompetence. In: *Multi-Agent-Based Simulation XI, Lecture Notes in Computer Science* 6532, pp. 76–92 (2011).