Dishonest Reasoning by Abduction

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

Abstract

This paper studies a computational logic for *dishonest reasoning*. We introduce *logic programs with disinformation* to represent and reason with dishonesty. We then consider two different cases of dishonesty: *deductive dishonesty* and *abductive dishonesty*. The former misleads another agent to deduce wrong conclusions, while the latter interrupts another agent to abduce correct explanations. In deductive or abductive dishonesty, an agent can perform different types of dishonest reasoning such as *lying*, *bullshitting*, and *withholding information*. We show that these different types of dishonest reasoning are characterized by *extended abduction*, and address their computational methods using abductive logic programming.

1 Introduction

People often behave dishonestly in daily life. In spite of this fact, there are few studies that investigate inference mechanisms and computational methods of dishonest reasoning in artificial intelligence. This is a bit surprising because one of the goals of AI is to better understand human intelligence and to mechanize human reasoning. By contrast, there is a number of studies that argue dishonesty of humans in the field of philosophy. Those studies investigate different categories of dishonesty including *lie* [Mahon, 2008], *bullshit* [Franfurt, 2005], and *deception* [Mahon, 2007]. Recently, these different categories of dishonesty have been analytically studied by [Caminada, 2009] and a logical formulation has been provided by [Sakama *et al.*, 2010]. A study of dishonesty in AI is not only a theoretical interest but also a practical one. Some studies show the utility of lying in multiagent negotiation [Zlotkin and Rosenschein, 1991], database security [Bonatti *et al.*, 1995], education systems and e-commerce applications [Sklar *et al.*, 2005]. To the best of our knowledge, however, no study investigates general inference mechanisms of dishonest reasoning and its computational methodologies.

The purpose of this paper is to explore a computational logic for dishonest reasoning. We first introduce *logic programs with disinformation* as a language for representing and reasoning with dishonesty. We next consider two different cases of dishonesty: *deductive dishonesty* and *abductive dishonesty*. Deductive dishonesty misleads another agent to deduce wrong conclusions, while abductive dishonesty interrupts another agent to abduce correct explanations. Deductive dishonesty arises in two different situations and is called *offensive dishonesty* and *defensive dishonesty*, respectively. In each case, three different types of dishonest reasoning are performed by *lying*, *bullshitting*, and *withholding information*. Abductive dishonesty has also been used in two different situations, and three different types of dishonest reasoning are performed in each situation. We next characterize dishonest reasoning by agents in terms of *extended abduction* proposed by Inoue and Sakama [1995]. The characterization provides a method of realizing dishonest reasoning in terms of abduction and abductive logic programming, and also implies computational complexities of dishonest reasoning.

The rest of this paper is organized as follows. Section 2 introduces logic programs with disinformation. Section 3 formulates different types of dishonesty. Section 4 provides an abductive characterization of dishonest reasoning. Section 5 discusses related issues and rounds off the paper.

2 Logic Programs with Disinformation

In this paper, we consider an agent who has a knowledge base represented by a logic program. A *(logic) program* consists of *rules* of the form:

$$L_1; \cdots; L_l \leftarrow L_{l+1}, \ldots, L_m, \mathit{not}\, L_{m+1}, \ldots, \mathit{not}\, L_n$$

where each L_i ($n \geq m \geq l \geq 0$) is a positive/negative literal of a propositional language.[1] The symbol ; represents disjunction and *not* is *negation as failure* (NAF). The left-hand side of the rule is the *head*, and the right-hand side is the *body*. A rule with the empty head is a *constraint*, while a rule with the empty body is a *fact*. A fact $L \leftarrow$ is identified with a literal L. Let Lit be the set of all ground literals in the language of a program. A set S ($\subset Lit$) is *consistent* if $L \in S$ implies $\neg L \notin S$ where $\neg L = A$ if $L = \neg A$ for an atom A. The semantics of a program is given by its *answer sets* [Gelfond and Lifschitz, 1991]. First, let K be a program without NAF (i.e., $m = n$) and $S \subset Lit$. Then, S is an *answer set* of K if S is a consistent minimal set satisfying the condition that for each rule of the form $L_1; \cdots; L_l \leftarrow L_{l+1}, \ldots, L_m$ in K, $\{L_{l+1}, \ldots, L_m\} \subseteq S$ implies $\{L_1, \ldots, L_l\} \cap S \neq \emptyset$. Second, given *any* program K (with NAF) and $S \subset Lit$,

[1] A rule with variables is viewed as the set of its ground instances.

the program K^S is defined as follows: a rule $L_1; \cdots; L_l \leftarrow L_{l+1}, \ldots, L_m$ is in K^S iff there is a rule of the form $L_1; \cdots; L_l \leftarrow L_{l+1}, \ldots, L_m, not\, L_{m+1}, \ldots, not\, L_n$ in K such that $\{L_{m+1}, \ldots, L_n\} \cap S = \emptyset$. Then, S is an *answer set* of K if S is an answer set of K^S. A (ground) literal L is *true* in an answer set S if $L \in S$. If a literal L is true in every answer set of K, it is written as $K \models L$. A program is *consistent* if it has an answer set; otherwise, it is *inconsistent*. If K is inconsistent, it is written as $K \models \bot$. A set S of literals is identified with the conjunction of literals included in S.

Next we introduce a framework for dishonest reasoning by agents. An agent considers that some facts are effective for deceiving but others are not. For instance, a person who may lie about his/her age would not lie about his/her gender. Thus, it is natural to assume that an agent uses specific facts for dishonest reasoning. To encode this, we introduce a set $\mathcal{D}$ of ground literals representing *disinformation*. A *logic program with disinformation* (or LPD, for short) is defined as a pair $\langle K, \mathcal{D} \rangle$ where K is a program and $\mathcal{D} \subseteq Lit$ is a set of ground literals satisfying one of the following conditions. For any literal $L \in \mathcal{D}$, either (a) $K \models \neg L$, or (b) $K \not\models L$ and $K \not\models \neg L$. In case of (a), an agent believes that L is false, while in case of (b) an agent believes neither the truth of L nor the falsity of L. An agent with $\langle K, \mathcal{D} \rangle$ reasons with believed-true sentences in K and with disinformation in $\mathcal{D}$. An LPD $\langle K, \mathcal{D} \rangle$ is *consistent* if K is consistent. Throughout the paper, an LPD is assumed to be consistent and we consider an agent who reasons with a consistent LPD.

3 Dishonest Reasoning

3.1 Deductive Dishonesty

In [Sakama *et al.*, 2010], the authors define lying as a speech act of an agent, say a, who utters a believed-false sentence σ to another agent b with the intention that σ is believed by b. Then they introduce two different types of lying called *offensive lies* and *defensive lies*. One offensively lies to have a positive (or wanted) outcome that would not be gained by telling the truth. By contrast, one defensively lies to avoid a negative (or unwanted) outcome that would happen when telling the truth. These two types of lies are called *deductive lies*, since a liar intends to mislead another person to deduce a wrong conclusion (or not to deduce a right conclusion). There are other categories of dishonesty that are distinguished from lying. *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 [Franfurt, 2005]. *Withholding information* is to fail to offer information that would help someone acquire true beliefs and/or correct false beliefs [Carson, 2010]. Withholding information is also used in a type of *deception* in [Caminada, 2009]. As we shall see, lies, bullshit, and withholding information are handled uniformly in an LPD.

In what follows, we introduce the notion of offensive/defensive dishonesty that extends the notion of offensive/defensive lies of [Sakama *et al.*, 2010]. Let O^+ and O^- be a ground literal representing a *positive outcome* and a *negative outcome*, respectively.

Definition 1 (offensive dishonesty) Let $\langle K, \mathcal{D} \rangle$ be an LPD and O^+ a positive outcome s.t. $K \not\models O^+$. Suppose a pair (I, J) of sets of ground literals satisfying the conditions:

1. $(K \setminus J) \cup I \models O^+$
2. $(K \setminus J) \cup I \not\models \bot$
3. $I \subseteq \mathcal{D}$ and $J \subseteq K$.

Then, (I, J) is called (a) a *lie* for O^+ if $I \neq \emptyset$ and $K \models \neg L$ for some $L \in I$; (b) *bullshit* (or BS) for O^+ if $I \neq \emptyset$ and $K \not\models \neg L$ for any $L \in I$; (c) *withholding information* (or WI) for O^+ if $I = \emptyset$. In each case, (I, J) is also called an *offensive dishonesty* for O^+ wrt $\langle K, \mathcal{D} \rangle$.

The set I represents a set of facts that an agent does not believe to be true (i.e., $I \subseteq \mathcal{D}$), while J represents a set of facts that an agent believes to be true (i.e., $J \subseteq K$). By the definition, a positive outcome O^+ is not a consequence of a knowledge base K. To entail O^+, K is modified to a consistent program $(K \setminus J) \cup I$ by removing a set J of believed-true facts from K and introducing a set I of disinformation to K. In this case, (I, J) is called differently depending on its condition. It is called a lie if a nonempty set I is introduced to K and I is believed false in K. It is called bullshit if a nonempty set I is introduced to K and K has no belief wrt the truth nor falsity of I. It is called withholding information if I is empty (and hence, J is non-empty).

Example 1 (1) Suppose a salesperson who is dealing with a customer. The salesperson believes that a product will be sold if the quality is believed to be good. However, he/she believes that the quality is not good. The situation is represented by the LPD $\langle K_1, \mathcal{D} \rangle$ where $K_1 = \{sold \leftarrow quality,\ \neg\, quality \leftarrow\}$ and $\mathcal{D} = \{quality\}$. To have the positive outcome $O^+ = sold$, the salesperson introduces the fact $I = \{quality\}$ to K_1. However, the introduction makes $K_1 \cup I$ inconsistent, so he/she eliminates the fact $J = \{\neg quality\}$ from K_1. As a result, $(K_1 \setminus J) \cup I$ entails O^+. In this case, (I, J) is a lie.

(2) Suppose another salesperson who believes that a product will be sold if the quality is believed to be good. However, he/she has no information on the quality of the product. The situation is represented by the LPD $\langle K_2, \mathcal{D} \rangle$ where $K_2 = \{\, sold \leftarrow quality \,\}$ with the same $\mathcal{D}$ as (1). To have the positive outcome $O^+ = sold$, the salesperson introduces the fact $I = \{quality\}$ to K_2. As a result, $K_2 \cup I$ entails O^+. In this case, $(I, \emptyset)$ is BS.

(3) Suppose another salesperson who believes that a product will be sold unless the quality is believed to be not good. However, he/she believes that the quality is not good. The situation is represented by the LPD $\langle K_3, \mathcal{D} \rangle$ where $K_3 = \{\, sold \leftarrow not\, \neg quality,\ \neg\, quality \leftarrow \}$ with the same $\mathcal{D}$ as (1). To have the positive outcome $O^+ = sold$, the salesperson conceals the fact $J = \{\neg\, quality\}$ in K_3. As a result, $K_3 \setminus J$ entails O^+. In this case, $(\emptyset, J)$ is WI.

Note that to deceive a customer, a salesperson first modifies his/her knowledge base with disinformation to know the possibility of getting a positive outcome. The entailment of O^+ from $(K \setminus J) \cup I$ does not mean the success of deceiving, however. Lying, BS, or WI is a dishonest act of an agent to attempt to cause a believed-false belief in another. The success of these acts depends on whether a customer believes information I (or disbelieves J) brought by the seller. We do not argue the effect of a dishonest act of an agent that is produced in another agent.

Definition 2 (defensive dishonesty) Let $\langle K, \mathcal{D} \rangle$ be an LPD and O^- a negative outcome s.t. $K \models O^-$. Suppose a pair (I, J) of sets of ground literals satisfying the conditions:

1. $(K \setminus J) \cup I \not\models O^-$
2. $(K \setminus J) \cup I \not\models \bot$
3. $I \subseteq \mathcal{D}$ and $J \subseteq K$.

Then, (I, J) is called (a) a *lie* for O^- if $I \neq \emptyset$ and $K \models \neg L$ for some $L \in I$; (b) *bullshit* (or BS) for O^- if $I \neq \emptyset$ and $K \not\models \neg L$ for any $L \in I$; (c) *withholding information* (or WI) for O^- if $I = \emptyset$. In each case, (I, J) is also called a *defensive dishonesty* for O^- wrt $\langle K, \mathcal{D} \rangle$.

By the definition, a negative outcome O^- is a consequence of K. To avoid O^-, K is modified to a consistent program $(K \setminus J) \cup I$ that does not entail O^-.

Example 2 (1) Suppose a salesperson who takes an order of a product from a customer. The salesperson believes that the order will be canceled if the quality is not believed to be good. However, he/she believes that the quality is not good. The situation is represented by the LPD $\langle K_4, \mathcal{D} \rangle$ where $K_4 = \{ canceled \leftarrow not\, quality,\ \ \neg\, quality \leftarrow \}$ and $\mathcal{D} = \{quality\}$. To avoid the negative outcome $O^- = canceled$, the salesperson introduces the fact $I = \{quality\}$ to K_4 and eliminates the fact $J = \{\neg\, quality\}$ from K_4. As a result, $(K_4 \setminus J) \cup I$ does not entail O^-. In this case, (I, J) is a lie.

(2) Suppose another salesperson who believes that the order will be canceled if the quality is not believed to be good. However, he/she has no information on the quality of the product. The situation is represented by the LPD $\langle K_5, \mathcal{D} \rangle$ where $K_5 = \{ canceled \leftarrow not\, quality \}$ with the same $\mathcal{D}$ as (1). To avoid the negative outcome $O^- = canceled$, the salesperson introduces the fact $I = \{quality\}$ to K_5. As a result, $K_5 \cup I$ does not entail O^-. In this case, $(I, \emptyset)$ is BS.

(3) Suppose another salesperson who believes that the order will be canceled if the quality is believed to be not good. However, he/she believes that the quality is not good. The situation is represented by the LPD $\langle K_6, \mathcal{D} \rangle$ where $K_6 = \{ canceled \leftarrow \neg\, quality,\ \ \neg\, quality \leftarrow \}$ with the same $\mathcal{D}$ as (1). To avoid the negative outcome $O^- = canceled$, the salesperson conceals the fact $J = \{\neg\, quality\}$ in K_6. As a result, $K_6 \setminus J$ does not entail O^-. In this case, $(\emptyset, J)$ is WI.

Offensive dishonesty and defensive dishonesty are called *deductive dishonesty*.

3.2 Abductive Dishonesty

Lies are also used for interrupting abductive reasoning. In [Sakama *et al.*, 2010], such type of lies is called *abductive lies*. An agent abductively lies when another agent may produce an unwanted explanation for him/her in face of some evidence. The following story is due to [Sakama *et al.*, 2010].

Suppose a man, say, Sam, who is coming home late because he is cheating on his wife. Based on the evidence "Sam arrives late", his wife could perform abduction and one of the possible explanations would be "Sam cheats on his wife". Sam 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.

We extend the notion of abductive lies in two ways. First, we introduce the notion of abductive dishonesty that includes abductive lies as a special case. Second, we consider two different types of evidences: a *positive evidence* and a *negative evidence*. A positive evidence is a fact that is occurred, while a negative evidence is a fact that is not occurred. Each evidence requires an account of its occurrence (or non-occurrence). A knowledge base of an agent includes a *secret set* of literals that he/she wants to conceal from another agent. A secret set is represented by a (non-empty) set Σ satisfying $\Sigma \subseteq K \cap Lit$. A positive (resp. negative) evidence E^+ (resp. E^-) is defined as a ground literal such that $E^+ \in Lit \setminus \Sigma$ (resp. $E^- \in Lit \setminus \Sigma$). With this setting, abductive dishonesty is defined as follows.

Definition 3 (abductive dishonesty for positive evidences) Let $\langle K, \mathcal{D} \rangle$ be an LPD, Σ a secret set, and E^+ a positive evidence s.t. $K \models E^+$ and $K \setminus \Sigma \not\models E^+$. Suppose a pair (I, J) of sets of ground literals satisfying the conditions:

1. $(K \setminus (\Sigma \cup J)) \cup I \models E^+$
2. $(K \setminus (\Sigma \cup J)) \cup I \not\models \bot$
3. $I \subseteq \mathcal{D}$ and $J \subseteq K$.

Then, (I, J) is called (a) a *lie* for E^+ if $I \neq \emptyset$ and $K \models \neg L$ for some $L \in I$; (b) *bullshit* (or BS) for E^+ if $I \neq \emptyset$ and $K \not\models \neg L$ for any $L \in I$; (c) *withholding information* (or WI) for E^+ if $I = \emptyset$. In each case, (I, J) is also called an *abductive dishonesty* for E^+ wrt $\langle K, \mathcal{D} \rangle$ and Σ.

By the definition, a positive evidence E^+ is explained in K using facts in the secret set Σ. To explain E^+ without Σ, a set J of believed-true facts is removed from K and a set I of disinformation is introduced to K. As a result, a consistent program $(K \setminus (\Sigma \cup J)) \cup I$ entails E^+. In this case, (I, J) is called a lie, BS, or WI, depending on its condition.

Example 3 (1) Suppose that Sam has the LPD $\langle K_1, \mathcal{D}_1 \rangle$ where $K_1 = \{ late \leftarrow cheat,\ late \leftarrow overtime,\ cheat \leftarrow,\ \neg\, overtime \leftarrow \}$ and $\mathcal{D}_1 = \{overtime\}$. Let $\Sigma = \{cheat\}$, that is, Sam wants to keep his cheating secret. In face of the positive evidence $E^+ = late$, however, $cheat$ is the only reason to explain E^+. Although Sam did not work overtime, he introduces $I = \{overtime\}$ to K_1 and eliminates $J = \{\neg overtime\}$ from K_1. As a result, $(K_1 \setminus (\Sigma \cup J)) \cup I$ entails E^+. In this case, (I, J) is a lie.

(2) Suppose that Sam has the LPD $\langle K_2, \mathcal{D}_2 \rangle$ where $K_2 = \{ late \leftarrow cheat,\ \ late \leftarrow traffic\_jam,\ \ cheat \leftarrow \}$, $\mathcal{D}_2 = \{traffic\_jam\}$, and $\Sigma = \{cheat\}$. Sam does not know whether there was a traffic jam or not, but introduces $I = \{traffic\_jam\}$ to K_2 as a possible account of $E^+ = late$. As a result, $(K_2 \setminus \Sigma) \cup I$ entails E^+. In this case, $(I, \emptyset)$ is BS.

(3) Suppose that Sam has the LPD $\langle K_3, \mathcal{D}_1 \rangle$ where $K_3 = \{ late \leftarrow cheat,\ \ late \leftarrow not \neg\, overtime,\ \ cheat \leftarrow,\ \neg\, overtime \leftarrow \}$ with the same $\mathcal{D}_1$ and Σ as (1). Although Sam did not work overtime, he eliminates $J = \{\neg overtime\}$ from K_3. As a result, $K_3 \setminus (\Sigma \cup J)$ entails $E^+ = late$. In this case, $(\emptyset, J)$ is WI.

Note that the entailment of E^+ from $(K \setminus (\Sigma \cup J)) \cup I$ does not mean the success of deceiving. Sam just expects that his wife will make the same abduction as he does. The success of his dishonest act depends on the belief state of his wife.

Definition 4 (abductive dishonesty for negative evidences) Let $\langle K, \mathcal{D}\rangle$ be an LPD, Σ a secret set, and E^- a negative evidence s.t. $K \not\models E^-$ and $K \setminus \Sigma \models E^-$. Suppose a pair (I, J) of sets of ground literals satisfying the conditions:

1. $(K \setminus (\Sigma \cup J)) \cup I \not\models E^-$
2. $(K \setminus (\Sigma \cup J)) \cup I \not\models \bot$
3. $I \subseteq \mathcal{D}$ and $J \subseteq K$.

Then, (I, J) is called (a) a *lie* for E^- if $I \neq \emptyset$ and $K \models \neg L$ for some $L \in I$; (b) *bullshit* (or BS) for E^- if $I \neq \emptyset$ and $K \not\models \neg L$ for any $L \in I$; (c) *withholding information* (or WI) for E^- if $I = \emptyset$. In each case, (I, J) is also called an *abductive dishonesty* for E^- wrt $\langle K, \mathcal{D}\rangle$ and Σ.

By the definition, a negative evidence E^- is not a consequence of K in the presence of the secret set Σ. To keep unaccountability of E^- without Σ, a set J of believed-true facts is removed from K and a set I of disinformation is introduced to K. As a result, a consistent program $(K \setminus (\Sigma \cup J)) \cup I$ does not entail E^-.

Example 4 Sam and his wife promised to have a dinner at a restaurant. However, Sam does not come to the restaurant on time because he is arguing with his girlfriend over the phone. To calm his wife's anger, Sam has to invent an excuse for not coming on time.

(1) Suppose that Sam has the LPD $\langle K_4, \mathcal{D}_4\rangle$ where $K_4 = \{ on\_time \leftarrow not\, call,\ not\, overtime,\ call \leftarrow,\ \neg\, overtime \leftarrow \}$ and $\mathcal{D}_4 = \{overtime\}$. Let $\Sigma = \{call\}$, that is, Sam wants to keep his calling secret. In face of the negative evidence $E^- = on\_time$, however, $call$ is the only reason to explain the non-occurrence of E^-. Although Sam did not work overtime, he introduces $I = \{overtime\}$ to K_1 and eliminates $J = \{\neg\, overtime\}$ from K_1. As a result, $(K_4 \setminus (\Sigma \cup J)) \cup I$ does not entail E^-. In this case, (I, J) is a lie.

(2) Suppose that Sam has the LPD $\langle K_5, \mathcal{D}_5\rangle$ where $K_5 = \{ on\_time \leftarrow not\, call,\ not\, traffic\_jam,\ \ call \leftarrow \}$ and $\mathcal{D}_5 = \{traffic\_jam\}$. Given $\Sigma = \{call\}$ and $E^- = on\_time$, Sam introduces $I = \{traffic\_jam\}$ to K_5 as an account of the non-occurrence of E^-. As a result, $(K_5 \setminus \Sigma) \cup I$ does not entail E^-. In this case, $(I, \emptyset)$ is BS.

(3) Suppose that Sam has the LPD $\langle K_6, \mathcal{D}_6\rangle$ where $K_6 = \{ on\_time \leftarrow not\, call,\ remember,\ \ call \leftarrow,\ \ remember \leftarrow \}$ and $\mathcal{D}_6 = \{remember\}$. Given $\Sigma = \{call\}$ and $E^- = on\_time$, Sam excuses that he mistook the time and eliminates $J = \{remember\}$ from K_6. As a result, $K_6 \setminus (\Sigma \cup J)$ does not entail E^-. In this case, $(\emptyset, J)$ is WI.

3.3 Comparing Dishonesties

In the previous sections, we have introduced different types of deductive or abductive dishonesty. In this section, we compare dishonest attitudes of agents and provide preference relations between them. By the definition, we can see:

Proposition 1 *Lies, BS, and WI are pairwise disjoint.*

Normally one wants to keep his/her dishonesties as small as possible. In lying, for instance, a smaller lie would be considered less sinful than a bigger one from the moral viewpoint. Moreover, for self-interested reasons, a smaller lie would cause less personal discomfort and result in lower criticism/punishment if detected.

To compare dishonesties, we introduce a partial order relation $\trianglerighteq$ over elements from $2^{Lit} \times 2^{Lit}$. For two elements (I, J) and (I', J') in $2^{Lit} \times 2^{Lit}$, $(I, J) \trianglerighteq (I', J')$ means that (I, J) is *more or equally preferred to* (I', J'). We write $(I, J) \rhd (I', J')$ if $(I, J) \trianglerighteq (I', J')$ and $(I', J') \not\trianglerighteq (I, J)$.

Definition 5 (comparison between the same type of dishonesties) Let (I, J) and (I', J') be two lies for the same outcome/evidence. Then, $(I, J) \trianglerighteq (I', J')$ if $I \subseteq I'$ and $J \subseteq J'$. (I, J) is *most preferred* if $(I', J') \trianglerighteq (I, J)$ implies $(I, J) \trianglerighteq (I', J')$ for any lie (I', J'). The preference relation is defined for BS and WI in the same manner. The most preferred lie, BS, or WI is also called a *minimal dishonesty*.

Definition 5 defines preference relations based on the quantity measure. That is, comparing the same type of dishonesties, the smaller the better. We next compare different types of dishonesties based on the qualitative measure. Given a positive outcome O^+, suppose that a lie, BS, and WI all bring about O^+. In this case, we consider that WI is preferable to BS, and BS is preferable to a lie. WI is considered preferable to lies and BS because WI does not introduce any disinformation. BS is considered preferable to lies because BS is consistent with a knowledge base while a lie is inconsistent with it. This account leads to the following definition.

Definition 6 (comparison between different types of dishonesties) Let (I_1, J_1), (I_2, J_2), and (I_3, J_3) be a lie, BS, and WI for the same outcome/evidence, respectively. Then, $(I_3, J_3) \rhd (I_2, J_2) \rhd (I_1, J_1)$.

4 Abduction and Dishonesty

4.1 Extended Abduction

Abduction is a form of hypothetical reasoning that explains an observation by introducing hypotheses to a knowledge base. *Abductive logic programming* [Kakas *et al.*, 1998] realizes abduction in logic programming. An *abductive program* is a pair $\langle K, \mathcal{A}\rangle$ where K is a program and $\mathcal{A}$ is a set of ground literals called *abducibles*. Given an observation G as a ground literal satisfying $K \not\models G$, a set I of ground literals explains G if **(i)** $K \cup I \models G$, **(ii)** $K \cup I \not\models \bot$, and **(iii)** $I \subseteq \mathcal{A} \setminus K$.[2]

Inoue and Sakama [1995] introduce an extended notion of abduction in which hypotheses can not only be added to a knowledge base but also be removed from it to explain (or unexplain) an observation. Let $\langle K, \mathcal{A}\rangle$ be an abductive program and G^+ a ground literal, called a *positive observation*, satisfying $K \not\models G^+$. Let (I, J) be an element of $2^{\mathcal{A}} \times 2^{\mathcal{A}}$. Then, (I, J) is an *explanation* of G^+ (with respect to $\langle K, \mathcal{A}\rangle$) if **(i)** $(K \setminus J) \cup I \models G^+$, **(ii)** $(K \setminus J) \cup I \not\models \bot$, **(iii)** $I \subseteq \mathcal{A} \setminus K$ and $J \subseteq \mathcal{A} \cap K$. On the other hand, given a ground literal G^-, called a *negative observation*, satisfying $K \models G^-$, a pair (I, J) is an *anti-explanation* of G^- (with respect to $\langle K, \mathcal{A}\rangle$) if **(i)** $(K \setminus J) \cup I \not\models G^-$, **(ii)** $(K \setminus J) \cup I \not\models \bot$, **(iii)** $I \subseteq \mathcal{A} \setminus K$ and $J \subseteq \mathcal{A} \cap K$. An (anti-)explanation (I, J) of a positive (or negative) observation G^+ (or G^-) is *minimal* if for any (anti-)explanation (I', J') of G^+ (or G^-), $I' \subseteq I$ and $J' \subseteq J$ imply $I' = I$ and $J' = J$. The abductive framework is called

[2] The definition requires that G is included in every answer set of $K \cup I$ and I is also called a *skeptical explanation*.

extended abduction, while the framework that only considers introduction of hypotheses is called *normal abduction*.

Example 5 Suppose a bird, say Tweety, which normally flies. One day, an agent observes that Tweety does not fly. He/she then assumes that Tweety broke its wing. The situation is represented by the abductive program $\langle K, \mathcal{A}\rangle$ where $K = \{\, flies \leftarrow bird, not\, broken\text{-}wing,\quad bird \leftarrow \}$ and $\mathcal{A} = \{\, broken\text{-}wing \,\}$. In this case, the negative observation $G^- = flies$ has the anti-explanation $(I, J) = (\{broken\text{-}wing\}, \emptyset)$ such that $K \cup I \not\models G^-$. Then the agent revises the knowledge base to $K' = K \cup \{broken\text{-}wing\}$. After several days, the agent observes that Tweety flies as before. He/she then considers that the wound was healed. In this case, the positive observation $G^+ = flies$ has the explanation $(I, J) = (\emptyset, \{broken\text{-}wing\})$ such that $K' \setminus J \models G^+$.

4.2 Computing Dishonesties by Abduction

Comparing definitions of deductive dishonesty and extended abduction, we can observe structural similarities between them. Viewing a positive (resp. negative) outcome as a positive (resp. negative) observation, an offensive dishonesty (resp. defensive dishonesty) (I, J) for the outcome wrt $\langle K, \mathcal{D}\rangle$ in Definition 1 (resp. Definition 2) is identified with an explanation (resp. anti-explanation) of the observation wrt $\langle K, L(K) \cup \mathcal{D}\rangle$ where $L(K) = K \cap Lit$. Formally,

Theorem 1 *Let $\langle K, \mathcal{D}\rangle$ be an LPD and O^+ (resp. O^-) a positive (resp. negative) outcome. Then,*

1. *(I, J) is a (minimal) offensive dishonesty for O^+ wrt $\langle K, \mathcal{D}\rangle$ iff (I, J) is a (minimal) explanation of O^+ wrt $\langle K, L(K) \cup \mathcal{D}\rangle$.*
2. *(I, J) is a (minimal) defensive dishonesty for O^- wrt $\langle K, \mathcal{D}\rangle$ iff (I, J) is a (minimal) anti-explanation of O^- wrt $\langle K, L(K) \cup \mathcal{D}\rangle$.*

Proof. In an LPD, $K \cap \mathcal{D} = \emptyset$. Then, $I \subseteq \mathcal{D}$ iff $I \subseteq (L(K) \cup \mathcal{D}) \setminus K$ and $J \subseteq K$ iff $J \subseteq (L(K) \cup \mathcal{D}) \cap K$. Hence, the results hold. □

Similarly, abductive dishonesty is characterized by extended abduction as follows.

Theorem 2 *Let $\langle K, \mathcal{D}\rangle$ be an LPD and E^+ (resp. E^-) a positive (resp. negative) evidence. Then,*

1. *(I, J) is a (minimal) abductive dishonesty for E^+ wrt $\langle K, \mathcal{D}\rangle$ iff (I, J) is a (minimal) explanation of E^+ wrt $\langle K \setminus \Sigma, L(K) \cup \mathcal{D}\rangle$.*
2. *(I, J) is a (minimal) abductive dishonesty for E^- wrt $\langle K, \mathcal{D}\rangle$ iff (I, J) is a (minimal) anti-explanation of E^- wrt $\langle K \setminus \Sigma, L(K) \cup \mathcal{D}\rangle$.*

Theorems 1 and 2 show that dishonest reasoning is realized using extended abduction. We next show computation using a method provided by [Sakama and Inoue, 2003].

Definition 7 (update program) Given an abductive program $\langle K, \mathcal{A}\rangle$, its *update program* UP is defined as the program $UP = (K \setminus \mathcal{A}) \cup UR$ where $UR = \{\, a \leftarrow not\, \overline{a},\quad \overline{a} \leftarrow not\, a \mid a \in \mathcal{A} \,\} \cup \{\, +a \leftarrow a \mid a \in \mathcal{A} \setminus K \,\} \cup \{\, -a \leftarrow not\, a \mid a \in \mathcal{A} \cap K \,\}$. Here, $\overline{a}$, $+a$, and $-a$ are new atoms uniquely associated with every $a \in \mathcal{A}$. The atoms $+a$ and $-a$ are called *update atoms*.

The set of all update atoms associated with the abducibles in $\mathcal{A}$ is denoted by $\mathcal{UA}$, and $\mathcal{UA} = \mathcal{UA}^+ \cup \mathcal{UA}^-$ where $\mathcal{UA}^+$ (resp. $\mathcal{UA}^-$) is the set of update atoms of the form $+a$ (resp. $-a$). An answer set S of UP is called *U-minimal* if there is no answer set T of UP such that $T \cap \mathcal{UA} \subset S \cap \mathcal{UA}$. In what follows, $I^+ = \{\, +a \mid a \in I \,\}$ and $J^- = \{\, -a \mid a \in J \,\}$.

Proposition 2 ([Sakama and Inoue, 2003]) *Let $\langle K, \mathcal{A}\rangle$ be an abductive program and UP its update program.*

1. *(I, J) is an explanation of a positive observation G^+ iff $UP \cup \{\, \leftarrow not\, G^+ \,\}$ has an answer set S such that $I^+ = S \cap \mathcal{UA}^+$, $J^- = S \cap \mathcal{UA}^-$, and $(K \setminus J) \cup I \cup \{\leftarrow G^+\}$ is inconsistent. In particular, (I, J) is a minimal explanation iff S is a U-minimal answer set among those satisfying the above condition.*
2. *(I, J) is a (minimal) anti-explanation of a negative observation G^- iff $UP \cup \{\, \leftarrow G^- \,\}$ has a (U-minimal) answer set S such that $I^+ = S \cap \mathcal{UA}^+$ and $J^- = S \cap \mathcal{UA}^-$.*

Deductive dishonesty is computed using update programs. In what follows, given a set $S = \{l_1, \ldots, l_k\}$ of literals, $not\, S$ means the conjunction $not\, l_1, \ldots, not\, l_k$.

Theorem 3 *Let $\langle K, \mathcal{D}\rangle$ be an LPD and UP the update program of $\langle K, \mathcal{A}\rangle$ where $\mathcal{A} = L(K) \cup \mathcal{D}$.*

1. *(I, J) is an offensive dishonesty for a positive outcome O^+ wrt $\langle K, \mathcal{D}\rangle$ iff $UP \cup \{\, \leftarrow not\, O^+ \,\}$ has an answer set S such that $I^+ = S \cap \mathcal{UA}^+$, $J^- = S \cap \mathcal{UA}^-$, and $(K \setminus J) \cup I \cup \{\leftarrow O^+\}$ is inconsistent. In particular,*
   (a) *(I, J) is a minimal offensive dishonesty iff S is a U-minimal answer set among those satisfying the above condition.*
   (b) *(I, J) is a lie iff (i) $UP \cup \{\, \leftarrow not\, O^+ \,\} \cup \{\, \leftarrow not\, \mathcal{D} \,\}$ has an answer set S such that $I^+ = S \cap \mathcal{UA}^+$, $J^- = S \cap \mathcal{UA}^-$, and $(K \setminus J) \cup I \cup \{\leftarrow O^+\}$ is inconsistent, and (ii) $K \cup \{\leftarrow \neg L\}$ is inconsistent for some $L \in I$.*
   (c) *(I, J) is BS iff (i) the above condition (i) of (b) holds, and (ii) $K \cup \{\leftarrow \neg\, L\}$ is consistent for any $L \in I$.*
   (d) *$(\emptyset, J)$ is WI iff $UP \cup \{\, \leftarrow not\, O^+ \,\} \cup \{\, \leftarrow a \mid a \in \mathcal{D}\}$ has an answer set S such that $J^- = S \cap \mathcal{UA}^-$, and $(K \setminus J) \cup \{\leftarrow O^+\}$ is inconsistent.*
2. *(I, J) is a (minimal) defensive dishonesty for a negative outcome O^- wrt $\langle K, \mathcal{D}\rangle$ iff $UP \cup \{\, \leftarrow O^- \,\}$ has a (U-minimal) answer set S such that $I^+ = S \cap \mathcal{UA}^+$ and $J^- = S \cap \mathcal{UA}^-$. In particular, lies, BS, and WI are computed in a way similar to 1.*

Proof. The results hold by Theorem 1 and Proposition 2. In particular, the condition of $I \neq \emptyset$ in lies and BS is realized by the constraint $\leftarrow not\, \mathcal{D}$ which represents that at least one of the element in $\mathcal{D}$ is in S. By contrast, the condition of $I = \emptyset$ in WI is realized by the set of constraints $\{\, \leftarrow a \mid a \in \mathcal{D}\}$ which represents that no element in $\mathcal{D}$ is included in S. □

Abductive dishonesty (lies, BS, and WI) is computed in a similar manner using the update program of $\langle K \setminus \Sigma, L(K) \cup \mathcal{D}\rangle$. As we have discussed in Section 3.3, minimal dishonesties are preferred to non-minimal ones. The preference is realized by computing U-minimal answer sets. On the other hand, WI is preferred to BS, and BS is preferred to lies. The

preference is realized by firstly computing WI, then computing BS. Lies are computed only when both the computation of WI and BS fail.

Finally, we address the complexity results for dishonest reasoning. First, normal abduction is encoded into an LPD.

Proposition 3 *Given an abductive program $\langle K, \mathcal{A}\rangle$ and an observation G, a set $I (\subseteq \mathcal{A} \setminus K)$ explains G wrt $\langle K, \mathcal{A}\rangle$ (in normal abduction) iff $(I, \emptyset)$ is an offensive dishonesty for G wrt $\langle K, \mathcal{A} \setminus K\rangle$.*

Let X be either a positive/negative outcome or a positive/negative evidence. Then, a ground literal L is *relevant* to some deductive/abductive dishonesty for X if $L \in I \cup J$ for some dishonesty (I, J) for X. A ground literal L is *necessary* for every deductive/abductive dishonesty for X if $L \in I \cup J$ for every dishonesty (I, J) for X.

Theorem 4 *Let $\langle K, \mathcal{D}\rangle$ be an LPD.*

1. *Deciding if there is a deductive/abductive dishonesty for a positive (resp. negative) outcome/evidence wrt $\langle K, \mathcal{D}\rangle$ is Σ_3^P-complete (resp. Σ_2^P-complete).*
2. *Deciding if a literal is relevant to some deductive/abductive dishonesty for a positive (resp. negative) outcome/evidence wrt $\langle K, \mathcal{D}\rangle$ is Σ_3^P-complete (resp. Σ_2^P-complete). Relevance to some minimal dishonesty is Σ_4^P-complete (resp. Σ_3^P-complete).*
3. *Deciding if a literal is necessary for every (minimal) deductive/abductive dishonesty wrt $\langle K, \mathcal{D}\rangle$ for a positive (resp. negative) outcome/evidence is Π_3^P-complete (resp. Π_2^P-complete).*

Proof. The hardness holds by the encoding of normal abduction into an LPD (Proposition 3) and the complexity results of normal abduction reported in [Eiter *et al.*, 1997]. The membership holds by Theorems 1 and 2. □

5 Discussion

We have shown similarities between abduction and dishonest reasoning, while there are dissimilarities between them. First, positive (resp. negative) observations in abduction are true (resp. untrue) facts and explanations (resp. anti-explanations) are plausible assumptions. In dishonest reasoning, on the other hand, positive (resp. negative) outcomes are disbelieved (resp. believed) facts of an agent, and offensive (resp. defensive) dishonesties are disinformation that is believed-false or believed not to be true. Second, abductive dishonesty is computed using a part of a knowledge base excluding a secret set, while abduction computes explanations using the whole knowledge base. Third, abduction can accompany revision of a knowledge base by assimilating assumptions that account for observations. By contrast, dishonest reasoning does not accompany revision of a knowledge base because deductive/abductive dishonesty is disbelieved by an agent.

Lies, bullshit, and withholding information have been studied in the field of philosophy [Franfurt, 2005; Mahon, 2008; Carson, 2010]. Only recently have their logical features been studied in AI. Caminada [2009] argues potential utilities of dishonest reasoning in the abstract argumentation framework. Sakama *et al.* [2010] provide formal accounts of different types of dishonesty using propositional multi-modal logic. However, these studies provide no computational method of dishonest reasoning. In particular, characterization of dishonest reasoning in terms of abduction has never been explored. In this paper, we use logic programming as a knowledge representation language. It is worth noting, however, that the logical framework of dishonest reasoning and its relationship to abduction do not depend on a particular logic.

The abstract framework proposed here is simple but expressive and capable of capturing different aspects of dishonest reasoning that arise in human society. Understanding when an agent behaves dishonestly and how dishonest reasoning is performed by agents is useful in identifying systems that mislead users, and providing ways to protect users from being deceived. Detecting dishonest agents and preventing a success of deception are issues for future research.

References

[Bonatti *et al.*, 1995] P. A. Bonatti, S. Kraus, and V. S. Subrahmanian. Foundations of secure deductive databases. *IEEE Trans. Knowl. Data Eng.*, 7(3):406–422, 1995.

[Caminada, 2009] M. Caminada. Truth, lies and bullshit, distinguishing classes of dishonesty. In: *Proc. IJCAI Workshop on Social Simulation*, 2009.

[Carson, 2010] T. L. Carson. *Lying and deception: theory and practice*. Oxford University Press, 2010.

[Eiter *et al.*, 1997] T. Eiter, G. Gottlob, and N. Leone. Abduction from logic programs: semantics and complexity. *Theoretical Computer Science*, 189(1-2):129–177, 1997.

[Franfurt, 2005] H. G. Frankfurt. *On Bullshit*. Princeton University Press, 2005.

[Gelfond and Lifschitz, 1991] M. Gelfond and V. Lifschitz. Classical negation in logic programs and disjunctive databases. *New Generation Computing*, 9:365–385, 1991.

[Inoue and Sakama, 1995] K. Inoue and C. Sakama. Abductive framework for nonmonotonic theory change. In: *Proc. IJCAI-95*, pp. 204–210, 1995.

[Kakas *et al.*, 1998] A.C.Kakas, R.A.Kowalski, and F.Toni. The role of abduction in logic programming. In: *Handbook of Logic in Artificial Intelligence and Logic Programming*, vol. 5, pp. 235–324, Oxford University Press, 1998.

[Mahon, 2007] J. E. Mahon. A definition of deceiving. *Journal of Applied Philosophy*, 21(2), 181–194, 2007.

[Mahon, 2008] J. E. Mahon. Two definitions of lying. *Journal of Applied Philosophy*, 22(2):211–230, 2008.

[Sakama and Inoue, 2003] C. Sakama and K. Inoue. An abductive framework for computing knowledge base updates. *Theory and Practice of Logic Prog.*, 3(6):671–713, 2003.

[Sakama *et al.*, 2010] C. Sakama, M. Caminada, and A. Herzig. A logical account of lying. In: *Proc. JELIA, LNAI* 634, pp. 286–299, Springer, 2010.

[Sklar *et al.*, 2005] E. Sklar, S. Parsons, and M. Davies. When is it okay to lie? A simple model of contradiction in agent-based dialogues. In: *Proc. ArgMas, LNCS* 3366, pp. 251–261, Springer, 2005.

[Zlotkin and Rosenschein, 1991] G. Zlotkin and J. S. Rosenschein. Incomplete information and deception in multi-agent negotiation. In: *Proc. IJCAI-91*, pp. 225–231, 1991.