A Logical Account of Lying

Chiaki Sakama[1], Martin Caminada[2], and Andreas Herzig[3]

[1] Wakayama University, Japan
sakama@sys.wakayama-u.ac.jp
[2] University of Luxembourg, Luxembourg
martin.caminada@uni.lu
[3] Université Paul Sabatier, France
herzig@irit.fr

Abstract. This paper aims at providing a formal account of *lying* – a dishonest attitude of human beings. We first formulate lying under propositional modal logic and present basic properties for it. We then investigate why one engages in lying and how one reasons about lying. We distinguish between offensive and defensive lies, or deductive and abductive lies, based on intention behind the act. We also study two weak forms of dishonesty, *bullshit* and *deception*, and provide their logical features in contrast to lying. We finally argue dishonesty postulates that agents should try to satisfy for both moral and self-interested reasons.

1 Introduction

Lying can be considered to be one of the basic behaviors of human beings. In spite of its familiarity to most of us, the question of "What is lying?" has been studied by a number of philosophers (for instance, [5, 9, 14] and references therein). Surprisingly, however, the topic has been almost completely ignored in artificial intelligence. There are several reasons why the study of lying is important in AI. First, lying is a linguistic behavior inherent to human beings, that requires intelligence and thinking. Studies on lying can thereby contribute to better understand human intelligence. Second, elucidating the mechanism of lying opens possibilities to develop computers that lie [16]. For instance, we can imagine a nurse robot who knows that a patient has a serious cancer but informs the patient that he/she is not in a serious state. Some potential applications of lying in AI and knowledge engineering are also addressed in [3, 20]. Third, lying is an act of social interaction. Hence, studying the act in the context of multiagent systems is necessary for designing intelligent agents. A recent study reports that an intelligent agent could behave dishonestly to win a debate in formal argumentation systems [4]. Lying has a distinctive feature as a speech act. According to Searle [19], a speech act is *sincere* if a speaker utters a believed-true sentence. This basic attitude is not applied to lying, that is, *a liar utters a believed-false sentence*. Saint Augustine, who was a Berber philosopher and theologian, says that "the heart of a liar is said to be double, that is, twofold in its thinking: one part consisting of that knowledge which he knows or thinks to be true, yet does not so express it; the other part consisting of that knowledge which he knows or thinks to be false, yet expresses as true" [2, p.55].

Providing a formal account of lying requires one to overcome various difficulties. First of all, there is no universally accepted definition of lying and even the definition

in the Oxford English Dictionary is problematic[4] [14]. Furthermore, formal logics are usually employed for formulating the truth of sentences and the correctness of inferences, whereas lies contradict the truth [22]. Thus, a formal account of lying is still an open and challenging topic in AI.

The purpose of this paper is to provide a logical account of lying. We formulate various forms of lies using propositional modal logic and investigate formal properties. We also characterize other types of dishonesty and compare them with lying. We propose basic postulates for dishonesty that agents should try to satisfy. The rest of this paper is organized as follows. Section 2 introduces a modal language for belief and intention and provides a logical framework of lying. Section 3 investigates different types of lying and argues their properties. Section 4 formulates *bullshit* and *deception* as weaker forms of dishonesty. Section 5 discusses related issues and Section 6 concludes the paper.

2 Liars' Logic

2.1 A Simple Logic for Belief and Intention

In this paper, we consider a propositional modal logic of intentional communication [7]. A propositional modal language L_0 is built from a finite set of propositional constants $\{p, q, r, \ldots\}$ on the logical connectives $\neg$, $\vee$, $\wedge$, $\supset$, $\equiv$, and on two families of modal operators, $(B_a)_{a\in A}$ and $(I_a)_{a\in A}$, where A is a finite set of agents. Well-formed formulas (or *sentences*) in L_0 are defined as usual as those belonging to a multi-modal propositional logic. Sentences in L_0 will be denoted by the small Greek letters, and parentheses are employed as usual to clarify the structure of sentences. $\top$ and $\bot$ represent valid and contradictory sentences, respectively. The set of all sentences in L_0 is denoted by Φ and $\Phi^\star = \Phi \setminus \{\top, \bot\}$. A finite set of sentences is identified with the conjunction of all sentences included in the set. The intuitive reading of $B_a\phi$ and $I_a\phi$ are that an agent a believes that ϕ and intends that ϕ, respectively. A Kripkean semantics is defined for L_0, although we omit the details here.[5] A logic BI_0 is defined over L_0, that is an extension of KD45$_n$ [11] and has the following axioms and inference rules:

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

($\mathbf{K_B}$) $B_a\phi \wedge B_a(\phi \supset \psi) \supset B_a\psi$ and ($\mathbf{K_I}$) $I_a\phi \wedge I_a(\phi \supset \psi) \supset I_a\psi$.

($\mathbf{D_B}$) $B_a\phi \supset \neg B_a\neg\phi$ and ($\mathbf{D_I}$) $I_a\phi \supset \neg I_a\neg\phi$.

($\mathbf{4_B}$) $B_a\phi \supset B_aB_a\phi$ and ($\mathbf{4_{IB}}$) $I_a\phi \supset B_aI_a\phi$.

($\mathbf{5_B}$) $\neg B_a\phi \supset B_a\neg B_a\phi$ and ($\mathbf{5_{IB}}$) $\neg I_a\phi \supset B_a\neg I_a\phi$.

$$(\mathbf{MP})\ \frac{\phi \quad \phi \supset \psi}{\psi}, \qquad (\mathbf{N_B})\ \frac{\phi}{B_a\phi}, \qquad (\mathbf{N_I})\ \frac{\phi}{I_a\phi}.$$

To represent a speech act of an agent, we introduce the unary predicate $utter_{xy}$ defined over sentences in L_0 with $x, y \in A$. An expression $utter_{ab}(\sigma)$ means that an agent a expresses a sentence σ to another agent b. A language L_0^U is defined as L_0

[4] The OED definition of lying is: to make a false statement with the intention to deceive.

[5] Informally speaking, $B_a\phi$ (resp. $I_a\phi$) holds iff ϕ is true in all states of affairs compatible with a's current beliefs (resp. intentions).

together with the predicate $utter_{xy}$. If an agent utters something, he/she intends the speech act and is aware of his/her utterance. This is expressed by the next axiom:

$$(\mathbf{U_{IB}}) \quad utter_{ab}(\sigma) \supset I_a(utter_{ab}(\sigma)) \wedge B_a(utter_{ab}(\sigma)).$$

The system BI_0^U, defined over L_0^U, is the weakest extension of BI_0 containing the axiom $(\mathbf{U_{IB}})$. If a sentence ϕ is a theorem of BI_0^U, we write $\vdash \phi$. An agent a has a *knowledge base* K_a as a finite set of believed-true sentences from L_0^U. Each agent believes that other agents follow the same logic BI_0^U in their beliefs and intentions. Thus, $B_aB_b\phi \supset B_a\neg B_b\neg\phi$ and $B_a(I_b\phi \wedge I_b(\phi \supset \psi)) \supset B_aI_b\psi$, for instance. Given two sentences σ and λ in Φ, we write $\sigma \succeq \lambda$ if $\vdash \sigma \supset \lambda$. In this case, we say that σ is *stronger than or equal to* λ (or λ is *weaker than or equal to* σ). We write $\sigma \succ \lambda$ if $\sigma \succeq \lambda$ and $\lambda \not\succeq \sigma$, and say that σ is *stronger than* λ (or λ is *weaker than* σ).

2.2 Lying

Lying can be seen as a speech act of an agent (a speaker) towards another agent (a hearer). For our purpose, we will use a relatively simple definition of lying which seems to be well-accepted in the literature.

> *To lie (to another person) is: to make a believed-false statement (to another person) with the intention that that statement be believed to be true (by the other person).* – [12] and (L6) of [14]

We can then provide a formal definition of lying in L_0^U as follows.

Definition 2.1 (lie) Let a and b be two agents and $\sigma \in \Phi$. Then, define

$$LIE_{ab}(\sigma) \stackrel{def}{=} utter_{ab}(\sigma) \wedge B_a\neg\sigma \wedge I_aB_b\sigma. \tag{1}$$

In this case, we say that a *lies* to b on the sentence σ. σ is also called a *lie*.

By the definition, a lies to b if a utters a believed-false sentence σ to b with the intention that σ is believed by b. Some researchers argue that lying does not necessarily require the use of words [14], but here we consider lying as a statement of a sentence. Note also that the speaker believes $\neg\sigma$, but that the truth of $\neg\sigma$ is not actually required. That is,"a person is to be judged as lying or not lying according to the intention of his own mind, not according to the truth or falsity of the matter itself" [2, p.55]. Lying is not simply saying something that one believes to be false, but involves an intention to deceive. Thus, if one says something manifestly false as a joke or a metaphor, it is not a lie.[6] Lying on valid or contradictory sentences is meaningless.

Proposition 2.1 $\vdash LIE_{ab}(\top) \supset \bot$ *and* $\vdash LIE_{ab}(\bot) \supset \bot$.

Proof. $LIE_{ab}(\top)$ implies $B_a\bot$ that implies $\neg B_a\top$ $(\mathbf{D_B})$, while $\top$ implies $B_a\top$ $(\mathbf{N_B})$. Contradiction. Next, $LIE_{ab}(\bot)$ implies $I_aB_b\bot$, while $B_b\top$ implies $\neg B_b\bot$ $(\mathbf{D_B})$ that implies $I_a\neg B_b\bot$ $(\mathbf{N_I})$ then $\neg I_aB_b\bot$ $(\mathbf{D_I})$. Contradiction. □

If an agent lies, he/she is aware of his/her dishonest act.

Proposition 2.2 $\vdash LIE_{ab}(\sigma) \supset B_a(LIE_{ab}(\sigma))$ *for any* $\sigma \in \Phi$.

Proof. The result holds by Def. 2.1(1) and the axioms $(\mathbf{U_{IB}})$, $(\mathbf{4_B})$, and $(\mathbf{4_{IB}})$. □

Lying to oneself leads to contradiction.[7]

[6] Some philosophers argue that an intention to deceive is not a necessary condition of lying, however [5].

[7] "In short, self-deception involves an inner conflict, perhaps the existence of contradiction" [8].

Proposition 2.3 $\vdash LIE_{aa}(\sigma) \supset \bot$ *for any* $\sigma \in \Phi$.

Proof. $LIE_{aa}(\sigma)$ implies $B_a\neg\sigma \wedge I_a B_a \sigma$. $B_a\neg\sigma$ implies $\neg B_a\sigma$ ($\mathbf{D_B}$), which implies $I_a\neg B_a\sigma$ ($\mathbf{N_I}$). On the other hand, $I_a B_a\sigma$ implies $\neg I_a\neg B_a\sigma$ ($\mathbf{D_I}$). Contradiction. □

Note that when lying, a speaker does in general not care about the belief state of the hearer. If a speaker a believes that a hearer b believes σ, $LIE_{ab}(\sigma)$ would have the effect of strengthening the incorrect belief of the hearer. On the other hand, if a believes that b disbelieves σ, $LIE_{ab}(\sigma)$ might cause belief revision of the hearer.

3 Various Forms of Lying

3.1 Offensive Lie vs. Defensive Lie

One has motives for lying and several reasons are considered behind the act. Here we consider two typical cases. First, one lies to have a positive (or wanted) outcome that would not be gained by telling the truth. Second, one lies to avoid a negative (or unwanted) outcome that would happen when telling the truth. An example of the first case is that a salesperson lies about the quality of a product, which leads a customer to make a (wrong) decision of buying the product. An example of the second case is that a child lies about his/her good performance in the exam to avoid punishment by his/her parents. We say that the first case of lying is *offensive*, while the second case is *defensive*. A *positive outcome* or a *negative outcome* is an effect expected by a speaker with respect to the result of reasoning by a hearer. Thus, in offensive/defensive lying a speaker reasons about what a hearer believes in the context of discourse.

Definition 3.1 (offensive/defensive lie) Let a and b be two agents and $\sigma, \phi, \psi \in \Phi$. Then, define

$$O\text{-}LIE_{ab}(\sigma,\phi) \stackrel{def}{=} I_a B_b\,\phi \wedge \neg B_a B_b(\neg\sigma \supset \phi) \wedge B_a B_b(\sigma \supset \phi) \wedge LIE_{ab}(\sigma). \quad (2)$$

In this case, a *offensively lies* to b on σ to have the positive outcome ϕ. σ is also called an *offensive lie* for ϕ. Next, define

$$D\text{-}LIE_{ab}(\sigma,\psi) \stackrel{def}{=} I_a\neg B_b\,\psi \wedge \neg B_a\neg B_b(\neg\sigma \wedge \psi) \wedge B_a\neg B_b(\sigma \wedge \psi) \wedge LIE_{ab}(\sigma). \quad (3)$$

In this case, a *defensively lies* to b on σ to avoid the negative outcome ψ. σ is also called a *defensive lie* for ψ.

Intuitive meanings of the definition are as follows. In (2), a offensively lies on σ if a has an intention to make b believe ϕ. And a disbelieves that the believed-true sentence $\neg\sigma$ leads b to believe a positive outcome ϕ, while a believes that the believed-false sentence σ does. With these conditions, a lies to b on σ. In (3) a defensively lies on σ if a has an intention to make b disbelieve ψ. And a considers it possible that b believes that the believed-true sentence $\neg\sigma$ and a negative outcome ψ hold at the same time, while a does not consider it possible that b believes that the believed-false sentence σ and ψ hold simultaneously. With these conditions, a lies to b on σ. As a special case, a may lie to b on the sentence ϕ (resp. $\neg\psi$) to make b believe ϕ (resp. disbelieve ψ).

Proposition 3.1 *Let a and b be two agents and $\phi, \psi \in \Phi$.*

(i) $O\text{-}LIE_{ab}(\phi, \phi) \equiv \neg B_a B_b\, \phi \wedge LIE_{ab}(\phi)$.
(ii) $D\text{-}LIE_{ab}(\neg\psi, \psi) \equiv \neg B_a \neg B_b\, \psi \wedge LIE_{ab}(\neg\psi)$.

Proof. The result (i) directly follows by the definition. (ii) also follows by the fact that $I_a \neg B_a \psi$ is implied by $I_a B_a \neg\psi$ of $LIE_{ab}(\neg\psi)$ by $(\mathbf{D_B})$, $(\mathbf{N_I})$ and $(\mathbf{K_I})$. □

In $O\text{-}LIE_{ab}(\phi, \phi)$, the condition $\neg B_a B_b\, \phi$ means that a has motives for offensive lying when a disbelieves that b believes the positive outcome ϕ. In $D\text{-}LIE_{ab}(\neg\psi, \psi)$, the condition $\neg B_a \neg B_b\, \psi$ means that a has motives for defensive lying when a considers it possible that b believes the negative outcome ψ. Thus, the definitions of offensive and defensive lies are stronger than the definition of lies of Definition 2.1. This is due to the fact that offensive (resp. defensive) lying has additional objectives to have positive outcomes (resp. avoid negative outcomes), while lying in general is only aimed at making the hearer believe the uttered statement itself.

Example 3.1 Suppose that a salesperson a is dealing with a customer b, and that a is requested to provide b with information about the quality of the product. The salesperson believes $\neg\, high\_quality$ and also believes that the customer has the knowledge base $K_b = \{\, high\_quality \supset buy \,\}$. When the salesperson has the positive outcome $\phi = buy$, telling the true belief does not lead b to buy the product. In this case, a offensively lies to b on $\sigma = high\_quality$. Next, suppose that a child a and his/her mother b talk about examination, and a is requested to provide b with information about the score and the rank. The child believes $\neg(high\_score \wedge high\_rank)$, and also believes that mother has the knowledge base $K_b = \{\, \neg\, (high\_score \wedge high\_rank) \supset punish \,\}$. When the child has the negative outcome $\psi = punish$, telling the true belief leads b to punish a. In this case, a defensively lies to b on $\sigma = high\_score \wedge high\_rank$.

When an agent lies to another agent, the success of the act depends on the belief state or knowledgeability of the hearer. For instance, it is easier to mislead children than adults, and it is more difficult to mislead experts than novices. We next consider how different degrees of lies are used depending on a hearer's belief state that is believed by a speaker. An agent b is *knowledgeable not less than* another agent c if $B_c\, \phi \supset B_b\, \phi$ holds for any formula $\phi \in \Phi$. Suppose that there are three agents a, b and c, and a believes that b is knowledgeable not less than c. Then, in offensive lying Def.3.1(2), (i) $\neg B_a B_b (\neg\sigma \supset \phi)$ implies $\neg B_a B_c (\neg\sigma \supset \phi)$, and (ii) $B_a B_b (\sigma \supset \phi)$ does not imply $B_a B_c (\sigma \supset \phi)$. By (i) if a disbelieves that a positive outcome ϕ is not gained by telling the believed-true sentence $\neg\sigma$ to b, then a also has the same disbelief for c. This means that a's motive of offensively lying to c is not less than the motive of offensively lying to b. By (ii) even if a believes that a lie σ leads b to a positive outcome ϕ, a does not believe that the same lie leads c to ϕ. This means that, to have a positive outcome ϕ from c, a has to craft a lie that is not weaker than σ in general. In case of defensive lying Def.3.1(3), (iii) $\neg B_a \neg B_b (\neg\sigma \wedge \psi)$ does not imply $\neg B_a \neg B_c (\neg\sigma \wedge \psi)$, and (iv) $B_a \neg B_b (\sigma \wedge \psi)$ implies $B_a \neg B_c (\sigma \wedge \psi)$. By (iii) even if a considers it possible that b believes that the believed-true sentence $\neg\sigma$ and a negative outcome ψ hold simultaneously, a does not have the same belief for c. This means that a's motive of defensively lying to c is not more than the motive of defensively lying to b. By (iv) if a does not consider it possible

that b believes the believed-false sentence σ and ψ hold simultaneously, then a also has the same belief for c. This means that, to avoid a negative outcome ψ from c, a can craft a lie that is not stronger than σ in general. We next formulate the situation.

Let σ be an offensive lie for a positive outcome ϕ. If $\sigma' \succeq \sigma$ implies $\sigma \succeq \sigma'$ for any offensive lie σ' for ϕ, then σ is called a *strongest* offensive lie (denoted by σ_s). By contrast, if $\sigma \succeq \sigma'$ implies $\sigma' \succeq \sigma$ for any offensive lie σ' for ϕ, then σ is called a *weakest* offensive lie (denoted by σ_w). The notion of the strongest/weakest defensive lies is similarly defined.

Proposition 3.2 *Suppose that there are three agents a, b and c, and a believes that b is knowledgeable not less than c. Let ϕ and ψ be sentences in Φ. Then,*

(i) $\vdash (O\text{-}LIE_{ab}(\sigma_w, \phi) \wedge O\text{-}LIE_{ac}(\lambda, \phi)) \supset \bot$ *for any* $\lambda \in \Phi$ *such that* $\sigma_w \succ \lambda$.
(ii) $\vdash (D\text{-}LIE_{ab}(\sigma_s, \psi) \wedge D\text{-}LIE_{ac}(\lambda, \psi)) \supset \bot$ *for any* $\lambda \in \Phi$ *such that* $\lambda \succ \sigma_s$.

Proof. (i) Suppose that $O\text{-}LIE_{ab}(\sigma_w, \phi) \wedge O\text{-}LIE_{ac}(\lambda, \phi)$ and $\sigma_w \succ \lambda$ hold. As a believes that b is knowledgeable not less than c, $B_a B_c(\lambda \supset \phi)$ implies $B_a B_b(\lambda \supset \phi)$ $(*)$. Next, assume that $B_a B_b(\neg\lambda \supset \phi)$ $(**)$. By $(*)$ and $(**)$, it holds that $B_a B_b\, \phi$, which implies $B_a B_b(\neg\sigma_w \supset \phi)$ $(\dagger)$. As $O\text{-}LIE_{ab}(\sigma_w, \phi)$, it holds that $\neg B_a B_b(\neg\sigma_w \supset \phi)$ which contradicts $(\dagger)$. So $\neg B_a B_b(\neg\lambda \supset \phi)$ $(\ddagger)$. The facts $(*)$ and $(\ddagger)$ imply that a can offensively lie to b on the sentence λ for the outcome ϕ. As σ_w is the weakest lie, $\sigma_w \not\succ \lambda$. Contradiction. (ii) is proved in a similar way. □

Example 3.2 (cont. Example 3.1) Suppose that the salesperson a deals with another customer c. a notices that c is more cautious than b in making decisions, and believes that c will buy the product if it is valuable as well as good in quality. But a believes that the product is neither of these $\neg high\_quality \wedge \neg valuable$, and also believes that $K_c = \{\, (high\_quality \wedge valuable) \supset buy \,\}$ where $B_c(K_c) \supset B_b(K_c)$ holds. To have the positive outcome $\phi = buy$, a has to lie offensively on the sentence $\lambda = high\_quality \wedge valuable$, which is stronger than σ, to convince c to buy the product. Next, suppose that a child a has a dialogue with his/her father c. a knows that father is more generous than mother, and believes that he is only concerned about the score. But a believes $\neg\, high\_score$, and also believes that $K_c = \{\, \neg high\_score \supset punish \,\}$ where $B_c(K_c) \supset B_b(K_c)$ holds. To avoid the negative outcome $\psi = punish$, a lies defensively on the sentence $\lambda = high\_score$, which is weaker than σ, to persuade father not to punish him/her.

3.2 Deductive Lie vs. Abductive Lie

By an offensive lie (resp. a defensive lie), a speaker intends to mislead a hearer to deduce a wrong conclusion (resp. not to deduce a right conclusion). We call these types of lies *deductive lies*. By contrast, a person often lies in order to block another person for generating assumptions. For instance, suppose a man, say, Sam, who is coming home late because he is cheating on his wife. Based on the observation "Sam arrives late", his wife could perform *abduction* and one of the possible explanations would be "Sam cheats on his wife". Sam, of course, does not want this abduction to take place, so he lies about a possible other reason, "I had to do overtime at work". Sam's hope is that once his wife has this incorrect information, her abductive reasoning process will stop.

She will no longer continue possible abduction, and will never even be aware of the possibility of Sam's cheating on her (if she trusts her husband).

Abduction is the process of forming an explanatory hypothesis from an observation [18]. Formally, let o be a sentence representing an *observation* and H a set of sentences representing a *hypothesis*. Given a knowledge base K and an observation o, a hypothesis H *explains* o in K if $K \wedge H \vdash o$ where $K \wedge H$ is consistent. An agent lies to interrupt abduction (by another agent) that produces an unwanted explanation for him/her. Let Σ_a ($\subseteq K_a$) be a set of sentences (called a *secret set*) which an agent a wants to conceal from another agent b. Abductive lie is then defined as follows.

Definition 3.2 (abductive lie) Let a and b be two agents and $o \in \Phi \setminus \Sigma_a$ a sentence observed by them. Also, let $\sigma \in \Phi \setminus \Sigma_a$ such that $\sigma \not\equiv o$. Then, define

$$A\text{-}LIE_{ab}(\sigma, o) \stackrel{def}{=} B_a\, o \wedge B_a \neg B_b(\Delta \supset o) \wedge B_a(B_b(\Gamma \supset o) \wedge \neg B_b \neg \Gamma) \\ \wedge B_a(B_b(\sigma \supset o) \wedge \neg B_b \neg \sigma) \wedge \bigwedge_{\gamma \in \Sigma_a} I_a \neg B_b\, \gamma \wedge LIE_{ab}(\sigma) \quad (4)$$

where Δ is any subset of $K_a \setminus \Sigma_a$ and Γ ($\subseteq \Phi$) is a set of sentences such that $\Gamma \cap \Sigma_a \neq \emptyset$. In this case, we say that the agent a *abductively lies* to another agent b on the sentence σ. σ is also said an *abductive lie* for the observation o.

In (4), $B_a\, o$ represents that a believes o. $B_a \neg B_b(\Delta \supset o)$ implies $B_a \neg B_b\, o$, so that a believes that b requires some explanation once he/she observes o. However, a believes that b does not explain o by believed-true sentences of a without some secret sentences (i.e., $B_a \neg B_b(\Delta \supset o)$). a believes that b explains o by either using some secret sentences of a (i.e., $B_a(B_b(\Gamma \supset o) \wedge \neg B_b \neg \Gamma)$ or some believed-false sentence σ of a (i.e., $B_a(B_b(\sigma \supset o) \wedge \neg B_b \neg \sigma)$), but a does not want b's believing any sentence γ in Σ_a (i.e., $\bigwedge_{\gamma \in \Sigma_a} I_a \neg B_b\, \gamma$). In this case, a abductively lies to b on σ for explaining o. Note that $\vdash A\text{-}LIE_{ab}(\sigma, \top) \supset \bot$ and $\vdash A\text{-}LIE_{ab}(\sigma, \bot) \supset \bot$ for any σ.

Example 3.3 Suppose that Sam has the knowledge base $K_a = \{\, cheat,\ \neg overtime,\ cheat \supset late,\ \ overtime \supset late \,\}$, and believes that his wife has the knowledge base $K_b = \{\, cheat \supset late,\ \ overtime \supset late \,\}$. Let $\Sigma_a = \{cheat\}$, that is, Sam wants to keep his cheating behavior secret. Given the observation $o = late$, Sam believes that his wife can abduce $\Gamma = \{cheat\}$ as a possible explanation for o. Then, Sam abductively lies on $\sigma = overtime$ which explains his late arrival and would stop her abducing the explanation $cheat$.

The effect of an abductive lie also depends on the belief state of a hearer. If a believes that an agent b is knowledgeable not less than another agent c, then the condition $B_a \neg B_c(\Delta \supset o)$ in $A\text{-}LIE_{ac}(\sigma, o)$ holds, while $B_a(B_c(\Gamma \supset o) \wedge \neg B_c \neg \Gamma)$ and $B_a(B_c(\sigma \supset o) \wedge \neg B_c \neg \sigma)$ do not necessarily hold. This means that a's motive of abductively lying to c is not more than the motive of abductively lying to b. For instance, if Sam believes that his daughter has the knowledge base $K_c = \{\, overtime \supset late \,\}$, then $B_a \neg B_c(cheat \supset late)$ and Sam does not need to lie her. When a abductively lies to c, however, a has to craft a lie that is not weaker than σ in general. If $K'_c = \{\, cheat \supset late \,\}$, then Sam has to make the stronger lie $\lambda = overtime \wedge (overtime \supset late)$, for instance. Given an observation o, the notion of a weakest abductive lie σ_w is defined in a way similar to a weakest offensive lie. Then we have the next result.

Proposition 3.3 *Suppose that there are three agents a, b and c, and a believes that b is knowledgeable not less than c. Let o be a sentence in $\Phi \setminus \Sigma_a$. Then,*
$\vdash (A\text{-}LIE_{ab}(\sigma_w, o) \wedge A\text{-}LIE_{ac}(\lambda, o)) \supset \bot$ *for any $\lambda \in \Phi$ such that $\sigma_w \succ \lambda$.*

Proof. Similar to the proof of Proposition 3.2(1). □

3.3 What are the Most Effective Lies?

In deductive lying and abductive lying, a number of candidate lies exist to achieve a speaker's goal. Then a question is how good liars select "best lies". As observed in Propositions 3.2 and 3.3, a speaker can select different degrees of lies according to the knowledgeability of a hearer. A stronger lie would be needed to have a positive outcome from a less knowledgeable hearer, while a weaker lie would be enough to avoid a negative outcome from the same hearer. A liar normally wants to keep his/her lie as small as possible. This is because, "The lie, to his immediate advantage, often results in an overall net loss of freedom in what he can do or say... The need to maintain the deception binds him" [12, p.119]. A stronger lie makes the liar less free, which he wants to avoid anyway. Besides, lies make the belief state of a hearer deviate from the objective reality (or, at least from the reality as believed by a speaker) and a stronger lie would increase such deviation. This is undesirable for a speaker because it increases the chance of the lie being detected. The best lie is a lie that does not have too much "collateral damage" on a hearer. We state a guideline for agents to satisfy in lying as the next postulate. Let λ, σ, o, ϕ, $\psi \in \Phi^\star$ and $\sigma \succeq \lambda$. Then, we have the next postulate.

Postulate I: Never tell an unnecessarily strong lie.
(i) $B_a(O\text{-}LIE_{ab}(\sigma, \phi) \supset B_b\phi) \wedge B_a(O\text{-}LIE_{ab}(\lambda, \phi) \supset B_b\phi) \supset \neg O\text{-}LIE_{ab}(\sigma, \phi)$.
(ii) $B_a(D\text{-}LIE_{ab}(\sigma, \psi) \supset \neg B_b\psi) \wedge B_a(D\text{-}LIE_{ab}(\lambda, \psi) \supset \neg B_b\psi) \supset \neg D\text{-}LIE_{ab}(\sigma, \psi)$.
(iii) $B_a(A\text{-}LIE_{ab}(\sigma, o) \supset B_b\, o) \wedge B_a(A\text{-}LIE_{ab}(\lambda, o) \supset B_b\, o) \supset \neg A\text{-}LIE_{ab}(\sigma, o)$.

4 Weak Form of Dishonesty

4.1 Bullshit

Frankfurt [10] studies a category of dishonesty, called *bullshit*, that is different from lies. Bullshit is a statement that "is grounded neither in a belief that it is true nor, as a lie must be, in a belief that it is not true" (ibid., p.33). As an example, consider a financial consultant paid by the hour to provide advice to his clients. The consultant gives advice to buy stocks, for instance, but he may or may not believe that buying stocks is the best strategy (due to the lack of expertise). Bullshit is a quite common phenomenon in daily life. Frankfurt states a reason for its occurrence as follows: "Bullshit is unavoidable whenever circumstances require someone to talk without knowing what he is talking about. Thus the production of bullshit is stimulated whenever a person's obligations or opportunities to speak about some topic exceed his knowledge of the facts that are relevant to that topic" (ibid., p.63). Bullshit can formally be defined as follows.

Definition 4.1 (bullshit) Let a and b be two agents and $\sigma \in \Phi$. Then,

$$BS_{ab}(\sigma) \stackrel{def}{=} utter_{ab}(\sigma) \wedge \neg B_a\sigma \wedge \neg B_a\neg\sigma. \tag{5}$$

In this case, we say that an agent a *bullshits* to another agent b on the sentence σ. σ is also called *bullshit* (shortly, BS).

In lying Def.2.1(1), the speaker a disbelieves σ but believes $\neg\sigma$. When bullshitting Def.4.1(5), on the other hand, a disbelieves $\neg\sigma$ either. In other words, a has no belief with respect to the truth value of σ. So one cannot bullshit about one's own beliefs.

Proposition 4.1 $\vdash BS_{ab}(B_a\sigma) \supset \bot$ *and* $\vdash BS_{ab}(\neg B_a\sigma) \supset \bot$ *for any* $\sigma \in \Phi$.

Proof. Both $BS_{ab}(B_a\sigma)$ and $BS_{ab}(\neg B_a\sigma)$ imply $\neg B_a B_a\sigma \wedge \neg B_a \neg B_a\sigma$. Here $\neg B_a B_a\sigma$ implies $\neg B_a\sigma$ ($\mathbf{4_B}$), which implies $B_a \neg B_a\sigma$ ($\mathbf{N_B}$). This contradicts $\neg B_a \neg B_a\sigma$. □

Bullshitting on valid or contradictory sentences is meaningless.

Proposition 4.2 $\vdash BS_{ab}(\top) \supset \bot$ *and* $\vdash BS_{ab}(\bot) \supset \bot$.

Proof. Both $BS_{ab}(\top)$ and $BS_{ab}(\bot)$ imply $\neg B_a \top$, but $\top$ implies $B_a \top$ ($\mathbf{N_B}$). □

Like lying, a bullshitter notices his/her act.

Proposition 4.3 $\vdash BS_{ab}(\sigma) \supset B_a(BS_{ab}(\sigma))$ *for any* $\sigma \in \Phi$.

There are some differences between lies and BS. First, bullshitting to oneself $BS_{aa}(\sigma)$ is possible in general. Second, $BS_{ab}(\sigma)$ does not contradict the belief of the speaker a. These facts imply that one cannot lie and bullshit on the same sentence.

Proposition 4.4 $\vdash LIE_{ab}(\sigma) \wedge BS_{ab}(\sigma) \supset \bot$ *for any* $\sigma \in \Phi$.

Proof. $LIE_{ab}(\sigma)$ implies $B_a\neg\sigma$, while $BS_{ab}(\sigma)$ implies $\neg B_a\neg\sigma$. □

Another important difference is that BS does not require the intention of a speaker a to make a hearer b believe σ. In the above example, the financial consultant has no interest in making the client believe that buying stocks is the best strategy or not. The only concern of the consultant is that the client believes that the statement is based on financial expertise. Since a has no belief with respect to σ, there is a freedom for a speaker to utter σ or $\neg\sigma$. The most effective BS is the one that is coherent with the speaker's belief. The choice whether to utter σ or $\neg\sigma$ is also decided by how likely it will be for a hearer to believe one of them (given some additional explanation). This is in contrast to lying where speakers have no freedom to make this choice because one of these options (either σ or $\neg\sigma$) will have consequences they might want to enjoy (or which they might want to avoid). A liar usually has an interest in creating a particular belief at a hearer. This is not always the case for BS, however.

On the other hand, there is BS that accompanies some intention. For instance, suppose a salesperson who is paid on commission basis, but does not really know the products that he is selling. The salesperson would make the claim that a product has a high quality, without having any knowledge on this. This is also an example of BS. However, making a client believe that the product has a high quality is preferred to making the client believe that the product has a low quality. The situation here differs from that of the financial consultant mentioned above (who is paid by the hour by the client, and hence has no intrinsic interest to advise to buy stocks or not). Such *intentional bullshit* is defined as

$$I\text{-}BS_{ab}(\sigma) \stackrel{def}{=} BS_{ab}(\sigma) \wedge I_a B_b\sigma. \tag{6}$$

By contrast, $BS_{ab}(\sigma)$ without $I_a B_b\sigma$ is called *unintentional*. In this paper, we will ignore this difference in cases where it is unimportant. Intentional BS (6) is similar to lies, so that offensive/defensive or deductive/abductive intentional BS could be considered. Different from unintentional BS, intentional BS to oneself is inconsistent.

Proposition 4.5 $\vdash I\text{-}BS_{aa}(\sigma) \supset \bot$ *for any* $\sigma \in \Phi$.

Proof. Similar to the proof of Proposition 2.3. □

Next we consider what is best BS. Any BS is dishonest, but the consequences of faking are generally less severe for a weak bullshitter than for a strong bullshitter. Suppose a salesperson who bullshits for selling specified products, say $high\_quality$, that is weaker than the bullshit $high\_quality \wedge valuable$. If a customer decides to buy the product, the strong bullshitter would be responsible for the value as well as the quality. Getting more responsibility is undesirable for a bullshitter anyway. We formulate the situation for *offensive* intentional BS. For $\sigma, \phi \in \Phi$, let us define

$$O\text{-}BS_{ab}(\sigma, \phi) \stackrel{def}{=} I_a B_b \phi \wedge \neg B_a B_b(\neg\sigma \supset \phi) \wedge B_a B_b(\sigma \supset \phi) \wedge I\text{-}BS_{ab}(\sigma).$$

Then we have the next postulate for BS.

Postulate II: Never tell unnecessarily strong BS. Let $\lambda, \sigma, \phi \in \Phi^\star$ and $\sigma \succeq \lambda$. Then, $B_a(O\text{-}BS_{ab}(\sigma, \phi) \supset B_b\phi) \wedge B_a(O\text{-}BS_{ab}(\lambda, \phi) \supset B_b\phi) \supset \neg O\text{-}BS_{ab}(\sigma, \phi)$.

Similar postulates are considered for defensive or abductive intentional BS. Lies and BS are two different forms of dishonesty, but lies are considered more sinful than BS.[8] This is because a liar intentionally implants wrong beliefs at the hearer, while a bullshitter spits out statements, intentionally or not, without knowing if they are true. As a result, "people do tend to be more tolerant of bullshit than of lies, perhaps because we are less inclined to take the former as a personal affront" [10, p.50]. This leads us to the next postulate.

Postulate III: Never lie if you can bullshit your way out of it. Let $\lambda, \sigma, \phi \in \Phi^\star$. Then, $B_a(O\text{-}BS_{ab}(\sigma, \phi) \supset B_b\phi) \wedge B_a(O\text{-}LIE_{ab}(\lambda, \phi) \supset B_b\phi) \supset \neg O\text{-}LIE_{ab}(\lambda, \phi)$.

4.2 Deception

Another form of dishonesty which we consider here is *deception*. There is no universally agreed definition of deception [6, 13], so we consider the one argued in [1]. Different from lying, there is no untruthfulness condition in deception. That is, a speaker makes a believed-true statement with the intention that a hearer misuses it to reach a wrong conclusion. For instance, John, who wants to marry his girlfriend Mary, tells her that he got a job at a company. Mary then considers that John has a stable income now and would agree to marry him. The company is almost bankrupt, however, and John believes that he would not get a stable income. But John does not tell Mary that his company is going bankrupt. In this speech act, John is telling the truth, while he expects that Mary will reach a conclusion "stable income" which he believes to be false. Thus, different from lies or BS, 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.[9] Caminada [4] captures the point as "With deception, one makes use of the *nonmonotonic* inference capabilities of the other person in order to implant wrong beliefs, without having to resort to lying ourselves". In the above example, John believes that Mary has the belief

[8] Some philosophers consider that bullshit is a class of lies [5, cf. L5].
[9] Some philosophers call this a "lie of omission" [14].

"$B_m((get\_job \wedge \neg B_m \neg stable) \supset stable)$". John then intends to make Mary believe get_job, while withholding $\neg stable$, which would result in Mary's believing $stable$. This is the effect of *default reasoning*. Now deception is formulated as follows.

Definition 4.2 (deception) Let a and b be two agents and $\delta, \sigma \in \Phi$ such that $\delta \not\equiv \sigma$. Then, define

$$DEC_{ab}(\sigma, \delta) \stackrel{def}{=} utter_{ab}(\sigma) \wedge B_a\sigma \wedge I_a B_b\sigma \wedge B_a B_b((\sigma \wedge \neg B_b \neg\delta) \supset \delta) \wedge B_a \neg B_b \neg\delta \wedge B_a \neg\delta \wedge I_a B_b\delta. \quad (7)$$

In this case, we say that an agent a *deceives* another agent b on the sentence σ. σ is also called *deception*.

In (7), the speaker a utters a believed-true sentence σ with the intention of making a hearer b believe it (i.e., $utter_{ab}(\sigma) \wedge B_a\sigma \wedge I_a B_b\sigma$). a believes that b uses σ to reach a default conclusion δ (i.e., $B_a B_b((\sigma \wedge \neg B_b\neg\delta) \supset \delta)$). a also believes that b disbelieves the falsity of δ (i.e., $B_a \neg B_b \neg\delta$), while a believes it (i.e., $B_a \neg\delta$). And believing δ by the hearer b is what the speaker a intends to achieve (i.e., $I_a B_b\delta$). Note that nonmonotonicity arises in $B_b((\sigma \wedge \neg B_b\neg\delta) \supset \delta)$. Compared with definitions of lies and bullshit, one can observe that the act of deception is a bit complicated. In fact, "The deceiver takes a more circuitous route to his success, where lying is an easier and more certain way to mislead" [1, p.440]. A reason for the complication is due to the fact that deception works by nonmonotonic reasoning.

Like lying and $I\text{-}BS$, the following properties hold.

Proposition 4.6 $\vdash DEC_{ab}(\bot, \delta) \supset \bot$ *for any* $\delta \in \Phi$.

Proposition 4.7 $\vdash DEC_{ab}(\sigma, \delta) \supset B_a(DEC_{ab}(\sigma, \delta))$ *for any* $\sigma, \delta \in \Phi$.

Proposition 4.8 $\vdash DEC_{aa}(\sigma, \delta) \supset \bot$ *for any* $\sigma, \delta \in \Phi$.

In contrast to lying and BS, $DEC_{ab}(\top, \delta)$ is consistent. In fact, it becomes

$$DEC_{ab}(\top, \delta) = utter_{ab}(\top) \wedge B_a B_b(\neg B_b\neg\delta \supset \delta) \wedge B_a \neg B_b \neg\delta \wedge B_a \neg\delta \wedge I_a B_b\delta.$$

In this case, a deceiver utters no meaningful information and just expects a hearer to reach a default conclusion δ. Different from lying and BS, a deceiver utters believed-true sentences. This implies that one cannot lie and deceive, nor bullshit and deceive, on the same sentence.

Proposition 4.9 $\vdash LIE_{ab}(\sigma) \wedge DEC_{ab}(\sigma, \delta) \supset \bot$ *and* $\vdash BS_{ab}(\sigma) \wedge DEC_{ab}(\sigma, \delta) \supset \bot$ *for any* $\sigma, \delta \in \Phi$.

As deception accompanies intention, offensive/defensive or deductive/abductive deception can also be defined. In lying and bullshitting, it is reasonable (and courteous to a hearer) not to lie and bullshit more than absolutely necessary (Postulates I and II). In case of deception, on the other hand, this is not necessarily the case. If an agent a deceives another agent b on the sentence $\sigma \wedge \lambda$, then the deception $\sigma \wedge \lambda$ is stronger than the deception σ. However, providing more information increases the knowledge of a hearer. For a speaker, providing more information implies concealing less information, which alleviates immoral feeling of the speaker. Thus, there is no reason to

prefer the weakest form of deception, so we do not have a postulate mandating it. On the other hand, deception is considered preferable to lies and BS as a speaker utters a believed-true sentence. This leads to the following postulate.

Postulate IV: Never lie nor bullshit if you can deceive your way out of it.
Let δ, λ, $\sigma \in \Phi^{\star}$. Then,
(i) $B_a(DEC_{ab}(\sigma, \delta) \supset B_b\delta) \wedge B_a(O\text{-}LIE_{ab}(\lambda, \delta) \supset B_b\delta) \supset \neg O\text{-}LIE_{ab}(\lambda, \delta)$.
(ii) $B_a(DEC_{ab}(\sigma, \delta) \supset B_b\delta) \wedge B_a(O\text{-}BS_{ab}(\lambda, \delta) \supset B_b\delta) \supset \neg O\text{-}BS_{ab}(\lambda, \delta)$.

The postulates I–IV are statements that agents should try to satisfy, both for moral reasons and for self-interested reasons (lower punishments if caught). If we assume that agents try to satisfy the dishonesty postulates, and that lying is worse than BS, which is again worse than deception, then one can characterize an agent by the worst level of dishonesty it is willing to commit in order to achieve a goal. For instance, a lawyer agent might be willing to deceive (providing only information favorable to his client) but not to BS nor to lie. So if one detects that an agent is deceiving, one cannot infer that it is also willing to BS or lie. However, the opposite is the case. If an agent is willing to lie, then from the dishonesty postulates, it can also be assumed to be willing to BS or to deceive. So an agent who is caught on deceiving can perhaps still be trusted not to lie (if trust is the default attitude), but an agent that is caught on lying cannot be trusted at all anymore (also regarding BS and deception). In multiagent systems if agents have implemented the dishonesty postulates, then this helps one to reason about the possible dishonesty of other agents, and about the extent to which they can still be trusted.

5 Discussion

Some attempts have been made to formulate lying using modal logic. O'Neill [17] provides logical definitions of lies and deception based on the logic of [7]. In contrast to our formulation with the logic BI_0, he uses the logic BI_2 which has four different modalities of belief, intention, common belief, and communication. His primary interest is to formulate various types of speech acts in an epistemic logic, and he does not investigate inference mechanisms behind the act of lying and other dishonesty. Different epistemic approaches are also reported in [22], but they just provide definitions of lies or deceptive utterances. Bonatti et al. [3] study databases that could lie to users to preserve security. They introduce a propositional modal logic to reason about databases, secrets, and users' beliefs. Their goal is formulating not lying but query answering in secure databases. Sklar et al. [20] formulate lying with argument-based dialogues. Their goal is capturing lies as contradictory dialogues, and they do not consider various types of lying, BS and deception. Caminada [4] provides a comparative study between lies, BS and deception and shows how these can be formalized using abstract argumentation. The paper provides philosophical arguments, but no logical theory is given.

This paper considered deductive and abductive lies, while lying can be combined with other types of inference. For instance, one may devise *inductive lies* by telling untrue evidences to make a hearer learn wrong inductive hypotheses. This paper focused attention in providing an ontology of dishonesty and explained how various forms of dishonesty are related to each other. It is also important to investigate how one can learn dishonesty attitudes in a multiagent society. Recent studies show that robots which

compete for foods learn to conceal food information [15]. Staab and Caminada [21] design and implement an MAS-based software simulator and observe that the incentives for dishonesty emerge for economical agents to have good performance.

6 Conclusion

We have provided a logical analysis of various concepts of dishonesty as they appear in the literature in philosophy and elsewhere. The issue of logical foundation of dishonesty is a topic that has received little attention until now. Our aim is to analyze this issue using a relatively simple logical formalization. Although some formal properties were provided, the strength of the current paper is conceptual rather than purely technical. The postulates can be seen as having a normative value, and should ideally be implemented for individual agents in multiagent systems. In future work, we elaborate the formulation and plan to build a formal system based on it.

References

1. Adler, J. E.: Lying, deceiving, or falsely implicating. J. Philosophy 94(9), 435–452 (1997)
2. Augustine, S.: Lying. In: Treatises on Various Subjects, Fathers of the Church, vol. 56, pp.45–110 (1952)
3. Bonatti, P. A., Kraus, S. and Subrahmanian, V. S.: Foundations of secure deductive databases. IEEE Transactions on Knowledge and Data Engineering 7(3), 406–422 (1995)
4. Caminada, M.: Truth, lies and bullshit, distinguishing classes of dishonesty. In: Proc. IJCAI Workshop on Social Simulation (2009)
5. Carson, T. L.: The definition of lying. Noûs 40(2), 284–306 (2006)
6. Chisholm, R. M. and Feehan, T. D.: The intent to deceive. Journal of Philosophy 74(3), 143–159 (1997)
7. Colombetti, M.: A modal logic of intentional communication. Mathematical Social Sciences 38, 171–196 (1999)
8. Demos, R.: Lying to oneself. Journal of Philosophy 57(18), 588–595 (1960)
9. Fallis, D.: What is lying? Journal of Philosophy 106(1), 29–56 (2009)
10. Frankfurt, H. G.: On Bullshit. Princeton Univ. Press (2005)
11. Halpern, J. and Moses, J.: A guide to completeness and complexity for modal logics of knowledge and belief. Artificial Intelligence 54, 349–379 (1992)
12. Kupfer, J.: The moral presumption against lying. Review of Metaphysics 36, 103–126 (1982)
13. Mahon, J. E.: A definition of deceiving. J. Applied Philosophy 21(2), 181–194 (2007)
14. Mahon, J. E.: Two definitions of lying. J. Applied Philosophy 22(2), 211–230 (2008)
15. Mitri, S., Floreano, D., and Keller, L.: The evolution of information suppression in communicating robots with conflicting interests. In: Proc. National Academy of Sciences 106(37):15786–15790 (2009)
16. Morris, J.: Can computers ever lie? Philosophy Forum 14, 389–401 (1976)
17. O'Neill, B.: A formal system for understanding lies and deceit. Jerusalem Conference on Biblical Economics (2003)
18. Peirce, C. S.: Collected Papers of Charles Sanders Peirce. Harvard University Press (1958)
19. Searle, J. R.: Speech Acts. Cambridge University Press (1969)
20. Sklar, E., Parsons, S. and Davies, M.: When is it okay to lie? A simple model of contradiction in agent-based dialogues. In: Proc. ArgMas, LNCS, vol. 3366, pp. 251–261, Springer (2005)
21. Staab, E. and Caminada, M.: Assessing the impact of informedness on a consultant's profit. In: Proc. 21st Benelux Conf. on AI (BNAIC'09), pp. 397–398, Eindhoven (2009)
22. Urchs, M.: Just lying. Logic and Logical Philosophy 15, 67–89 (2006)