A Logical Account of Lying

Chiaki Sakama
Wakayama University, Japan
Martin Caminada
University of Luxembourg, Luxembourg
Andreas Herzig
Universite' Paul Sabatier, France

JELIA 2010, Helsinki, Sept. 2010

Lying

- Lying is one of the basic behaviors of human beings.
- In spite of its familiarity, the question of "What is lying?" has been studied by a number of philosophers.
- Surprisingly, the topic has been almost completely ignored in artificial intelligence.

Why the study of Lying is important in AI?

- Lying is a linguistic behavior inherent to human beings, that requires intelligence and thinking.
- Understanding the mechanism of lying opens possibilities to develop computers that lie.
- Studying the act in the context of multiagent systems is necessary for designing intelligent agents.

Challenges for providing a formal account of lying

- There is no universally accepted definition of lying (even the definition in the OED is problematic).
- Little work exists for formulating lying.
- Formal logics are used for deriving sentences that are considered to be true, while this basic principle is not applied to an agent of lying.

Contributions of this study

- Various forms of lies are formulated using multimodal logic.
- Other categories of dishonesty, bullshit and deception, are formulated and contrasted with lying.
- Basic postulates for dishonesty are proposed for agents to satisfy for both moral and self-interested reasons.

A logic for belief and intention

- We consider a multimodal logic with two modalities: $B_a\phi$ (*a* believes a sentence ϕ) and $I_a\phi$ (*a* intends ϕ).
- The logic is an extension of $KD45_n$ and has the following axioms and inference rules.

(P) All propositional tautologies

(K_B) $B_a\phi \wedge B_a(\phi \supset \psi) \supset B_a\psi$ (K_I) $I_a\phi \wedge I_a(\phi \supset \psi) \supset I_a\psi$

(D_B) $B_a\phi \supset \neg B_a \neg \phi$ (D_I) $I_a\phi \supset \neg I_a \neg \phi$

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

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

(MP) $\phi, \phi \supset \psi / \psi$ (N_B) $\phi / B_a\phi$ (N_I) $\phi / I_a\phi$

- A speech act of an agent is represented by $utter_{ab}(\sigma)$ (*a* utters a sentence σ to *b*), which satisfies the axiom:
(U_{IB}) $utter_{ab}(\sigma) \supset I_a(utter_{ab}(\sigma)) \wedge B_a(utter_{ab}(\sigma))$

Lying: definition

- "To lie is to make a believed-false statement (to another person) with the intention that the statement be believed to be true (by the other person)."
  -- (Kupfer, 1982), (Mahon, 2008)

- Our formal definition: a, b: agents; σ: sentence
  $\mathrm{LIE}_{ab}(\sigma) = \mathrm{utter}_{ab}(\sigma) \wedge \mathrm{B}_a \neg\sigma \wedge \mathrm{I}_a \mathrm{B}_b \sigma$

- Note: the speaker believes $\neg\sigma$, but the truth of $\neg\sigma$ is not actually required.

Lying: properties

- One cannot lie on valid($\top$) or contradictory($\bot$) sentences.
  $LIE_{ab}(\top) \supset \bot$ and $LIE_{ab}(\bot) \supset \bot$
- A liar is aware of his/her act.
  $LIE_{ab}(\sigma) \supset B_a(LIE_{ab}(\sigma))$
- No one can lie to oneself.
  $LIE_{aa}(\sigma) \supset \bot$

When one has motives for lying?

- One lies to have a positive (or wanted) outcome that would not be gained by telling the truth.

  e.g. A salesperson lies about the quality of a product to convince a customer of buying the product.

- One lies to avoid a negative (or unwanted) outcome that would happen when telling the truth.

  e.g. A child lies about his performance in the exam to avoid punishment by his parents.

- The former is called an offensive lie, while the latter is called a defensive lie.

Offensive lie

- Def. *a, b:* agents; σ, ϕ: sentences.

$$\text{O-LIE}_{ab}(\sigma, \phi) = I_a B_b \phi \wedge \neg B_a B_b(\neg\sigma \supset \phi) \wedge B_a B_b(\sigma \supset \phi) \wedge \text{LIE}_{ab}(\sigma)$$

Offensive lie

- Def. *a, b:* agents; σ, ϕ: sentences.

$$\text{O-LIE}_{ab}(\sigma, \phi) = \text{I}_a\text{B}_b\,\phi \wedge \neg\text{B}_a\text{B}_b(\neg\sigma \supset \phi) \wedge \text{B}_a\text{B}_b(\sigma \supset \phi) \wedge \text{LIE}_{ab}(\sigma)$$

  - *a* has an intention to make *b* believe ϕ;

Offensive lie

- Def. *a, b:* agents; σ, ϕ: sentences.

$$\text{O-LIE}_{ab}(\sigma, \phi) = \text{I}_a\text{B}_b\,\phi \wedge \neg\text{B}_a\text{B}_b(\neg\sigma \supset \phi) \wedge \text{B}_a\text{B}_b(\sigma \supset \phi) \wedge \text{LIE}_{ab}(\sigma)$$

- *a* has an intention to make *b* believe ϕ;
- *a* disbelieves that the believed-true sentence ¬σ leads *b* to believe a positive outcome ϕ;

Offensive lie

- Def. *a, b:* agents; σ, ϕ: sentences.

$$\text{O-LIE}_{ab}(\sigma, \phi) = \text{I}_a\text{B}_b\,\phi \wedge \neg\text{B}_a\text{B}_b(\neg\sigma \supset \phi) \wedge \text{B}_a\text{B}_b(\sigma \supset \phi) \wedge \text{LIE}_{ab}(\sigma)$$

- *a* has an intention to make *b* believe ϕ;
- *a* disbelieves that the believed-true sentence ¬σ leads *b* to believe a positive outcome ϕ;
- *a* believes that the believed-false sentence σ leads *b* to believe ϕ;

Offensive lie

- Def. *a, b:* agents; σ, ϕ: sentences.

$$\text{O-LIE}_{ab}(\sigma, \phi) = I_a B_b \phi \wedge \neg B_a B_b(\neg\sigma \supset \phi) \wedge B_a B_b(\sigma \supset \phi) \wedge \text{LIE}_{ab}(\sigma)$$

- *a* has an intention to make *b* believe ϕ;
- *a* disbelieves that the believed-true sentence ¬σ leads *b* to believe a positive outcome ϕ;
- *a* believes that the believed-false sentence σ leads *b* to believe ϕ;
- *a* lies to *b* on σ.

Example

- A salesperson *a*'s belief:
  $B_a \neg$high_quality
- *a*'s belief wrt a customer *b*'s belief:
  $\neg B_a B_b(\neg$ high_quality $\supset$ buy)
  $\wedge$ $B_a B_b$(high_quality $\supset$ buy)
- *a* has the positive outcome ϕ=buy and have intention:
  $I_a B_b$ buy
- *a* lies to *b* on σ=high_quality:
  LIE_{ab}(high_quality)
- In this case, *a* offensively lies to *b* on σ to have ϕ:
  O-LIE_{ab}(high_quality, buy)

Defensive lie

- Def. a, b: agents; σ, ψ: sentences.

$$\text{D-LIE}_{ab}(\sigma, \psi) = \text{I}_a \neg \text{B}_b \psi \wedge \neg \text{B}_a \neg \text{B}_b (\neg \sigma \wedge \psi) \wedge \text{B}_a \neg \text{B}_b (\sigma \wedge \psi) \wedge \text{LIE}_{ab}(\sigma)$$

Defensive lie

- Def. a, b: agents; σ, ψ: sentences.

$$\text{D-LIE}_{ab}(\sigma, \psi) = \text{I}_a \neg \text{B}_b \psi \wedge \neg \text{B}_a \neg \text{B}_b (\neg \sigma \wedge \psi) \wedge \text{B}_a \neg \text{B}_b (\sigma \wedge \psi) \wedge \text{LIE}_{ab}(\sigma)$$

  - a has an intention to make b disbelieve ψ;

Defensive lie

- Def. a, b: agents; σ, ψ: sentences.

$$\text{D-LIE}_{ab}(\sigma, \psi) = I_a \neg B_b \psi \wedge \neg B_a \neg B_b (\neg\sigma \wedge \psi) \wedge B_a \neg B_b (\sigma \wedge \psi) \wedge \text{LIE}_{ab}(\sigma)$$

- a has an intention to make b disbelieve ψ;
- a considers it possible that b believes that the believed-true sentence $\neg\sigma$ and a negative outcome ψ hold at the same time;

Defensive lie

- Def. a, b: agents; σ, ψ: sentences.

$$\text{D-LIE}_{ab}(\sigma, \psi) = I_a \neg B_b \psi \wedge \neg B_a \neg B_b(\neg\sigma \wedge \psi) \wedge B_a \neg B_b(\sigma \wedge \psi) \wedge \text{LIE}_{ab}(\sigma)$$

- a has an intention to make b disbelieve ψ;
- a considers it possible that b believes that the believed-true sentence $\neg\sigma$ and a negative outcome ψ hold at the same time;
- a does not consider it possible that b believes that the believed-false sentence σ and ψ hold simultaneously;

Defensive lie

- Def. *a*, *b*: agents; σ, ψ: sentences.

$$\text{D-LIE}_{ab}(\sigma, \psi) = \text{I}_a \neg \text{B}_b \psi \wedge \neg \text{B}_a \neg \text{B}_b (\neg \sigma \wedge \psi) \wedge \text{B}_a \neg \text{B}_b (\sigma \wedge \psi) \wedge \text{LIE}_{ab}(\sigma)$$

- *a* has an intention to make *b* disbelieve ψ;
- *a* considers it possible that *b* believes that the believed-true sentence ¬σ and a negative outcome ψ hold at the same time;
- *a* does not consider it possible that *b* believes that the believed-false sentence σ and ψ hold simultaneously;
- *a* lies to *b* on σ.

Example

- A child *a*'s belief:
  $B_a \neg$high_score
- *a*'s belief wrt his mother *b*'s belief:
  $\neg B_a \neg B_b(\neg$ high_score $\wedge$ punish)
  $\wedge B_a \neg B_b$(high_score $\wedge$ punish)
- *a* has the negative outcome ψ=punish and have intention:
  $I_a \neg B_b$ punish
- *a* lies to *b* on σ= high_score :
  LIE_{ab}(high_score)
- In this case, *a* defensively lies to *b* on σ to avoid ψ:
  D-LIE_{ab}(high_score , punish)

O-Lies and D-Lies: remark

- To have a positive outcome ϕ, *a* can lie on ϕ
  $\text{O-LIE}_{ab}(\phi, \phi) \equiv \neg \text{B}_a \text{B}_b \phi \wedge \text{LIE}_{ab}(\phi)$
- To avoid a negative outcome ψ, *a* can lie on $\neg\psi$
  $\text{D-LIE}_{ab}(\neg\psi, \psi) \equiv \neg \text{B}_a \neg \text{B}_b \psi \wedge \text{LIE}_{ab}(\neg\psi)$
- O-LIE (resp. D-LIE) has additional objectives to have a positive outcome (resp. avoid a negative outcome).
- By contrast, lying in general is only aimed at making the hearer believe the uttered statement itself.

Different degrees of lies

- In lying, the success of the act depends on the belief state or knowledgeability of the hearer.
- For instance, it is easier to mislead children than adults, and it is more difficult to mislead experts than novices.
- Then, our question is: how different degrees of lies are used depending on a hearer's belief state that is believed by a speaker?

Stronger/Weaker lies

- A sentence σ is stronger than or equal to another sentence λ (or σ is weaker than or equal to λ) if $\sigma \supset \lambda$ (written $\sigma \geqq \lambda$).

- Let σ be an offensive lie for a positive outcome ϕ.
  - If $\sigma' \geqq \sigma$ implies $\sigma \geqq \sigma'$ for any offensive lie σ' for ϕ, σ is called a strongest offensive lie (denoted by σ_s).
  - If $\sigma \geqq \sigma'$ implies $\sigma' \geqq \sigma$ for any offensive lie σ' for ϕ, σ is called a weakest offensive lie (denoted by σ_w).

- The notion of strongest/weakest defensive lies is similarly defined.

Utilities of Different degrees of lies

- An agent *b* is knowledgeable not less than another agent *c* if $B_c\,\phi \supset B_b\,\phi$ holds for any ϕ.
- Suppose that *a* believes that *b* is knowledgeable not less than *c*. Then,
  $O\text{-}LIE_{ab}(\sigma_w,\phi) \wedge O\text{-}LIE_{ac}(\lambda,\phi) \supset \bot$ for any λ s.t. $\sigma_w > \lambda$
  "To have a positive outcome ϕ from *c*, *a* has to craft a lie that is not weaker than the weakest lie σ_w to *b*."

  $D\text{-}LIE_{ab}(\sigma_s,\psi) \wedge D\text{-}LIE_{ac}(\lambda,\psi) \supset \bot$ for any λ s.t. $\lambda > \sigma_s$
  "To avoid a negative outcome ψ from *c*, *a* can craft a lie that is not stronger than the strongest lie σ_s to *b*."

Example

- A salesperson *a*'s belief:
  B_a (¬high_quality ∧ ¬valuable)
- *a*'s belief wrt two customers *b* and *c*'s belief:
  B_b(high_quality ⊃ buy)
  B_c(high_quality ∧ valuable ⊃ buy)
  where $B_c\ \psi \supset B_b\ \psi$ holds for any ψ.
- To have the outcome ϕ=buy, *a* has to lie offensively to *c* on λ=high_quality ∧ valuable, which is stronger than σ=high_quality, to convince *c* to buy the product.

Deductive lies

- By an offensive lie (resp. a defensive lie), a speaker intends to mislead a hearer to deduce a wrong conclusion (resp. not to deduce a right conclusion).
- These types of lies are called deductive lies.
- By contrast, a person often lies in order to block another person for generating assumptions.

Example: Sam and his wife

- One day, Sam is coming home late because he is cheating on his wife.
- Based on the observation "Sam arrives late", his wife performs 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 will stop.

Abduction

- Abduction is the process of forming an explanatory hypothesis from an observation.
- Formally, let o be a sentence representing an observation and H a set of sentences representing a hypothesis.
- Given a background knowledge K and an observation o, a hypothesis H explains o in K if

$$K \wedge H \vdash o \text{ where } K \wedge H \text{ is consistent.}$$

Abductive lie

- K_a : a set of beliefs of an agent a
- Σ_a ($\subseteq K_a$): a secret set that a wants to conceal from b.
- A sentence $o \notin \Sigma_a$ is observed by two agents a and b.

A-Lie: definition

- $\text{A-LIE}_{ab}(\sigma, o) = B_a o \wedge B_a \neg B_b(\Delta \supset o)$
  $\wedge B_a(B_b(\Gamma \supset o) \wedge \neg B_b \neg \Gamma)$
  $\wedge B_a(B_b(\sigma \supset o) \wedge \neg B_b \neg \sigma)$
  $\wedge \bigwedge_{\gamma \in \Sigma_a} I_a \neg B_b \gamma \wedge \text{LIE}_{ab}(\sigma)$

  where $\Delta \subseteq K_a \setminus \Sigma_a$ and $\Gamma \cap \Sigma_a \neq \phi$.

A-Lie: definition

- $\text{A-LIE}_{ab}(\sigma, o) = B_a o \wedge B_a \neg B_b(\Delta \supset o)$
  $\wedge B_a(B_b(\Gamma \supset o) \wedge \neg B_b \neg \Gamma)$
  $\wedge B_a(B_b(\sigma \supset o) \wedge \neg B_b \neg \sigma)$
  $\wedge \bigwedge_{\gamma \in \Sigma_a} I_a \neg B_b \gamma \wedge \text{LIE}_{ab}(\sigma)$

  where $\Delta \subseteq K_a \setminus \Sigma_a$ and $\Gamma \cap \Sigma_a \neq \phi$.
- *a* believes o and that *b* does not explain o by believed-true sentences of *a* without some secret sentences;

A-Lie: definition

- $\text{A-LIE}_{ab}(\sigma, o) = B_a o \wedge B_a \neg B_b(\Delta \supset o)$
  $\wedge B_a(B_b(\Gamma \supset o) \wedge \neg B_b \neg \Gamma)$
  $\wedge B_a(B_b(\sigma \supset o) \wedge \neg B_b \neg \sigma)$
  $\wedge \bigwedge_{\gamma \in \Sigma_a} I_a \neg B_b \gamma \wedge \text{LIE}_{ab}(\sigma)$

  where $\Delta \subseteq K_a \setminus \Sigma_a$ and $\Gamma \cap \Sigma_a \neq \phi$.

- *a* believes o and that *b* does not explain o by believed-true sentences of *a* without some secret sentences;
- *a* believes that *b* explains o by either using some secret sentences of *a* or some believed-false sentence σ of *a*;

A-Lie: definition

- $\text{A-LIE}_{ab}(\sigma, o) = B_a o \wedge B_a \neg B_b(\Delta \supset o)$
  $\wedge B_a(B_b(\Gamma \supset o) \wedge \neg B_b \neg \Gamma)$
  $\wedge B_a(B_b(\sigma \supset o) \wedge \neg B_b \neg \sigma)$
  $\wedge \bigwedge_{\gamma \in \Sigma_a} I_a \neg B_b \gamma \wedge \text{LIE}_{ab}(\sigma)$

  where $\Delta \subseteq K_a \setminus \Sigma_a$ and $\Gamma \cap \Sigma_a \neq \phi$.

- *a* believes o and that *b* does not explain o by believed-true sentences of *a* without some secret sentences;
- *a* believes that *b* explains o by either using some secret sentences of *a* or some believed-false sentence σ of *a*;
- but *a* does not want *b*'s believing any sentence in Σ_a ;

A-Lie: definition

- $\text{A-LIE}_{ab}(\sigma, o) = B_a o \wedge B_a \neg B_b(\Delta \supset o)$
  $\wedge B_a(B_b(\Gamma \supset o) \wedge \neg B_b \neg \Gamma)$
  $\wedge B_a(B_b(\sigma \supset o) \wedge \neg B_b \neg \sigma)$
  $\wedge \bigwedge_{\gamma \in \Sigma_a} I_a \neg B_b \gamma \wedge \text{LIE}_{ab}(\sigma)$

  where $\Delta \subseteq K_a \setminus \Sigma_a$ and $\Gamma \cap \Sigma_a \neq \phi$.

- *a* believes o and that *b* does not explain o by believed-true sentences of *a* without some secret sentences;
- *a* believes that *b* explains o by either using some secret sentences of *a* or some believed-false sentence σ of *a*;
- but *a* does not want *b*'s believing any sentence in Σ_a ;
- *a* lies to *b* on σ.

Example

- Sam (a) has the set of beliefs:
  K_a ={ cheat, ¬overtime, cheat ⊃ late, overtime ⊃ late }
- Sam believes that his wife (b) has the belief:
  K_b ={ cheat ⊃ late, overtime ⊃ late }
- Sam wants to keep his cheating behavior secret:
  Σ_a ={ cheat }
- Given the observation o=late, Sam believes that his wife can abduce
  Γ_a ={ cheat }
- Then, Sam abductively lies to his wife on σ=overtime which explains late and would stop her abducing cheat.

Different degrees of abductive lies

- Suppose that *a* believes that *b* is knowledgeable not less than *c*, namely, $B_c\,\phi \supset B_b\,\phi$ for any ϕ. Then,

  $\text{A-LIE}_{ab}(\sigma_w, o) \wedge \text{A-LIE}_{ac}(\lambda, o) \supset \bot$ for any λ s.t. $\sigma_w > \lambda$

  "To explain o to *c*, *a* has to craft a lie that is not weaker than the weakest lie σ_w to *b*."

What are the most effective lies?

- In deductive/abductive lying, a number of lies exist to achieve a speaker's goal. Then a question is how good liars select “best lies”.
- A liar wants to keep his lie as small as possible because a stronger lie makes him less free in what he can do or say.
- Lies make the belief state of a hearer deviate from the objective reality and a stronger lie would increase such deviation. This is undesirable for a speaker because it increases the chance of the lie being detected.

Postulate I: Never tell an unnecessarily strong lie

σ, λ: sentences s.t. $\sigma > \lambda$.

- $B_a(\text{O-LIE}_{ab}(\sigma, \phi) \supset B_b\phi) \wedge B_a(\text{O-LIE}_{ab}(\lambda, \phi) \supset B_b\phi) \supset \neg\text{O-LIE}_{ab}(\sigma, \phi)$
- $B_a(\text{D-LIE}_{ab}(\sigma, \psi) \supset \neg B_b\psi) \wedge B_a(\text{D-LIE}_{ab}(\lambda, \psi) \supset \neg B_b\psi) \supset \neg\text{D-LIE}_{ab}(\sigma, \psi)$
- $B_a(\text{A-LIE}_{ab}(\sigma, o) \supset B_b o) \wedge B_a(\text{A-LIE}_{ab}(\lambda, o) \supset B_b o) \supset \neg\text{A-LIE}_{ab}(\sigma, o)$

Other categories of dishonesty

- 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" (Frankfurt 2005).

- "The production of bullshit is stimulated whenever a person's obligations or opportunities to speak about some topic exceed his knowledge of the facts that are relevant to that topic" (ibid).

Bullshit

- Def. a, b: agents; σ: sentence.

$$BS_{ab}(\sigma) = utter_{ab}(\sigma) \wedge \neg B_a\sigma \wedge \neg B_a\neg\sigma$$

- In contrast to lying,
  - the speaker a disbelieves $\neg\sigma$ as well as σ.
  - BS does not always require the intention of a speaker to make a hearer believe σ.
  - there is a freedom for a speaker to utter σ or $\neg\sigma$.

Example

- A financial consultant *a* is paid by the hour to provide advice to his client *b*.
- The consultant gives advice to buy stocks:
  $\text{utter}_{ab}(\text{buy\_stock})$
- Due to the lack of expertise, he has no belief concerning whether buying stocks is the best strategy or not:
  $\neg \text{B}_a\ \text{buy\_stock} \wedge \neg\ \text{B}_a \neg \text{buy\_stock}$
- In this case, *a* bullshits to *b* on σ:
  $\text{BS}_{ab}(\text{buy\_stock})$

Bullshit: properties

- One cannot BS about one's own (dis)beliefs.
  $BS_{ab}(B_a\sigma) \supset \bot$ and $BS_{ab}(\neg B_a\sigma) \supset \bot$
- One cannot BS on $\top$ or $\bot$.
  $BS_{ab}(\top) \supset \bot$ and $BS_{ab}(\bot) \supset \bot$
- A bullshitter is aware of his/her act.
  $BS_{ab}(\sigma) \supset B_a(BS_{ab}(\sigma))$
- BS and lies are mutually exclusive.
  $LIE_{ab}(\sigma) \wedge BS_{ab}(\sigma) \supset \bot$

Intentional BS vs. Unintentional BS

- Sometimes BS accompanies intention.

  e.g. A salesperson paid on commission basis may BS on the quality of products he is selling. He intends to make customers believe that the product has a high quality.

- Such intentional BS is defined as

  $$\text{I-BS}_{ab}(\sigma) = \text{BS}_{ab}(\sigma) \wedge \text{I}_a\text{B}_b\sigma$$

- In contrast to unintentional BS,
  - intentional BS to oneself is inconsistent.
  - *a* always utters σ that is intended to be believed by *b*.
  - like lying, offensive/defensive or deductive/abductive intentional BS could be considered.

What is best BS?

- Any BS is dishonest, but the consequence of faking is less severe for a weak BS than for a strong BS.
- The situation is formulated using offensive I-BS.

$$\text{O-BS}_{ab}(\sigma, \phi) = \text{I}_a\text{B}_b\,\phi \wedge \neg\text{B}_a\text{B}_b(\neg\sigma \supset \phi) \wedge \text{B}_a\text{B}_b(\sigma \supset \phi) \wedge \text{I-BS}_{ab}(\sigma)$$

- **Postulate II: Never tell unnecessary strong BS**

σ, λ: sentences s.t. $\sigma > \lambda$.

$$\text{B}_a\,(\text{O-BS}_{ab}\,(\sigma, \phi) \supset \text{B}_b\,\phi) \wedge \text{B}_a\,(\text{O-BS}_{ab}\,(\lambda, \phi) \supset \text{B}_b\,\phi) \supset \neg\text{O-BS}_{ab}\,(\sigma, \phi)$$

Lying vs. BS

- Lies and BS are two different forms of dishonesty, but lies are considered more sinful than BS.
- This is because a liar intentionally implants wrong beliefs at the hearer, while a bullshitter spits out statements, intentionally or not, without knowing if they are true.
- **Postulate III: Never lie if you can BS your way out of it**

$$B_a(\text{O-BS}_{ab}(\sigma, \phi) \supset B_b \phi) \wedge B_a(\text{O-LIE}_{ab}(\lambda, \phi) \supset B_b \phi) \supset \neg \text{O-LIE}_{ab}(\lambda, \phi)$$

Deception

- A speaker makes a believed-true statement with the intention that a hearer misuses it to reach a wrong conclusion (Adler 1997).
- A deceiver conceals something of the truth hoping that a hearer will make an incorrect inference based on incomplete beliefs.
- "With deception, one makes use of the nonmonotonic inference capabilities of the other person in order to implant wrong beliefs, without having to resort to lying ourselves" (Caminada 2009).

Deception

- *a*, *b*: agents; δ, σ: sentences s.t. $\delta \neq \sigma$.

$$\mathrm{DEC}_{ab}(\sigma, \delta) = \mathrm{utter}_{ab}(\sigma) \wedge \mathrm{B}_a\sigma \wedge \mathrm{I}_a\mathrm{B}_b\sigma \wedge \mathrm{B}_a\mathrm{B}_b\,((\sigma \wedge \neg\mathrm{B}_b\neg\delta) \supset \delta) \wedge \mathrm{B}_a\neg\mathrm{B}_b\neg\delta \wedge \mathrm{B}_a\neg\delta \wedge \mathrm{I}_a\mathrm{B}_b\delta$$

Deception

- *a*, *b*: agents; δ, σ: sentences s.t. $\delta \not\equiv \sigma$.

$$DEC_{ab}(\sigma, \delta) = utter_{ab}(\sigma) \wedge B_a\sigma \wedge I_aB_b\sigma \wedge B_aB_b\,((\sigma \wedge \neg B_b \neg \delta) \supset \delta) \wedge B_a \neg B_b \neg \delta \wedge B_a \neg \delta \wedge I_aB_b\delta$$

- *a* utters a believed-true sentence σ with the intention to make *b* believe it;

Deception

- *a*, *b*: agents; δ, σ: sentences s.t. $\delta \not\equiv \sigma$.

$$DEC_{ab}(\sigma, \delta) = utter_{ab}(\sigma) \wedge B_a\sigma \wedge I_aB_b\sigma \wedge B_aB_b\,((\sigma \wedge \neg B_b \neg \delta) \supset \delta) \wedge B_a \neg B_b \neg \delta \wedge B_a \neg \delta \wedge I_aB_b\delta$$

- *a* utters a believed-true sentence σ with the intention to make *b* believe it;
- *a* believes that *b* uses σ to reach a default conclusion δ;

Deception

- *a*, *b*: agents; δ, σ: sentences s.t. $\delta \not\equiv \sigma$.

$$DEC_{ab}(\sigma, \delta) = utter_{ab}(\sigma) \wedge B_a\sigma \wedge I_aB_b\sigma \wedge B_aB_b\,((\sigma \wedge \neg B_b \neg\delta) \supset \delta) \wedge B_a\neg B_b\neg\delta \wedge B_a\neg\delta \wedge I_aB_b\delta$$

- *a* utters a believed-true sentence σ with the intention to make *b* believe it;
- *a* believes that *b* uses σ to reach a default conclusion δ;
- *a* also believes that *b* disbelieves the falsity of δ while *a* believes it;

Deception

- *a*, *b*: agents; δ, σ: sentences s.t. $\delta \not\equiv \sigma$.

$$DEC_{ab}(\sigma, \delta) = utter_{ab}(\sigma) \wedge B_a\sigma \wedge I_aB_b\sigma \wedge B_aB_b\,((\sigma \wedge \neg B_b \neg \delta) \supset \delta) \wedge B_a \neg B_b \neg \delta \wedge B_a \neg \delta \wedge I_aB_b\delta$$

- *a* utters a believed-true sentence σ with the intention to make *b* believe it;
- *a* believes that *b* uses σ to reach a default conclusion δ;
- *a* also believes that *b* disbelieves the falsity of δ while *a* believes it;
- believing δ by *b* is what *a* intends to achieve.

Example

- John (a), who wants to marry his girlfriend Mary (b), tells her that he got a job at a company:
  $\text{utter}_{ab}(\text{get\_job}) \wedge B_a(\text{get\_job}) \wedge I_a B_b(\text{get\_job})$
- John intends that Mary will reach a default conclusion that John has a stable income (and she will agree with the marriage):
  $B_a B_b ((\text{get\_job} \wedge \neg B_b \neg \text{stable}) \supset \text{stable}) \wedge I_a B_b \text{ stable}$
- The company is almost bankrupt, however, and John believes that he would not get a stable income:
  $B_a \neg \text{stable}$
- But John does not tell Mary this fact, so that he believes that she disbelieves John's unstability:
  $B_a \neg B_b \neg \text{stable}$
- In this case, John deceives Mary on the fact get_job:
  $\text{DEC}_{ab}(\text{get\_job}, \text{stable})$

Deception: properties

- One cannot deceive on $\bot$ (but can deceive on $\top$).
  $DEC_{ab}(\bot, \delta) \supset \bot$
- A deceiver is aware of his/her act.
  $DEC_{ab}(\sigma, \delta) \supset B_a(DEC_{ab}(\sigma, \delta))$
- One cannot deceive oneself.
  $DEC_{aa}(\sigma, \delta) \supset \bot$
- Deception and lies/BS are mutually exclusive.
  $LIE_{ab}(\sigma) \wedge DEC_{ab}(\sigma, \delta) \supset \bot$
  $BS_{ab}(\sigma) \wedge DEC_{ab}(\sigma, \delta) \supset \bot$

What is best deception?

- In lying and bullshitting, it is reasonable not to lie and BS more than necessary (Postulates I & II).
- In deception, however, this is not necessarily the case. If *a* deceives *b* on the sentence $\sigma \wedge \lambda$, the deception is stronger than another deception σ.
- Providing more information implies concealing less information, which alleviates immoral feeling of the speaker.
- Thus, there is no reason to prefer the weakest form of deception, so we do not have a postulate mandating it.

Lying, BS, and deception

- Deception is considered preferable to lies and BS as a speaker utters a believed-true sentence.

- **Postulate IV: Never lie nor bullshit if you can deceive your way out of it.**

(i) $B_a\,(DEC_{ab}\,(\sigma, \delta) \supset B_b\,\delta) \;\wedge\; B_a\,(O\text{-}LIE_{ab}\,(\lambda, \delta) \supset B_b\,\delta) \supset \neg O\text{-}LIE_{ab}\,(\lambda, \delta)$

(ii) $B_a\,(DEC_{ab}\,(\sigma, \delta) \supset B_b\,\delta) \;\wedge\; B_a\,(O\text{-}BS_{ab}\,(\lambda, \delta) \supset B_b\,\delta) \supset \neg O\text{-}BS_{ab}\,(\lambda, \delta)$

Utilities of postulates

- The postulates I-IV are statements that agents should try to satisfy, both for moral reasons and for self-interested reasons (lower punishments if caught).
- If agents satisfy the dishonesty postulates, then one can characterize an agent by the worst level of dishonesty it is willing to commit to achieve a goal.
- For instance, an agent who is caught on deceiving can perhaps still be trusted not to lie, but an agent that is caught on lying cannot be trusted at all anymore.

Conclusion

- Formal analyses of various types of lies were provided and comparisons with different categories of dishonesty were made.
- Dishonesty postulates were provided, which, ideally implemented in multiagent systems, would help one to reason about the dishonesty of individual agents and about the extent to which they can still be trusted.
- Future work includes elaborating the formulation and building a formal system based on it.

Questions?