Learning Dishonesties

"Lying is related to intelligence"
Po Bronson, *Learning to Lie (2008)*

Chiaki Sakama

Wakayama University, Japan

ILP 2012, Dubrovnik

Background & Motivation

- According to a study in psychology, children lie by four years or earlier, in order to avoid punishment. Learning dishonesty is the process of **socialization** for children.
- Our question: **How could one model human acquisition of dishonesty using machine learning techniques in AI?**
- To the best of our knowledge, no attempt is made to formulate the process of learning dishonesty of humans using machine learning techniques.

Different Categories of Dishonesties

- A **lie** is an act of an agent who states a believed-false fact to another agent.
- **Bullshit** is a statement that is grounded neither in a belief that it is true nor in a belief that it is not true.
- **Withholding information** is to fail to offer information that would help someone acquire true beliefs and/or correct false beliefs.
- We study how such dishonest acts are obtained as behavioral rules of agents.

Logical Framework

- A **program** consists of rules of the form:

  $L_0 \leftarrow L_1, \ldots, L_m, \textbf{not } L_{m+1}, \ldots, \textbf{not } L_n$

  where each L_i is a positive/negative literal, and **not** is negation as failure.

- The semantics of a program is given by its **answer sets**. If a literal L is included in every answer set of a program **K**, it is written as $\mathbf{K} \models L$.

- A **logic program with disinformation (LPD)** is a pair $\langle \mathbf{K}, \mathbf{D} \rangle$ where **K** is a program and **D** is a set of ground literals s.t. for any $L \in \mathbf{D}$ either $\mathbf{K} \models \neg L$ or ($\mathbf{K} \not\models L$ and $\mathbf{K} \not\models \neg L$).

Dishonest Reasoning by LPD

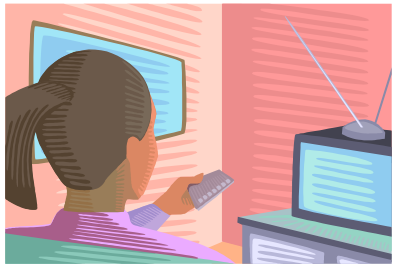
- $\langle K, D \rangle$: LPD, G: a ground literal representing a positive (or wanted) outcome s.t. $K \not\models G$
- Suppose a pair (I, J) of sets of ground literals satisfying:
  - $(K \setminus J) \cup I \models G$
  - $(K \setminus J) \cup I$ is consistent
  - $I \subseteq D$ and $J \subseteq K$
- Then, (I, J) is called an **offensive**
  - **lie** for G if $I \neq \emptyset$ and $K \models \neg L$ for some $L \in I$
  - **bullshit (or BS)** for G if $I \neq \emptyset$ and $K \not\models \neg L$ for any $L \in I$
  - **withholding information (or WI)** for G if $I = \emptyset$

Dishonest Reasoning by LPD

- **<K,D>**: LPD, G: a ground literal representing a negative (or unwanted) outcome s.t. **K** $\models$ G
- Suppose a pair **(I,J)** of sets of ground literals satisfying:
  - (**K** $\setminus$ **J**) $\cup$ **I** $\not\models$ G
  - (**K** $\setminus$ **J**) $\cup$ **I** is consistent
  - **I** $\subseteq$ **D** and **J** $\subseteq$ **K**
- Then, **(I,J)** is called an **defensive**
  - **lie** for G if **I** $\neq \emptyset$ and **K** $\models \neg$L for some L $\in$ **I**
  - **bullshit (or BS)** for G if **I** $\neq \emptyset$ and **K** $\not\models \neg$L for any L $\in$ **I**
  - **withholding information (or WI)** for G if **I** = $\emptyset$

Example

- One day, a child, Susie, watches TV. Mom asks whether she did her homework. Susie knows Mom permits her watching TV only when she finishes her work. Susie did not finish her work, but wants to keep watching TV.



- The belief state of Susie is represented by the LPD <**K**,**D**> such that

  K={ watchTV ← workDone, ¬workDone← },

  D={ workDone }.

- To have the positive outcome G=watchTV, Susie introduces the falsehood **I**={ workDone } to **K** and eliminates the fact **J**={¬workDone } from **K**.
- As a result, (**K** \ **J**) ∪ **I** ⊨ G and (**I**,**J**) is an offensive lie.

How Susie comes to a decision of lying?

- Susie knows that the positive outcome watchTV is not obtained in her background knowledge **K**.
- She seeks the possibility of getting watchTV and knows that the belief watchTV ← workDone in **K** would be used for this purpose.
- However, she also believes that workDone is false in **K**.
- Then, she lies on workDone in disinformation **D** to have the desired outcome watchTV.

Representing the act by a meta-rule

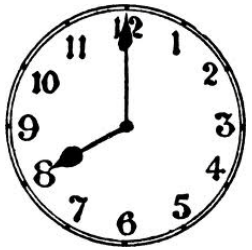
- The act of offensive lying by the child is represented by the meta-rule:

O-Lie(workDone) ← pos(watchTV), **not** prove(**K**, watchTV),
prove(**K**, watchTV ← workDone),
prove(**K**, ¬workDone), disinfo(workDone).

- pos(G) means a positive outcome G
- prove(**K**, F) holds iff **K** ⊨ F
- disinfo(F) means F ∈ **D**

Example (cont.)

- On another day, Susie watches TV, then Mom in the bathroom asks whether it is time to stop TV and go to bed. Susie knows she can watch TV if it is before 8 o'clock. Susie saw the clock and knows it is 8 now, but she wants to watch TV for a little while longer.



- The situation is represented by the LPD <**K′**,**D′**>

  K′={ watchTV ← ¬eight, eight← },

  D′={ ¬eight }.

- She then takes an action of lying using the meta-rule:

  O-Lie(¬eight) ← pos(watchTV), **not** prove(**K′**, watchTV),
  prove(**K′**, watchTV ← ¬eight),
  prove(**K′**, eight), disinfo(¬eight).

Learning Behavioral Rules

- The rules represent two different situations for a child to keep watching TV. Using these rules, Susie can **induce** the new behavioral rule:

  O-Lie(L) ← pos(watchTV), **not** prove(**K**, watchTV),
  prove(**K**, watchTV ← L), prove(**K**, ¬L), disinfo(L)

  where L is a variable representing any literal.

- Or she may induce a more general rule:

  O-Lie(L) ← pos(G), **not** prove(**K**, G),
  prove(**K**, G ← L), prove(**K**, ¬L), disinfo(L).

- The rule says if a positive outcome G is not proved in background knowledge **K** but is proved using a believed-false fact L, then **lie** on L.

Behavioral Rules for Dishonest Acts

- Similar rules are obtained for defensive lying, BS, and WI.

D-Lie(¬L) ← neg(G), prove(**K**, G←L),
prove(**K**, L), disinfo(¬L).

O-BS(L) ← pos(G), **not** prove(**K**, G), prove(**K**, G ← L),
not prove(**K**, L), **not** prove(**K**, ¬L), disinfo(L).

D-BS(L) ← neg(G), prove(**K**, G ← **not** L),
not prove(**K**, L), **not** prove(**K**, ¬L), disinfo(L).

O-WI(L) ← pos(G), **not** prove(**K**, G),
prove(**K**, G ← **not** L), prove(**K**, L).

D-WI(L) ← neg(G), prove(**K**, G ← L), prove(**K**, L).

where neg(G) means a negative outcome G.

O-Lie vs. O-BS

- O-Lie(L) has the condition prove(**K**, ¬L) while O-BS(L) has **not** prove(**K**, ¬L). Thus, a liar believes the falsefood of L while a bullshitter has no belief on L.

O-Lie(L) ← pos(G), **not** prove(**K**, G),
prove(**K**, G ← L), prove(**K**, ¬L), disinfo(L).

O-BS(L) ← pos(G), **not** prove(**K**, G), prove(**K**, G ← L),
not prove(**K**, L), **not** prove(**K**, ¬L), disinfo(L).

D-Lie vs. D-WI

- When a negative outcome G is proved by G←L and L in **K**, a liar states ¬L while a withholder just conceals L. Thus, a reasoner can select one of the two dishonest acts under the same condition.

D-Lie(¬L) ← neg(G), prove(**K**, G ← L), prove(**K**, L), disinfo(¬L).

D-WI(L) ← neg(G), prove(**K**, G ← L), prove(**K**, L).

D-BS vs. O-WI

- D-BS and O-WI are performed when background knowledge contains **nonmonotonic** rules.
- In D-BS(L), a negative outcome G is proved in **K** in the absence of L whose truth value is unknown. Then, a bullshitter states L to block the derivation of G.
- In O-WI(L), a positive outcome G is not proved in **K** by the presence of the true fact L. Then, a withholder conceals L to prove G.

D-BS(L) ← neg(G), prove(**K**, G ← **not** L),
not prove(**K**, L), **not** prove(**K**, ¬L), disinfo(L).

O-WI(L) ← pos(G), **not** prove(**K**, G),
prove(**K**, G ← **not** L), prove(**K**, L).

Remarks

- In the paper, we discussed preference rules for selecting "**best dishonest**" act, when different dishonest acts are possible to achieve a goal.
- In the longer version of this paper, we will develop an algorithm for computing behavioral rules for dishonest acts of agents.
- In this study, we focus on the very early stage of learning dishonesty. Learning more advanced skills (e.g., speculating the mental state of a hearer, planning most effective dishonest acts, etc) is left for future research.