Learning Dishonesty

Chiaki Sakama

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

Abstract. Children behave dishonestly as a way of managing problems in daily life. Then our primary interest of this paper is how children learn dishonesty and how one could model human acquisition of dishonesty using machine learning techniques. We first observe the structural similarities between dishonest reasoning and induction, and then characterize mental processes of dishonest reasoning using logic programming. We argue how one develops behavioral rules for dishonest acts and refines them to more advanced rules.

1 Introduction

"Lori, did you draw on your wall?" her mother asked, obviously upset.
"No," Lori answered, completely straight-faced.
"Well, who did it?"
"It wasn't me, Mommy," she replied, still the innocent angel.
"Was it a little ghost?" her mother asked sarcastically.
"Yeah, yeah," Lori said. "It was a ghost."

– Paul Ekman, "*Why kids lie*" [5]

Children learn dishonesty in their early ages. According to studies in psychology [2, 5], children lie by four years or earlier, mainly in order to avoid punishment. As they grow up, children use lies not only for protecting the self but for benefiting others. Victoria Talwar, who is an expert of children's lying behavior, says that "Lying is related to intelligence ... lying demands both advanced cognitive development and social skills that honesty simply doesn't require" [2]. Thus, learning dishonesty, including lying, is the process of socialization for children. Very young children start lying to their parents. In the dialog between a little girl and her mother at the beginning of the section, Lori, a three-and-a-half-year-old girl, scribbled on her bedroom wall with her crayons. She knows who did it, but denies her act in reply to the question of her mother.

The story is illustrative of several aspects of children's dishonest behavior. First, consider the reason why the little girl chose to lie. If her mother cheerfully asks the little girl that "Oh, what's a wonderful picture! Lori, did you draw it?", then her reply might be different. However, Lori observed that her mother appears displeased with the drawing. She then lied to avoid punishment. By this and other typical cases, we can say that *children come to behave dishonestly to*

avoid a unwanted outcome or to have a wanted outcome. Second, the little girl believes that she scribbled on the wall while behaves in a way that contradicts her belief. To resolve the inner conflict, she has to eliminate the believed-true fact "I scribbled on the wall" and instead introduce the believed-false fact "I did not scribble on the wall". Third, the little girl first just denied her act "It wasn't me". This is a simple and direct lie. Lori had no idea to whom she could impute the responsibility for the drawing. After Mom's sarcastic question, Lori put responsibility on a ghost. Compared with the first lie, "It was a ghost" is an indirect lie which is devised to cover the truth. Such an indirect lie requires advanced skills of creating a story. The little girl could not make the story herself but more aged children can do themselves.

Learning dishonesty is a process of human development in which children become adults. The topic has been studied by a number of researchers in the field of developmental psychology of children. On the other hand, the issue has been almost completely ignored in artificial intelligence and machine learning. Since the goal of machine learning research includes constructing a formal model of human learning and realizing the model on computers, it is important and meaningful to investigate the mechanism of learning dishonesty by humans. Then our question is: *How could one model human acquisition of dishonesty using machine learning techniques in AI?* The purpose of this study is providing an abstract framework for learning dishonesty. In this paper, we consider three different categories of dishonest acts. A *lie* is a statement of a believed-false sentence [1]. *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 [6]. *Withholding information* is to fail to offer information that would help someone acquire true beliefs and/or correct false beliefs [4]. We first characterize dishonest reasoning in terms of induction. We then argue how one acquires behavioral rules for dishonest acts and develops them to more advanced rules. To the best of our knowledge, this is the first attempt to formulate the process of learning dishonesty by humans using machine learning techniques.

The rest of this paper is organized as follows. Section 2 argues a logical formulation of dishonest reasoning. Section 3 provides connection between dishonest reasoning and induction. Section 4 characterizes behavioral rules for dishonest agents. Section 5 discusses related issues and Section 6 summarizes the paper.

2 Dishonest Reasoning

Dishonest reasoning made by a little girl in the introduction is represented in propositional logic. Lori initially believes that she draws something on the wall:

$$draw\_on\_wall. \tag{1}$$

She finds that Mom appears displeased with the drawing:

$$draw\_on\_wall \supset displeased\_Mom. \tag{2}$$

She experimentally believes that if she does something which makes Mom displeased, then she is scolded by Mom:

$$displeased\_Mom \supset scolded. \tag{3}$$

She wants to avoid to be scolded, then Lori decides to tell a lie. The process of reasoning by the little girl is formally presented as follows. Lori believes (1)–(3) which are represented as the background knowledge K:

$$\begin{aligned} K = \{\ & draw\_on\_wall, \\ & draw\_on\_wall \supset displeased\_Mom, \\ & displeased\_Mom \supset scolded\ \}. \end{aligned}$$

Mom asks whether she made the drawing, and Lori derives the fact $draw\_on\_wall$ in K. However, K also derives $scolded$ which she wants to avoid. Then, Lori removes the truth (1) and instead introduces the falsehood:

$$\neg\, draw\_on\_wall, \tag{4}$$

which results in the knowledge base K':

$$K' = (K \setminus \{\, draw\_on\_wall\,\}) \cup \{\neg\, draw\_on\_wall\}.$$

As a result, $K' \models \neg\, draw\_on\_wall$, and Lori replies "It wasn't me" to Mom.

Dishonest reasoning illustrated above is generalized in the following way. Suppose that a reasoner (or an agent) has background knowledge K and an unwanted outcome G (called *negative outcome*). When $K \models G$, the negative outcome G is obtained by honest reasoning. In this case, the agent tries to block the derivation of G by introducing disbelieved sentences I and removing believed sentences J:

$$(K \setminus J) \cup I \not\models G. \tag{5}$$

By contrast, suppose that an agent has background knowledge K and a wanted outcome G (called *positive outcome*). When $K \not\models G$, the positive outcome G is not obtained by honest reasoning. In this case, the agent tries to derive G by introducing disbelieved sentences I and removing believed sentences J:

$$(K \setminus J) \cup I \models G. \tag{6}$$

Here, I and J are sets of formulas which fill the gap between the current belief K of an agent and an (un)wanted outcome G. Dishonest reasoning is considered a process of revising K (not) to derive G. At this point, we observe some structural similarities between the process of dishonest reasoning and the problem of *induction* in machine learning. In fact, viewing G as a positive (resp. negative) evidence, the problem of constructing a new knowledge base $(K \setminus J) \cup I$ in (6) (resp. (5)) is considered a process of introducing a new hypothesis I and discarding an old belief J. [1]

[1] In the normal ILP setting, a hypothesis is only added to K to explain G, while it is more general to consider removal of current belief from K as well to explain G. This is especially the case when K is a nonmonotonic theory [8].

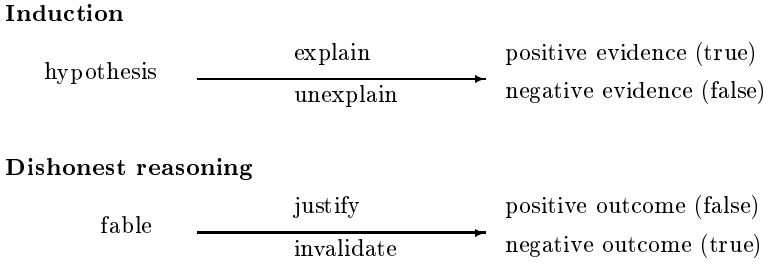

Fig. 1. induction vs. dishonest reasoning

An interesting contrast is that a positive evidence in induction is a true fact, and the goal is to construct a plausible hypotheses to explain it. On the other hand, a positive outcome in dishonest reasoning is a desired (and usually false) fact, and the goal is to invent a factitious story to justify it. Likewise, a negative evidence in induction is a false fact, and the goal is to construct a plausible hypotheses to unexplain it. On the other hand, a negative outcome in dishonest reasoning is an undesired (and usually true) fact, and the goal is to invent a factitious story to invalidate it. The correspondence between induction and dishonest reasoning is illustrated in Figure 1. Based on this intuition, we formulate dishonest reasoning as induction in the next section.

3 Dishonest Reasoning as Induction

An agent has a consistent propositional theory K as background knowledge. As usual, we identify a set of formulas with the conjunction of formulas included in the set. A consistent propositional formula is also called a *sentence*. A *positive outcome* is a sentence which an agent wants to obtain, while a *negative outcome* is a sentence which an agent wants to avoid.

Definition 3.1 (offensive/defensive dishonesty). Let K be background knowledge and G a positive outcome s.t. $K \not\models G$ (resp. a negative outcome s.t. $K \models G$). Suppose a pair (I, J) of sets of sentences satisfying the conditions:

1. $(K \setminus J) \cup I \models G$ (resp. $(K \setminus J) \cup I \not\models G$),
2. $(K \setminus J) \cup I$ is consistent.

Then,

(a) (I, J) (or simply I) is a *lie* for G if $I \neq \emptyset$ and $K \models \neg I$;
(b) (I, J) (or I) is *bullshit* (or BS) for G if $I \neq \emptyset$, $K \not\models \neg I$ and $K \not\models I$;
(c) (I, J) (or J) is *withholding information* (or WI) for G if $I = \emptyset$.

In each case, (I, J) is also called an *offensive dishonesty* for a positive outcome G (resp. a *defensive dishonesty* for a negative outcome G) wrt K.

In offensive dishonesty, a positive outcome G is not a consequence of a knowledge base K. To entail G, K is modified to a consistent theory $(K \setminus J) \cup I$ by removing a set J of believed sentences from K and introducing a set I of disbelieved sentences to K. In defensive dishonesty, on the other hand, a negative outcome G is a consequence of K. To invalidate G, K is modified to a consistent theory $(K \setminus J) \cup I$ that does not entail G.[2] In each case, (I, J) is called differently depending on its condition. I is called a lie if a nonempty set I is introduced to K and I is false in K. I is called bullshit if a nonempty set I is introduced to K and K has no belief wrt the truth nor falsity of I. J is called withholding information if I is empty (and hence, J is non-empty). Note that the entailment of a positive outcome (or non-entailment of a negative outcome) G 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 mislead another agent. The success of these acts depends on whether the opponent believes information I (or disbelieves J) brought by the proponent. We do not argue the effect of a dishonest act of an agent that is produced in another agent. By the definition, we have the following properties.

Proposition 3.1. *Lies, BS and WI are mutually exclusive.*

Proposition 3.2. *If G is a positive outcome, then $G \not\equiv false$. If G is a negative outcome, then $G \not\equiv true$.*

Proof. If a positive outcome is $G \equiv false$ or a negative outcome is $G \equiv true$, the conditions 1 and 2 of Definition 3.1 are contradictory with each other. □

Example 3.1. (a) Suppose the introductory example where Lori has the background knowledge:

$$\begin{aligned} K = \{ & \mathit{draw\_on\_wall}, \\ & \mathit{draw\_on\_wall} \supset \mathit{displeased\_Mom}, \\ & \mathit{displeased\_Mom} \supset \mathit{scolded} \}. \end{aligned}$$

To avoid the negative outcome $G = \mathit{scolded}$, Lori introduces the falsehood $I = \{\neg\, \mathit{draw\_on\_wall}\}$ to K and eliminates the fact $J = \{\mathit{draw\_on\_wall}\}$. As a result, $(K \setminus J) \cup I$ does not entail G. In this case, I is a defensive lie. Lori also creates another story $I' = \{\, \mathit{ghost\_did},\ \ \mathit{ghost\_did} \supset \neg\, \mathit{draw\_on\_wall}\,\}$. As $K \models \neg\, (\mathit{ghost\_did} \wedge (\mathit{ghost\_did} \supset \neg\, \mathit{draw\_on\_wall}))$, I' is also a defensive lie.

(b) Suppose another story that a little boy Tom is playing a game on the iPad. Mom asks him if there is any email from Dad. Tom considers that his game will be interrupted if an email has been received, but he wants to keep playing anyway. The belief state of Tom is represented by the background knowledge:

$$K = \{\, \mathit{mail} \supset \neg \mathit{playing},\ \ \neg\, \mathit{mail} \supset \mathit{playing}\,\}.$$

[2] When K is a propositional theory, the introduction of I monotonically increases theorems and does not eliminate G. On the other hand, when K is a nonmonotonic theory, the introduction of I can eliminate G. Although we consider (monotonic) propositional theory here, we provide a general setting for defensive dishonesty.

To have the positive outcome $G = playing$, Tom introduces the unknown fact $I = \{\neg\, mail\}$ to K. As a result, $K \cup I$ entails G. In this case, I is offensive BS. (Tom replies that there is no email without checking the mailbox.)

In Example 3.1(a), withholding the fact $J = \{draw\_on\_wall\}$ from K also has the effect of making G underived $K \setminus J \not\models G$. This is defensive WI. This corresponds to the situation where Lori says nothing in response to mother's question. Some researchers consider that there is not much difference between saying something false and concealing the truth [5]. However, these two acts are generally considered different dishonest acts [4], so we also distinguish them.

We have already observed structural similarities between dishonest reasoning and induction in Section 2. We reformulate computation of offensive/defensive dishonesty as induction problems as follows.

Offensive Dishonesty as Induction

Given : background knowledge K as a consistent propositional theory, and a positive outcome G as a positive evidence such that $K \not\models G$,
Find : a pair of sentences (I, J) such that $(K \setminus J) \cup I \models G$ where $(K \setminus J) \cup I$ is consistent.

Defensive Dishonesty as Induction

Given : background knowledge K as a consistent propositional theory, and a negative outcome G as a negative evidence such that $K \models G$,
Find : a pair of sentences (I, J) such that $(K \setminus J) \cup I \not\models G$ where $(K \setminus J) \cup I$ is consistent.

In each problem, we call (I, J) a *dishonest solution* for G under K.

With this setting, we can use induction algorithms of propositional theories for computing offensive/defensive dishonesty.

Proposition 3.3. *Given background knowledge K and a positive (resp. negative) outcome G, suppose that an induction algorithm finds a dishonest solution (I, J) for G under K. Then, if $I \neq \emptyset$ and $K \models \neg I$, then (I, J) is a lie. Else if $I \neq \emptyset$, $K \not\models \neg I$ and $K \not\models I$, then (I, J) is BS. Else if $I = \emptyset$, then (I, J) is WI.*

Proof. The result directly follows by Definition 3.1. □

Proposition 3.4. *Given background knowledge K and a positive (resp. negative) outcome G, suppose that an induction algorithm finds a dishonest solution (I, J) for G under K. Let (I', J') be any pair such that $I \subseteq I'$ and $J \subseteq J'$. Then,*

1. *If (I, J) is a lie, $(K \setminus J') \cup I' \models G$ for a positive outcome G (resp. $(K \setminus J') \cup I' \not\models G$ for a negative outcome G), and $(K \setminus J') \cup I'$ is consistent, then (I', J') is also a lie.*
2. *If (I, J) is BS, $(K \setminus J') \cup I' \models G$ for a positive outcome G (resp. $(K \setminus J') \cup I' \not\models G$ for a negative outcome G), and $(K \setminus J') \cup I'$ is consistent, then (I', J') is either a lie or BS.*

3. *If* $(\emptyset, J)$ *is a WI,* $K \not\models I'$, $(K \setminus J') \cup I' \models G$ *for a positive outcome* G *(resp.* $(K \setminus J') \cup I' \not\models G$ *for a negative outcome* G*), and* $(K \setminus J') \cup I'$ *is consistent, then* (I', J') *is either a lie, BS or WI.*

Proof. (1) If (I, J) is a lie, $K \models \neg I$ and $I \subseteq I'$ imply $K \models \neg I'$. Then, (I', J') is also a lie. (2) If (I, J) is BS, $K \not\models I$ and $I \subseteq I'$ imply $K \not\models I'$. Then, (I', J') is a lie if $K \models \neg I'$, otherwise, it is BS. (3) If $(\emptyset, J)$ is WI, there are three cases. When $I' = \emptyset$, (I', J') is WI. Else when $I' \neq \emptyset$, (I', J') is a lie if $K \models \neg I'$; otherwise, if $K \not\models \neg I'$, $K \not\models I'$ implies that (I', J') is BS. □

By Proposition 3.4, if a dishonest solution is computed, it could be extended for constructing another dishonest solution.

Example 3.2. In Example 3.1(a), a negative outcome G is proved in K. To compute (I, J) satisfying $(K \setminus J) \cup I \not\models G$, K is refined to K':

$$\begin{aligned} K' = \{\ & p \supset \mathit{draw\_on\_wall}, \\ & \mathit{draw\_on\_wall} \supset \mathit{displeased\_Mom}, \\ & \mathit{displeased\_Mom} \supset \mathit{scolded}\ \} \end{aligned}$$

where p is a newly introduced propositional sentence. Replacing $\mathit{draw\_on\_wall}$ by the sentence $p \supset \mathit{draw\_on\_wall}$ implies removing the fact $J = \{\mathit{draw\_on\_wall}\}$ from K, which is defensive WI $(\emptyset, \{\mathit{draw\_on\_wall}\})$. In this case, both $(\{\neg \mathit{draw\_on\_wall}\}, \{\mathit{draw\_on\_wall}\})$ and $(\{\mathit{ghost\_did},\ \mathit{ghost\_did} \supset \neg \mathit{draw\_on\_wall}\}, \{\mathit{draw\_on\_wall}\})$ are defensive lies.

In Example 3.1(b), a positive outcome G is not proved in K. To compute (I, J) satisfying $(K \setminus J) \cup I \models G$, by inverting the entailment relation as $(K \setminus J) \cup \{\neg G\} \models \neg I$, we can obtain $I = \{\neg \mathit{mail}\}$, which is offensive BS.

The refinement technique presented above is used in [11, 14], while the inverse entailment is used in [10].

4 Learning Dishonesty as Behavioral Rules

4.1 Behavioral rules

Children who dishonestly get away with a problem, are likely to behave dishonestly again to cope with similar problems. In the introductory example, Lori lied to her Mom to avoid punishment. If Mom overlooks her act and does not scold the little girl, then Lori would lie again to avoid punishment in another situation. Thus, children begin to act dishonestly in particular situations, then learn dishonest acts as behavioral rules to manage difficulties.

In this section, we characterize the process of learning behavioral rules of dishonesty using logic programming. Let us review the mental process of dishonest reasoning by the little girl in Example 3.1(a). First, Lori believes that the negative outcome *scolded* is deduced in her background knowledge K. She believes that the negative outcome is caused by her act *draw_on_wall*. She then

negates the fact and asserts $\neg draw\_on\_wall$. The act of defensive lying by the child is represented by the following meta-rule:

$$D\text{-}Lie(\neg draw\_on\_wall, scolded) \leftarrow neg(scolded), \\ prove(K,\ draw\_on\_wall \supset scolded),\ prove(K, draw\_on\_wall). \quad (7)$$

Here, $neg(G)$ means a negative outcome G, and $prove(K, F)$ holds iff $K \models F$. The rule (7) says if the negative outcome $scolded$ is deduced by $draw\_on\_wall$ in K, then defensively lie on the sentence $\neg draw\_on\_wall$ to avoid $scolded$.

Another day, Lori watches TV in the dining room. Mom noticed a puddle on the floor under her feet and asked if she wets her pants. She denies the fact to avoid punishment. The belief state of Lori is represented by the background knowledge:

$$K' = \{\ wet\_pants, \\ wet\_pants \supset displeased\_Mom, \\ displeased\_Mom \supset scolded\ \}.$$

To avoid the negative outcome $G = scolded$, Lori introduces the falsehood $I = \{\neg\, wet\_pants\}$ to K and eliminates the fact $J = \{wet\_pants\}$. Dishonest reasoning in this situation is represented by the meta-rule:

$$D\text{-}Lie(\neg wet\_pants, scolded) \leftarrow neg(scolded), \\ prove(K',\ wet\_pants \supset scolded),\ prove(K', wet\_pants). \quad (8)$$

Rules (7) and (8) represent two different situations for Lori not to be scolded. Using these rules, Lori can *induce* the new behavioral rule:

$$D\text{-}Lie(\neg F, scolded) \leftarrow neg(scolded),\ prove(K,\ F \supset scolded),\ prove(K, F) \quad (9)$$

where F is a variable representing any formula, or a more general rule:

$$D\text{-}Lie(\neg F, G) \leftarrow neg(G),\ prove(K,\ F \supset G),\ prove(K, F). \quad (10)$$

The rule (10) says if a negative outcome G is proved in background knowledge K using a believed-true sentence F, then lie on $\neg F$.[3] After obtaining such a behavioral rule, she can use the rule (10) to avoid negative outcomes in other circumstances. Suppose that a friend of Lori asks her to lend a toy. Lori does not want to lend it however. Then, she puts $neg(lend\_toy)$ and seeks a condition F which would deduce $lend\_toy$. If she finds the belief $have\_toy \wedge (have\_toy \supset lend\_toy)$ in her background knowledge, then she could lie on $\neg have\_toy$ ("I do not have it anymore"). Thus, once a behavioral rule is obtained by induction, it can be applied to individual situations by deduction.

A similar rule is obtained for offensive lying as

$$O\text{-}Lie(F, G) \leftarrow pos(G),\ not\ prove(K, G),\ prove(K,\ F \supset G),\ prove(K, \neg F) \quad (11)$$

[3] In (10), $prove(K, G)$ holds by $prove(K,\ F \supset G)$ and $prove(K, F)$.

where $pos(G)$ means a positive outcome G, and *not* is negation as failure to prove. The rule (11) says if a positive outcome G is not proved in background knowledge K but is proved using a believed-false sentence F, then lie on F. Rules are also constructed for offensive BS and defensive WI as follows:

$$O\text{-}BS(F,G) \leftarrow pos(G),\ not\ prove(K,G),\ prove(K,F \supset G),\ not\ prove(K,\neg F). \quad (12)$$

$$D\text{-}WI(F,G) \leftarrow neg(G),\ prove(K,F \supset G),\ prove(K,F). \quad (13)$$

Some facts are observed on these rules.

- Putting $F \equiv G$ in (10)–(13), we can obtain simplified rules:

  $$D\text{-}Lie(\neg G,G) \leftarrow neg(G),\ prove(K,G).$$
  $$O\text{-}Lie(G,G) \leftarrow pos(G),\ not\ prove(K,G),\ prove(K,\neg G).$$
  $$O\text{-}BS(G,G) \leftarrow pos(G),\ not\ prove(K,G),\ not\ prove(K,\neg G).$$
  $$D\text{-}WI(G,G) \leftarrow neg(G),\ prove(K,G).$$

  For instance, $D\text{-}Lie(\neg G,G)$ represents a defensive lie which just denies a negative outcome, and $O\text{-}Lie(G,G)$ represents an offensive lie which just asserts a positive outcome. These rules represent direct and unskillful lies. For instance, a child who has a homework but does not want to do it, lies "No homework".
- Comparing (11) and (12), $O\text{-}Lie(F,G)$ has the condition $prove(K,\neg F)$, while $O\text{-}BS(F,G)$ has the condition $not prove(K,\neg F)$. This means that a lier believes the falsehood of F, while a bullshitter has no belief on F.[4] A reasoner lies if $\neg F$ is proved in K, otherwise, he/she bullshits.
- Comparing (10) and (13), when a negative outcome G is proved using $F \supset G$ and F in K, a liar states $\neg F$ while a withholder just conceals F. That is, a reasoner can select one of the two dishonest acts under the same condition.

Note that we do not provide rules for defensive BS and offensive WI which do not happen in classical propositional theory. They would happen when background knowledge contains *nonmonotonic* rules. In defensive BS, $D\text{-}BS(F,G)$, a negative outcome G is proved in K in the absence of some formula F whose truth value is unknown. Then, a bullshitter states F to block the derivation of G. In offensive WI, $O\text{-}WI(F,G)$, on the other hand, a positive outcome G is not proved in K by the presence of a formula F. Then, a withholder conceals F to prove G. These two rules are useful when K is given as a nonmonotonic theory.

The meta-rules presented in this section instruct when to behave dishonestly. The process of taking each dishonest act is illustrated in Figure 2. In the figure, $A \rightarrow B$ means that if A holds then check B. We next consider rules for specifying how to behave dishonestly.

[4] In (12), $not\ prove(K,F)$ holds by $not\ prove(K,G)$ and $prove(K,F \supset G)$.

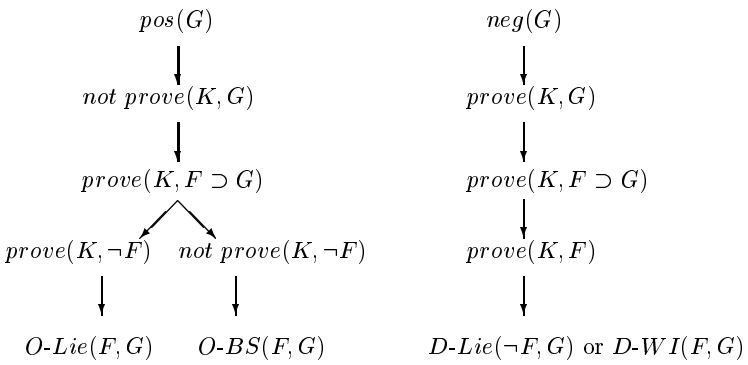

Fig. 2. behavioral rules for dishonest reasoning

4.2 Selecting the best dishonest act

In Example 3.1(a), to avoid the negative outcome $G = scolded$, Lori has two options for defensive lying: $I = \{\neg draw\_on\_wall\}$ and $I' = \{ghost\_did,\ ghost\_did \supset \neg draw\_on\_wall\}$. Then, which lie is better than the other? Comparing two lies, we can observe two facts. First, I' is stronger than I in the sense that I' implies I. Second, I' is more presumptive than I in the sense that I' requires an extra assumption $ghost\_did$. A liar normally wants to keep his/her lie as small as possible. This is because, 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. Moreover, a smaller lie is less sinful than a bigger one. In [12], it is represented as the maxim: "*Never tell an unnecessarily strong lie*". Then, the rule of defensive lying (10) would be refined to the following rule:

$$\begin{aligned} \textit{D-Lie}(\neg F, G) \leftarrow\ & \textit{neg}(G),\ \textit{prove}(K,\ F \supset G),\ \textit{prove}(K, F), \\ & \textit{prove}(K,\ \neg F \supset \neg H),\ \textit{not D-Lie}(\neg H, G). \end{aligned} \tag{14}$$

The rule (14) says if a negative outcome G is proved in background knowledge K using F and there is no defensive lie $\neg H$ for G which is weaker than $\neg F$, then defensively lie on $\neg F$. Children would acquire such a rule by empirically learning that a stronger (or a bigger) lie is more detectable than a weaker (or a smaller) one.

Similarly, preference rules for *O-Lie*, *O-BS*, and *D-WI* are constructed as follows.

$$\begin{aligned} \textit{O-Lie}(F, G) \leftarrow\ & \textit{pos}(G),\ \textit{not prove}(K, G),\ \textit{prove}(K,\ F \supset G), \\ & \textit{prove}(K, \neg F),\ \textit{prove}(K,\ F \supset H),\ \textit{not O-Lie}(H, G). \end{aligned} \tag{15}$$

$$\begin{aligned} \textit{O-BS}(F, G) \leftarrow\ & \textit{pos}(G),\ \textit{not prove}(K, G),\ \textit{prove}(K,\ F \supset G), \\ & \textit{not prove}(K, \neg F),\ \textit{prove}(K,\ F \supset H),\ \textit{not O-BS}(H, G). \end{aligned} \tag{16}$$

$$D\text{-}WI(F,G) \leftarrow neg(G),\ prove(K,\ F \supset G),\ prove(K,F),\\ prove(K,\ F \supset H),\ not\ D\text{-}WI(H,G). \tag{17}$$

As mentioned in the previous section, one can choose *D-Lie* (10) or *D-WI* (13) to avoid a negative outcome. Then, a question is which dishonest act should one to choose? Suppose that Lori believes that both a lie and WI are effective to avoid punishment. In this case, WI is considered preferable to lies because WI introduces no disinformation. Hence, $WI = (\emptyset, \{draw\_on\_wall\})$ is preferred to lies $(\{\neg draw\_on\_wall\}, \{draw\_on\_wall\})$ or $(\{ghost\_did,\ ghost\_did \supset \neg draw\_on\_wall\}, \{draw\_on\_wall\})$. Such preference is represented as the maxim: "*Never lie if you can have your way by withholding information*". The rule (10) is then modified as

$$D\text{-}Lie(\neg F,G) \leftarrow neg(G),\ prove(K,\ F \supset G),\\ prove(K,F),\ not\ D\text{-}WI(F,G). \tag{18}$$

The rule (18) represents that one selects the action $D\text{-}Lie(\neg F,G)$ if a negative outcome G is proved in background knowledge K and the action $D\text{-}WI(F,G)$ is not taken. Children would acquire such a preference by learning that providing false information is immoral and would lead to punishment when detected. Withholding information is also immoral in some cases, but is considered less sinful than lying. Such preferences are also considered between BS and lying. That is, BS is preferred to lying as the former provides unknown information, while the latter provides false information. In other words, "*Never lie if you can bullshit your way out of it*" [12].[5] The rule (11) is then modified as

$$O\text{-}Lie(F,G) \leftarrow pos(G),\ not\ prove(K,G),\\ prove(K,\ F \supset G),\ not\ O\text{-}BS(F,G). \tag{19}$$

The rule (19) represents that one selects the action $O\text{-}Lie(F,G)$ if a positive outcome G is not proved in background knowledge K and the action $O\text{-}BS$ is not taken. That is, if $not\ prove(K,\neg F)$ holds then take the action of $O\text{-}BS(F,G)$, otherwise, take the action of $O\text{-}Lie(F,G)$.

The rules (14)–(17) represent qualitative guidelines for selecting effective dishonest acts, while the rules (18) and (19) represent quantitative guidelines for selecting best dishonest acts.

4.3 Reasoning about the belief state of a hearer

As they grow up, children become more skilled and careful in acting dishonestly. Paul Ekman says: "A successful liar considers the perspective of the target being lied to. Taking the role of the other person, considering what will seem credible or suspicious to that person, allows the liar to consider the impact of his own behavior on the target and to fine-tune and adjust his behavior accordingly. ...

[5] A similar imperative is mentioned in [6].

Preschoolers aren't very good at this because at such early ages children don't realize that there is more than one perspective—theirs—on an event. They think everyone thinks the way they do. As they move toward adolescence, kids become much more able to put themselves in someone else's shoes" [5].

The behavioral rules in Section 4.1 describe when to behave dishonestly based on the speaker's belief base, while they do not take into account of the belief of the hearer. In Ekman's viewpoint, those rules characterize a very early-stage of dishonest acts by children in which they identify their own belief with the hearer's one. We then consider how to develop those behavioral rules in a way which distinguishes a hearer's belief base from the speaker's one and reasons about the belief state of a hearer.

Example 4.1. Suppose a school child, John, got a bad score at an exam. Mom asks whether he did well on the exam. John then considers: if Mom knows that he got a bad score, then Mom will make him study hard. To this end, Mom may restrict him from watching TV. However, John does not want this restriction. The belief state of John is represented by

$$\begin{aligned} K_O &= \{\ bad\_score\ \}, \\ K_S &= \{\ bad\_score\ \supset\ study\_hard, \\ &\qquad study\_hard\ \supset\ restrict\_TV\ \} \end{aligned}$$

where K_O represents objective facts, while K_S represents subjective beliefs with respect to the belief state of Mom. If John informs Mom of the objective fact bad_score, then $K_S \cup \{bad\_score\}$ derives $restrict\_TV$. To avoid this, he would not inform Mom of the truth.

Dishonest reasoning by this school boy is more advanced than the one by the little girl Lori. John distinguishes his belief on the objective fact (K_O) and the subjective belief on the belief state of Mom (K_S). In his objective belief K_O, John believes that he got a bad score. He then surmises the belief state of Mom in his subjective belief K_S and conjectures that Mom will restrict his watching TV if she knows the score. As a result, John would make a decision to act dishonestly. Definition 3.1 of offensive/defensive dishonesty is modified for such an advanced reasoner as follows. Let K_O be a knowledge base representing an agent's belief on the objective fact, and let K_S be a knowledge base representing an agent's subjective belief on the belief state of another agent. Suppose a positive outcome G s.t. $K_O \cup K_S \not\models G$ (resp. a negative outcome s.t. $K_O \cup K_S \models G$). Then, a pair (I, J) of sets of sentences satisfying the conditions:

1. $(K_O \setminus J) \cup K_S \cup I \models G$ (resp. $(K_O \setminus J) \cup K_S \cup I \not\models G$),
2. $(K_O \setminus J) \cup K_S \cup I$ is consistent

is called an *offensive dishonesty* for a positive outcome G (resp. a *defensive dishonesty* for a negative outcome G) wrt $K_O \cup K_S$. With this modified definition, the results of Section 3 hold as they are by identifying K with $K_O \cup K_S$.

The behavioral rules of such an advanced reasoner take the mental state of a hearer into consideration and will become more sophisticated than those

presented in Section 4.1. Suppose that there are two agents, a speaker a and a hearer b. To distinguish objective facts and subjective beliefs in the speaker's background knowledge K_a, we represent objective facts by sentences and subjective beliefs on the belief state of the hearer by $believe(b, F)$ which means "b believes F". With this setting, the behavioral rule for defensive lying is represented by the following rule:

$$D\text{-}Lie(\neg F, G) \leftarrow neg(G),\ prove(K_a, believe(b, F \supset G)),\ prove(K_a, F). \quad (20)$$

The rule (20) says if a believes that the believed-true sentence F leads b to believe a negative outcome G, then a defensively lies to b on $\neg F$. Comparing (10) and (20), $prove(K, F \supset G)$ in (10) is replaced by $prove(K_a, believe(b, F \supset G))$ in (20). This reflects the change that a speaker a considers the belief state of a hearer b and reasons by the hearer's viewpoint. Similarly, the behavioral rule of offensive lying is represented as the rule:

$$\begin{aligned} O\text{-}Lie(F, G) \leftarrow pos(G),\ & not\, prove(K_a, believe(b, G)), \\ & prove(K_a, believe(b, F \supset G)),\ prove(K_a, \neg F). \end{aligned} \quad (21)$$

The rule (21) says if a disbelieves that b believes a positive outcome G, and a believes that the believed-false sentence F leads b to believe G, then a offensively lies to b on F. Comparing (11) and (21), $not\, prove(K, G)$ in (11) is replaced by $not\, prove(K_a, believe(b, G))$ in (21), and $prove(K, F \supset G)$ in (11) is replaced by $prove(K_a, believe(b, F \supset G))$ in (21). Behavioral rules are also constructed for offensive BS and defensive WI as follows:[6]

$$\begin{aligned} O\text{-}BS(F, G) \leftarrow\ & pos(G),\ not\ prove(K_a, believe(b, G)), \\ & prove(K_a, believe(b, F \supset G)),\ not\ prove(K_a, F),\ not\ prove(K_a, \neg F). \end{aligned} \quad (22)$$

$$D\text{-}WI(F, G) \leftarrow neg(G),\ prove(K_a, believe(b, F \supset G)),\ prove(K_a, F). \quad (23)$$

As mentioned by Ekman, very young children do not distinguish between their own belief and others' belief. Then, they identify $prove(K_a, believe(b, F))$ with $prove(K_a, F)$. With this identification, the rules (21)–(23) reduce to those rules (11)–(13) of Section 4.1. Children come to realize that different people have different belief bases and another person does not always think the way they do. To achieve their own goals, children need to reason about belief of the hearer as done in this section and will acquire advanced rules for dishonest acts.

5 Discussion

5.1 Scientific Hypotheses and Dishonesty

In Section 2 we argued structural similarities between dishonest reasoning and induction. Those similarities are not a coincidence because hypotheses are sentences which have no logical reason to believe. Scientific discoveries have been

[6] In (22) "$not\ prove(K_a, F)$" represents that the agent a has no belief on the truth of F. This is in contrast to (12) in which this fact is implicitly represented in the rule (cf. footnote 4).

achieved by inventing hypotheses. However, history shows that a number of hypotheses turned out false or incorrect—Aristotle's theory of elements, Ptolemaic model of the universe, and Dalton's atomic theory, to name a few. Are those philosophers or scientists are liars? We could answer negatively to this question because it is unlikely that they believed the falsity of their own hypotheses. If a person states a believed-true sentence, which is in fact false, the person is not lying. Saint Augustine, who was a Berber philosopher and theologian, says: "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" [1, p.55].[7]

If they are not liars, then are they bullshitters? To answer this question, consider how scientists build hypotheses. Carl G. Hempel says: "Perhaps there are no objective criteria of confirmation; perhaps the decision as to whether a given hypothesis is acceptable in the light of a given body of evidence is no more subject to rational, objective rules than is the process of inventing a scientific hypothesis or theory; perhaps, in the last analysis, it is a "sense of evidence", or feeling of plausibility in view of the relevant data, which ultimately decides whether a hypothesis is scientifically acceptable. ... it is often deceptive, and can certainly serve neither as a necessary nor as a sufficient condition for the soundness of the given assertion" [7]. The truth of any hypothesis is unknown; if it turns true/false, it is not a hypothesis anymore. In this respect, one could argue that "They are bullshit, in the precise sense that we cannot prove them to be true, as they are things we have to *assume* so we *can* prove things to be (provisionally) true" [3]. Nevertheless, we consider that scientific hypotheses are distinguished from bullshit. Generally speaking, scientists are kind of people who are interested in the truth of the world. Then, they build hypotheses to account for observed data in the world. By contrast, a bullshitter "does not care whether the things he says describe reality correctly" [6]. Harry G. Frankfurt says: "It is just this lack of connection to a concern with truth—this indifference to how things really are—that I regard as of the essence of bullshit" [6]. As such, there is no aspect of truth-seeking in bullshit. This is the difference between scientific hypotheses and bullshit of dishonest reasoners.

5.2 Related Work

Sakama *et al.* [12] have introduced a formal model of dishonesty. The study formulate lies, bullshit and deception using a propositional multimodal logic, and compare their formal properties. In his continuous study, Sakama [13] characterizes dishonest reasoning using *extended abduction* [8] and provides a method of computing dishonest reasoning in terms of abductive logic programming. In [13] the notion of *logic programs with disinformation* (LPD) is introduced, which is defined as a pair $\langle K, \mathcal{D} \rangle$ where K is a program representing believed-true infor-

[7] It is interesting to note that young children consider lies differently. "Up until about eight years of age, children consider any false statement a lie, regardless of whether the person who said it knew it was false. Intention is not the issue—only whether information is false or true" [5].

mation and $\mathcal{D}$ is a set of disinformation representing believed-false or disbelieved-true facts. Then, lies, BS and WI are characterized by introducing facts from $\mathcal{D}$ and removing facts from K. The logical framework of [13] is similar to Definition 3.1 of this paper, but there is an important difference. An LPD assumes the pre-specified set $\mathcal{D}$ of ground literals and selects appropriate disinformation from this set for dishonest reasoning. This is also the case in abductive logic programming where an abductive logic program is given as a pair $\langle K, \mathcal{A} \rangle$ where $\mathcal{A}$ is a set of *abducibles*. The paper [13] restricts possible disinformation to the ground facts in $\mathcal{D}$, which enables to compute dishonest acts of agents using abductive logic programming. The logical framework of this paper relaxes the condition and does not prepare possible disinformation in advance. We use induction instead of abduction which makes it possible to fabricate a story as a set of sentences like $I' = \{ ghost\_did, \quad ghost\_did \supset \neg draw\_on\_wall \}$ of Example 3.1. We show possible computation of dishonest reasoning using existing induction algorithms (cf. Example 3.2), while appropriate inductive bias is needed for computing effective dishonest solutions. The paper [13] does not study how an agent develops behavioral rules of dishonesty nor how one refines those rules to more advanced ones.

To the best of our knowledge, there is no study which attempts to formulate the process of human acquisition of dishonesty. On the other hand, a recent study [9] simulates evolution and natural selection in robot learning. In this study, a group of robots have the task of finding a food source in a field. Once a robot found the food, it stays nearby and emits blue light, which informs other robots of the location and results in overcrowding around the food. After a few generations, robots become more secretive and learn to conceal food information for their own survival. The study shows the possibility of designing artificial agents which would acquire dishonest attitudes in their environment.

6 Conclusion

This paper focuses on the problem of children's learning to be dishonest. We related the process of fabricating a story in dishonest reasoning to the process of building a plausible hypothesis in induction. With this correspondence, there is a possibility of using induction algorithms for implementing learning dishonesty. We have also provided behavioral rules of dishonest reasoners which characterize when and how they would act dishonestly. Although those rules might represent the very early stage of dishonest acts by young children, those rules could be refined and sophisticated as argued in this paper. We also recognize that successful dishonest acts require various contingency plans including consideration of a plot to cover the truth and make a story credible, evaluation of the positive/negative effects of a dishonest act, preparation of responses to suspicious questions, etc. Such sophisticated planning matures with age, and children develop the ability as they grow older. The present paper focuses on the very early stage of learning dishonest acts and does not discuss the process of acquiring the ability of sophisticated planning. The issue is related to planning by dishonest agents and

left for future research. This paper targets an issue that has not been thoroughly addressed from an AI point of view and the present study is at an early stage of the research. Further study is needed for development of the proposed model and its evaluation in reality.

References

1. Augustine. Lying. In: *Treatises on Various Subjects, Fathers of the Church*, vol. 56, pp.45–110, 1952.
2. P. Bronson. Learn to lie. *New York Magazine*, February 2008.
3. R. G. Brown. Why science (natural philosophy) is bullshit. In: *Axioms*. `http://www.phy.duke.edu/~rgb/Philosophy/axioms/axioms/node44.html`, 2012.
4. T. L. Carson. *Lying and deception: theory and practice*. Oxford University Press, 2010.
5. P. Ekman. *Why kids lie: how parents can encourage truthfulness*. Scribner, 1989.
6. H. G. Frankfurt. *On Bullshit*. Princeton University Press, 2005.
7. C. G. Hempel. Studies in the logic of confirmation. In: B. A. Brody (Ed.), *Readings in the Philosophy of Science*, McGraw-Hill, 1970.
8. K. Inoue and C. Sakama. Abductive framework for nonmonotonic theory change. In: *Proc. 14th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 204–210, 1995.
9. S. Mitri, D. Floreano and L. Keller. The evolution of information suppression in communicating robots with conflicting interests. In: *Proc. National Academy of Sciences* 106(37), pp. 15786–15790, 2009.
10. S. Muggleton. Inverse entailment and Progol. *New Generation Computing* 13:245–286, 1995.
11. R. Quinlan. Learning logical definitions from relations. *Machine Learning* 5:239–266, 1990.
12. C. Sakama, M. Caminada and A. Herzig. A logical account of lying. In: *Proc. 12th European Conference on Logics in Artificial Intelligence (JELIA), Lecture Notes in Artificial Intelligence* 6341, pp. 286–299, Springer, 2010.
13. C. Sakama. Dishonest reasoning by abduction. In: *Proc. 22nd International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1063–1068, 2011.
14. E. Y. Shapiro. *Algorithmic Program Debugging*. MIT Press, Cambridge, MA, 1983.