DISHONEST REASONING BY ABDUCTION

Chiaki Sakama
Wakayama University

IJCAI-2011, Barcelona, July 2011

Background and Motivation

- People often behave **dishonestly** in daily life

- Few studies investigate inference mechanisms and computational methods of dishonest reasoning in AI

- This is a bit surprise because one of the goals of AI is to better understand human intelligence and to mechanize human reasoning

Contribution

- Exploring a computational logic for **dishonest reasoning**

- Formulating different types of dishonesty such as **lie, bullshit** and **withholding information**

- Characterizing dishonest reasoning in terms of **extended abduction**

Logic Programs with Disinformation

- **Logic Program**
  - A program **K** consists of rules of the form:

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

    where L_i is a literal and ***not*** is **negation as failure**
  - The semantics of **K** is given by its **answer sets**
  - **K** |= L if L is included in every answer set of **K**;
    K |= ⊥ if **K** has no answer set (or inconsistent)
- **Logic Program with Disinformation (or LPD)**
  - <**K**, **D**> with a program **K** and a set **D** of ground literals
  - For any L in **D**, either (a) **K** |= ¬L or (b) **K** |≠ L and **K** |≠ ¬L

Deductive Dishonesty

- Misleading another agent to **deduce** wrong conclusion
- Two different types of deductive dishonesty
  - **offensive dishonesty:** behave dishonestly to have a positive (or wanted) outcome that would not be gained by telling the truth
  - **defensive dishonesty:** behave dishonestly to avoid a negative (or unwanted) outcome that would not be gained by telling the truth
- In each case, an agent can perform different categories of dishonest reasoning
  - **Lie**: to tell a believed-false sentence
  - **Bullshit**: to tell a sentence that is neither believed to be true nor believed to be false
  - **Withholding Information**: to fail to offer information that would help someone acquire true beliefs and/or correct false beliefs

Offensive Dishonesty

- <**K**,**D**>: LPD, O^+: a ground literal representing a positive outcome s.t. $\mathbf{K} \not\models O^+$
- Suppose a pair (**I**,**J**) of sets of ground literals satisfying:
  - $(\mathbf{K} \setminus \mathbf{J}) \cup \mathbf{I} \models O^+$
  - $(\mathbf{K} \setminus \mathbf{J}) \cup \mathbf{I} \not\models \perp$
  - $\mathbf{I} \subseteq \mathbf{D}$ and $\mathbf{J} \subseteq \mathbf{K}$
- Then, (**I**,**J**) is called
  - **lie** for O^+ if $\mathbf{I} \neq \emptyset$ and $\mathbf{K} \models \neg L$ for some $L \in \mathbf{I}$
  - **bullshit (or BS)** for O^+ if $\mathbf{I} \neq \emptyset$ and $\mathbf{K} \not\models \neg L$ for any $L \in \mathbf{I}$
  - **withholding information (or WI)** for O^+ if $\mathbf{I} = \emptyset$

Example

- A salesperson believes that a product will be sold if the quality is good. However, he believes that the quality is not good.
- The situation is represented by an LPD $<\mathbf{K},\mathbf{D}>$ where

  $\mathbf{K}=\{\ sold \leftarrow quality.\ \ \neg quality \leftarrow .\ \}$ and $\mathbf{D}=\{quality\}$
- To have a positive outcome $O^+=\{sold\}$, he introduces $\mathbf{I}=\{quality\}$ to **K** and eliminates $\mathbf{J}=\{\neg quality\}$ from **K**.
- As a result, $(\mathbf{K} \setminus \mathbf{J}) \cup \mathbf{I} \models O^+$. In this case, $(\mathbf{I},\mathbf{J})$ is an offensive lie.

Defensive Dishonesty

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

Example

- A salesperson believes that an order will be canceled if the quality is not good. However, he has no information on the quality of the product.
- The situation is represented by an LPD $<\mathbf{K},\mathbf{D}>$ where

  $\mathbf{K}=\{\text{ canceled} \leftarrow \text{not quality }\}$ and $\mathbf{D}=\{\text{quality}\}$
- To avoid a negative outcome $O^{-}=\{\text{canceled}\}$, he introduces $\mathbf{I}=\{\text{quality}\}$ to **K**.
- As a result, $\mathbf{K} \cup \mathbf{I} \not\models O^{-}$. In this case, $(\mathbf{I},\emptyset)$ is defensive BS.

Abductive Dishonesty

- Interrupting another agent to **abduce** correct explanations
- Two different types of abductive dishonesty
  - **abductive dishonesty for positive evidences:** behave dishonestly to explain a positive evidence that is occurred
  - **abductive dishonesty for negative evidences:** behave dishonestly to explain a negative evidence that is not occurred
- In each case, an agent can perform different categories of dishonest reasoning – **Lie**, **BS** or **WI**
- A knowledge base of an agent includes a **secret set** of literals that he wants to conceal from another agent

Abductive Dishonesty for Positive Evidences

- <**K**,**D**>: LPD, Σ: a secret set, E^+: a ground literal representing a positive evidence s.t. $\mathbf{K} \models E^+$ and $\mathbf{K} \setminus \Sigma \not\models E^+$
- Suppose a pair (**I**,**J**) of sets of ground literals satisfying:
  - $(\mathbf{K} \setminus (\Sigma \cup \mathbf{J})) \cup \mathbf{I} \models E^+$
  - $(\mathbf{K} \setminus (\Sigma \cup \mathbf{J})) \cup \mathbf{I} \not\models \bot$
  - $\mathbf{I} \subseteq \mathbf{D}$ and $\mathbf{J} \subseteq \mathbf{K}$
- Then, (**I**,**J**) is called
  - **lie** for E^+ if $\mathbf{I} \neq \emptyset$ and $\mathbf{K} \models \neg L$ for some $L \in \mathbf{I}$
  - **bullshit (or BS)** for E^+ if $\mathbf{I} \neq \emptyset$ and $\mathbf{K} \not\models \neg L$ for any $L \in \mathbf{I}$
  - **withholding information (or WI)** for E^+ if $\mathbf{I} = \emptyset$

Example

- Sam is coming home late because he is cheating on his wife. Observing the late arrival, his wife might abduce his cheating. Sam does not want this abduction to take place, so he makes up another reason: he did overtime at work. He hopes this disinformation will stop her abduction.
- The situation is represented by an LPD $<\mathbf{K},\mathbf{D}>$ where

  $\mathbf{K}=\{\ late \leftarrow cheat.\ \ late \leftarrow overtime.\ \ cheat \leftarrow.\ \ \neg overtime \leftarrow.\ \}$, $\mathbf{D}=\{overtime\}$ and the secret set $\mathbf{\Sigma}=\{cheat\}$
- In face of the positive evidence $E^+ = late$, he introduces $\mathbf{I}=\{overtime\}$ to **K** and eliminates $\mathbf{J}=\{\neg overtime\}$ from **K**.
- As a result, $(\mathbf{K} \setminus (\mathbf{\Sigma} \cup \mathbf{J})) \cup \mathbf{I} \models E^+$. In this case, $(\mathbf{I},\mathbf{J})$ is an abductive lie.

Abductive Dishonesty for Negative Evidences

- $<\mathbf{K},\mathbf{D}>$: LPD, $\mathbf{\Sigma}$: a secret set, E^{-}: a ground literal representing a negative evidence s.t. $\mathbf{K} \not\models E^{-}$ and $\mathbf{K} \setminus \mathbf{\Sigma} \models E^{-}$
- Suppose a pair $(\mathbf{I},\mathbf{J})$ of sets of ground literals satisfying:
  - $(\mathbf{K} \setminus (\mathbf{\Sigma} \cup \mathbf{J})) \cup \mathbf{I} \not\models E^{-}$
  - $(\mathbf{K} \setminus (\mathbf{\Sigma} \cup \mathbf{J})) \cup \mathbf{I} \not\models \bot$
  - $\mathbf{I} \subseteq \mathbf{D}$ and $\mathbf{J} \subseteq \mathbf{K}$
- Then, $(\mathbf{I},\mathbf{J})$ is called
  - **lie** for E^{-} if $\mathbf{I} \neq \emptyset$ and $\mathbf{K} \models \neg L$ for some $L \in \mathbf{I}$
  - **bullshit (or BS)** for E^{-} if $\mathbf{I} \neq \emptyset$ and $\mathbf{K} \not\models \neg L$ for any $L \in \mathbf{I}$
  - **withholding information (or WI)** for E^{-} if $\mathbf{I} = \emptyset$

Example

- Sam and his wife promised to have a dinner at a restaurant. But Sam does not come to the restaurant on time because he is arguing with his girlfriend over the phone. Sam then excuses that he mistook the time.
- The situation is represented by an LPD <**K**,**D**> where

  K={ on-time←not call,remember. call←. remember←. }, **D**={remember} and the secret set **Σ**={call}
- In face of the negative evidence E^{-}= on-time, he eliminates **J**={remember} from **K**.
- As a result, $\mathbf{K} \setminus (\mathbf{\Sigma} \cup \mathbf{J}) \not\models E^{-}$. In this case, (∅, **J**) is an abductive WI.

Preference between Dishonesties

- **Quantitative Measure**: Comparing the same type of dishonesties, **the smaller the better**
- Let (**I**,**J**) and (**I'**,**J'**) be two lies/BS/WI for the same outcome/evidence. Then, (**I**,**J**) is **more or equally preferred to** (**I'**,**J'**) (written (**I**,**J**)≧(**I'**,**J'**)) if **I** ⊆ **I'** and **J** ⊆ **J'** .
  The most preferred one is called a **minimal dishonesty**.
- **Qualitative Measure**: Comparing different types of dishonesties,
  - **WI is preferable to BS and Lies**, since WI introduces no disinformation
  - **BS is preferable to Lies**, since BS is consistent with an agent's belief
- Let (**I1**,**J1**), (**I2**,**J2**) and (**I3**,**J3**) be a lie, BS and WI for the same outcome/evidence, respectively. Then, (**I3**,**J3**)>(**I2**,**J2**)>(**I1**,**J1**)

Extended Abduction [Inoue & Sakama, IJCAI-95]

- An **abductive program** is a pair <**K**,**A**> where **K** is a logic program and **A** is a set of ground literals representing hypotheses (called **abducibles**)
- Given a **positive observation** G^+ as a ground literal satisfying $\mathbf{K} \not\models G^+$, a pair (**I**,**J**) of sets of ground literals is an **explanation** of G^+ if (i) $(\mathbf{K} \setminus \mathbf{J}) \cup \mathbf{I} \models G^+$, (ii) $(\mathbf{K} \setminus \mathbf{J}) \cup \mathbf{I} \not\models \bot$, (iii) $\mathbf{I} \subseteq \mathbf{A} \setminus \mathbf{K}$ and $\mathbf{J} \subseteq \mathbf{A} \cap \mathbf{K}$
- Given a **negative observation** G^- as a ground literal satisfying $\mathbf{K} \models G^-$, a pair (**I**,**J**) of sets of ground literals is an **anti-explanation** of G^- if (i) $(\mathbf{K} \setminus \mathbf{J}) \cup \mathbf{I} \not\models G^-$, (ii) $(\mathbf{K} \setminus \mathbf{J}) \cup \mathbf{I} \not\models \bot$, (iii) $\mathbf{I} \subseteq \mathbf{A} \setminus \mathbf{K}$ and $\mathbf{J} \subseteq \mathbf{A} \cap \mathbf{K}$
- An (anti-)explanation (**I**,**J**) is **minimal** if $\mathbf{I'} \subseteq \mathbf{I}$ and $\mathbf{J'} \subseteq \mathbf{J}$ imply $\mathbf{I'}=\mathbf{I}$ and $\mathbf{J'}=\mathbf{J}$ for any (anti-)explanation (**I'**,**J'**)

Example

- Tweety is a bird and normally flies. One day an agent observes that Tweety does not fly. He then assumes that Tweety broke its wing.
- The situation is represented by an abductive program **<K,A>** where **K**={ flies←bird, not broken-wing. bird←.}, **A**={broken-wing}
- In this case, the negative observation G^{-}= flies has the anti-explanation (**I**,**J**)=({broken-wing}, ∅) s.t. **K** ∪ **I** $\not\models G^{-}$
- The agent revises **K** to **K'**=**K** ∪ {broken-wing}. After several days, the agent observes that Tweety flies as before. He then considers that the wound was healed.
- In this case, the positive observation G^{+}= flies has the explanation (**I**,**J**)=(∅, {broken-wing}) s.t. **K'** \ **J** $\models G^{+}$

Computing Dishonesties by Abduction

- There are structural similarities between deductive dishonesty and extended abduction
  - Viewing a positive outcome as a positive observation, an offensive dishonesty $(\mathbf{I},\mathbf{J})$ for the outcome wrt $\langle \mathbf{K},\mathbf{D} \rangle$ is identified with an explanation of the observation wrt $\langle \mathbf{K}, L(\mathbf{K}) \cup \mathbf{D} \rangle$ where $L(\mathbf{K})=\mathbf{K} \cap \mathbf{Lit}$ and $\mathbf{Lit}$ is the set of all ground literals in the language
  - Viewing a negative outcome as a negative observation, a defensive dishonesty $(\mathbf{I},\mathbf{J})$ for the outcome wrt $\langle \mathbf{K},\mathbf{D} \rangle$ is identified with an anti-explanation of the observation wrt $\langle \mathbf{K}, L(\mathbf{K}) \cup \mathbf{D} \rangle$
- Similar correspondences are observed between abductive dishonesty and extended abduction

Deductive Dishonesty vs. Extended Abduction

- <**K**,**D**>: LPD, O^+: positive outcome, O^-: negative outcome.
  - (**I**,**J**) is a (minimal) offensive dishonesty for O^+ wrt <**K**,**D**> iff (**I**,**J**) is a (minimal) explanation of O^+ wrt <**K**, L(**K**) ∪ **D**>
  - (**I**,**J**) is a (minimal) defensive dishonesty for O^- wrt <**K**,**D**> iff (**I**,**J**) is a (minimal) anti-explanation of O^- wrt <**K**, L(**K**) ∪ **D**>

Abductive Dishonesty vs. Extended Abduction

- <**K**,**D**>: LPD, E^+: positive evidence, E^-: negative evidence.
  - (**I**,**J**) is a (minimal) abductive dishonesty for E^+ wrt <**K**,**D**> iff (**I**,**J**) is a (minimal) explanation of E^+ wrt <**K** \ **Σ**, L(**K**) ∪ **D**>
  - (**I**,**J**) is a (minimal) abductive dishonesty for E^- wrt <**K**,**D**> iff (**I**,**J**) is a (minimal) anti-explanation of E^- wrt <**K** \ **Σ**, L(**K**) ∪ **D**>

Computational Complexities

- The following 3 decision problems are considered. Given a propositional LPD <**K**,**D**> and a positive/negative outcome/evidence X,
  - Does there **exist** a deductive/abductive dishonesty (**I**,**J**) for X?
  - Is a literal L is **relevent** to some (minimal) deductive/abductive dishonesty for the outcome/evidence? (i.e., $L \in \mathbf{I} \cup \mathbf{J}$ for some (**I**,**J**) for X)
  - Is a literal L is **necessary** for every (minimal) deductive/abductive dishonesty for the outcome/evidence? (i.e., $L \in \mathbf{I} \cup \mathbf{J}$ for every (**I**,**J**) for X)

Summary of Complexity Results

	Deductive Dishonesty Positive Outcome	Deductive Dishonesty Negative Outcome	Abductive Dishonesty Positive Evidence	Abductive Dishonesty Negative Evidence
Existence	Σ^P_3	Σ^P_2	Σ^P_3	Σ^P_2
Relevance (minimal)	Σ^P_3 (Σ^P_4)	Σ^P_2 (Σ^P_3)	Σ^P_3 (Σ^P_4)	Σ^P_2 (Σ^P_3)
Necessity (minimal)	Π^P_3 (Π^P_3)	Π^P_2 (Π^P_2)	Π^P_3 (Π^P_3)	Π^P_2 (Π^P_2)

*Each entry represents completeness for the respective class.

Final Remark

- Extended abduction is computed using **answer set programming** (ASP) [Sakama & Inoue, TPLP 2003]
- (Correspondence between dishonest reasoning and extended abduction)
  + (computation of extended abduction in ASP)
  ⇒ (computation of dishonest reasoning in ASP).
  (The method is provided in the paper).
- The logical framework of dishonest reasoning and its relationship to abduction do **not** depend on a particular logic.

Related Studies by the author and his colleagues

- Chiaki Sakama, Martin Caminada and Andreas Herzig
  A Logical Account of Lying.
  12th European Conference on Logics in AI (JELIA),
  LNAI 6341 , 2010.

- Chiaki Sakama and Martin Caminada
  The Many Faces of Deception.
  Thirty Years of Nonmonotonic Reasoning (NonMon@30),
  Lexington, KY, USA, October 2010.

- Chiaki Sakama
  Logical Definitions of Lying.
  14th International Workshop on Trust in Agent Societies,
  Taipei, Taiwan, May 2, 2011

Related Studies by the author and his colleagues

- Ngoc-Hieu Nguyen, Tran Son, Enrico Pontelli and Chiaki Sakama
  ASP-Prolog for Negotiation Among Dishonest Agents.
  11th International Conference on Logic Programming and Nonmonotonic Reasoning (LPNMR), LNAI 6645, 2011.
- Chiaki Sakama, Tran Son and Enrico Pontelli
  A Logical Formulation for Negotiation Among Dishonest Agents.
  IJCAI-2011, Barcelona (presented on Friday afternoon)
- Papers are available at the author's home page:

  http://www.wakayama-u.ac.jp/~sakama