A Formal Model of Dishonest Communication

Chiaki Sakama
Wakayama University, Japan

European Summer School in Logic, Language and Information (ESSLLI)
Workshop on Formal Models of Communication
August 6-10, 2012, Opole, Poland

"Men are able to trust one another, knowing the exact degree of dishonesty they are entitled to expect."

- Paul Stephen Leacock (1869-1944, Canadian Humorist)

Dishonesty in Communication

- Dishonesty appears in human communication to benefit the self or avoid conflict.
- A study in social psychology shows that lying is a fact of daily life and 20% - 30% of conversation are lies.
- There are different categories of dishonesty: lies, bullshit, withholding information, half-truth, deception, etc.

Sincerity Rule

- In classical speech act theory (Searle, 1969), a speaker who makes an assertion is assumed to obey the sincerity rule: "the speaker commits himself to a belief in the truth of the expressed proposition".
- In the FIPA ACL, an agent is assumed to believe what it communicates.
  "In Arcol, an agent must make only sincere contributions and may assume that other agents also make only sincere contributions. Consequently, you cannot use Arcol in settings where sincerity cannot be taken for granted — for example, in electronic commerce or, broadly, in negotiation of any kind." (M.P. Singh: "Agent communication languages: rethinking the principles", 1998.)

Questions

- How to provide a formal model of dishonesty?
- What are the differences among different categories of dishonesty?
- Are there any behavioral rules which agents should obey in dishonest communication?

Why the study of dishonesty is important in AI?

- Dishonest behaviors are inherent to human beings, that require intelligence and thinking.

  "*Lying is related to intelligence … lying demands both advanced cognitive development and social skills that honesty simply doesn't require.*"

  - Victoria Talwar (an expert of children's behavior)

- Studies on dishonesty can contribute to better understanding human intelligence and take us one step closer to realizing "human-like" AI.

Why the study of dishonesty is important in AI?

- Understanding the mechanism of dishonesty opens possibilities to develop computers behaving dishonestly.

  - a robot for medical care which does not inform a patient of the true state of affairs
  - a database system which provides incorrect information to preserve security in a multi-user environment
  - an education system which presents false solutions in order to encourage students to find errors

Why the study of dishonesty is important in AI?

- Studying dishonest acts in the context of multiagent systems is necessary for designing "social" agents.

  - agents who conceal one's true position in automated negotiation
  - agents who behave economically to minimize costs and/or maximize benefits
  - agents who use dishonest arguments as strategies to win a debate game

Challenges for providing a formal model of dishonesty

- There are a number of philosophical studies on dishonesty, but relatively little work exists for formulating dishonesty.
- The notion of dishonesty is conceptual, which gives rise to the different definitions in the literature.
- A formal logic derives sentences that are true in the belief of an agent, while a dishonest agent behaves in a way that contradicts his/her belief.

Plan of this talk

- To formulate dishonest communication, we introduce a propositional multi-modal logic that can represent belief, intention and communication of an agent.
- We formulate lies, bullshit, withholding information and half-truths. We investigate formal properties and argue their connection to deception.
- Maxims for speech-acts of dishonesty are proposed that should ideally be satisfied by agents.

Logic for belief, intention and communication

- A propositional multimodal logic **BIC** has 3 modalities

  - $B_a\phi$: an agent *a* believes a sentence ϕ
  - $I_a\phi$: an agent *a* intends a sentence ϕ
  - $C_{ab}\phi$: an agent *a* communicates a sentence ϕ to (another) agent *b*

- The semantics of **BIC** is given by the Kripke semantics for normal modal operators.

Axiomatic system of BIC

1. All propositional tautologies
2. The axioms for B (the system KD45)

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

   (D_B) $\neg B_a \bot$

   (4_B) $B_a\phi \supset B_aB_a\phi$

   (5_B) $\neg B_a\phi \supset B_a\neg B_a\phi$

3. The axioms for I and C (the system KD)

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

   (D_I) $\neg I_a \bot$

   (K_C) $C_{ab}\phi \wedge C_{ab}(\phi \supset \psi) \supset C_{ab}\psi$

   (D_C) $\neg C_{ab} \bot$

Axiomatic system of BIC (cont.)

4. The bridge axioms among B, I and C:

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

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

(4_{CB}) $C_{ab}\phi \supset B_aC_{ab}\phi$

(5_{CB}) $\neg C_{ab}\phi \supset B_a\neg C_{ab}\phi$

(4_{CI}) $C_{ab}\phi \supset I_aC_{ab}\phi$

(5_{CI}) $\neg C_{ab}\phi \supset I_a\neg C_{ab}\phi$

5. Rules of inference:

(MP) If $\vdash \phi$ and $\vdash \phi \supset \psi$ then $\vdash \psi$

(N_B) If $\vdash \phi$ then $\vdash B_a\phi$

(N_I) If $\vdash \phi$ then $\vdash I_a\phi$

(N_C) If $\vdash \phi$ then $\vdash C_{ab}\phi$

where $\vdash \phi$ iff a sentence ϕ is a theorem of BIC.

Remark

- (N_I) says that every theorem holds at all state of affairs that *a* might intend to bring about. (N_C) says that every theorem is unconditionally communicated from *a* to *b*.

- The axiomatic system is sound and complete wrt the Kripke semantics.

- Each agent believes that other agents follow the same logic.

Lying

- There is no universally accepted definition of lying (even the definition in the OED is problematic).

"To make a false statement with the intention to deceive." (OED definition)

- "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"

  - Saint Augustine: 354-430, a Barber philosopher

Four Necessary Conditions for Lying

(J.E.Mahon: "The definition of lying and deception", *Stanford Encyclopedia of Philosophy*, 2008)

- Lying requires that a person make a statement (statement condition)
- Lying requires that the person believe the statement to be false (untruthfulness condition)
- Lying requires that the untruthful statement be made to another person (addressee condition)
- Lying requires that the person intend that that other person believe the untruthful statement to be true (intention to deceive addressee condition)

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)."
  (J.E. Mahon: "Two definitions of lying", 2008)

- Def. a, b: agents; σ: sentence
  $\mathrm{LIE}_{ab}(\sigma) =^{\mathrm{def}} \mathrm{C}_{ab}\,\sigma \wedge \mathrm{B}_a \neg\sigma \wedge \mathrm{I}_a \mathrm{B}_b \sigma$
  An agent a communicates a believed-false sentence σ to another agent b with the intention that b believes σ.

- The agent a believes $\neg\sigma$, but the truth of $\neg\sigma$ is not actually required (cf. OED definition).

Sincerity Condition

- A speech act of an agent is sincere if an agent communicates what he/she believes to be true.
- The sincerity condition is represented by the formula:

$$Sinc_{ab}(\sigma) =^{def} C_{ab}\,\sigma \supset B_a\sigma$$

- Lying is insincere.

$$\vdash LIE_{ab}(\sigma) \wedge Sinc_{ab}(\sigma) \supset \bot$$

Lying: properties

- One cannot lie on valid($\top$) or contradictory($\bot$) sentences.
  $\vdash \mathrm{LIE}_{ab}(\top) \supset \bot$ and $\vdash \mathrm{LIE}_{ab}(\bot) \supset \bot$
- Lying on combined sentences.
  $\vdash \mathrm{LIE}_{ab}(\sigma) \wedge \mathrm{LIE}_{ab}(\lambda) \supset \mathrm{LIE}_{ab}(\sigma \wedge \lambda)$
- Lying on contrary sentences.
  $\vdash \mathrm{LIE}_{ab}(\sigma) \wedge \mathrm{LIE}_{ab}(\neg\sigma) \supset \bot$
- A liar is aware of his/her act.
  $\vdash \mathrm{LIE}_{ab}(\sigma) \supset \mathrm{B}_{a}(\mathrm{LIE}_{ab}(\sigma))$

Lying to oneself

- An agent is self-consistent if it satisfies the formula:

$$Cons_a(\sigma) =^{def} B_a\sigma \supset \neg I_a B_a \neg\sigma$$

: if one believes something, then he/she does not intend to make oneself believe the contrary.

- When an agent is self-consistent, lying to oneself leads to contradiction.

$$\vdash LIE_{aa}(\sigma) \wedge Cons_a(\sigma) \supset \bot$$

Motives for lying

- An agent has a desired outcome that he/she wants to have.
- But he/she believes that the outcome would not be gained by telling true beliefs.
- On the other hand, he/she believes that the outcome would be gained by telling false beliefs.
- In this case, an agent has an incentive to lie.

Lying with objectives

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

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

Lying with objectives

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

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

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

Lying with objectives

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

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

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

Lying with objectives

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

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

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

Lying with objectives

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

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

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

Example

- A salesperson *a* is dealing with a customer *b*.
- *a* has an intention to make *b* buy some product, but disbelieves that *b* will buy it:
  $I_a B_b$ buy $\wedge \neg B_a B_b$ buy
- *a* believes *b* will buy the product if it has a high quality:
  $B_a B_b$(high_quality $\supset$ buy)
- *a* believes $\neg$ high_quality, then lies to *b* on the quality:
  LIE_{ab}(high_quality)
- In this case, O-LIE_{ab}(high_quality, buy) holds.

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".

  (H.G. Frankfurt: "On Bullshit", 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

  $BS_{ab}(\sigma) =^{def} C_{ab}\,\sigma \wedge \neg B_a\sigma \wedge \neg B_a\neg\sigma$

- 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 communicate 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:
  C_{ab} buy_stock
- Due to the lack of expertise, he has no belief concerning whether buying stocks is the best strategy or not:
  $\neg B_a$ buy_stock $\wedge \neg B_a \neg$buy_stock
- In this case, *a* bullshits to *b* on buying stocks:
  BS_{ab}(buy_stock)

Bullshit: properties

- One cannot BS about one's own (dis)beliefs.
  $\vdash BS_{ab}(B_a\sigma) \supset \bot$ and $\vdash BS_{ab}(\neg B_a\sigma) \supset \bot$
- One cannot BS on $\top$ or $\bot$.
  $\vdash BS_{ab}(\top) \supset \bot$ and $\vdash BS_{ab}(\bot) \supset \bot$
- BS is insincere.
  $\vdash BS_{ab}(\sigma) \wedge Sinc_{ab}(\sigma) \supset \bot$
  where $Sinc_{ab}(\sigma) = C_{ab}\,\sigma \supset B_a\sigma$
- A bullshitter is aware of his/her act.
  $\vdash BS_{ab}(\sigma) \supset B_a(BS_{ab}(\sigma))$

Bullshit: properties

- BS on combined sentences

  $\nvdash \ BS_{ab}(\lambda) \wedge BS_{ab}(\sigma) \supset BS_{ab}(\lambda \wedge \sigma)$

  (i.e., $B_a(\neg\lambda \vee \neg\sigma)$ is consistent with $BS_{ab}(\lambda) \wedge BS_{ab}(\sigma)$ but inconsistent with $BS_{ab}(\lambda \wedge \sigma)$.)

- BS on contrary sentences

  $\vdash \ BS_{ab}(\sigma) \wedge BS_{ab}(\neg\sigma) \supset \bot$

- BS and lies are mutually exclusive.

  $\vdash \ LIE_{ab}(\sigma) \wedge BS_{ab}(\sigma) \supset \bot$

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

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

- In contrast to unintentional BS, *a* always communicates σ that is intended to be believed by *b*.

Intentional BS with objectives

- Like lying, intentional BS with objectives is considered.
- Def. *a, b:* agents; σ, ϕ: sentences.

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

- *a* has an intention to make *b* believe ϕ;
- *a* disbelieves that *b* believes ϕ;
- *a* believes that the unknown sentence σ leads *b* to believe ϕ;
- *a* intentionally bullshits to *b* on σ.

Withholding Information

- "to withhold information is to fail to offer information that would help someone acquire true beliefs and/or correct false beliefs".
  (T.L. Carson: "Lying and Deception: theory and practice", Oxford Univ. Press, 2010)

- Def. a, b: agents; σ: sentence

  $WI_{ab}(\sigma) =^{def} \neg C_{ab}\,\sigma \wedge B_a\,\sigma \wedge I_a B_b \neg \sigma$

- In contrast to lying and BS, a does not communicate σ to b.

Withholding Information: properties

- One cannot WI on $\top$ or $\bot$.
  $\vdash \mathrm{WI}_{ab}(\top) \supset \bot$ and $\vdash \mathrm{WI}_{ab}(\bot) \supset \bot$
- WI is not insincere.
  $\nvdash \mathrm{WI}_{ab}(\sigma) \wedge \mathrm{Sinc}_{ab}(\sigma) \supset \bot$
- WI on combined sentences
  $\vdash \mathrm{WI}_{ab}(\lambda) \wedge \mathrm{WI}_{ab}(\sigma) \supset \mathrm{WI}_{ab}(\lambda \wedge \sigma)$
- WI on contrary sentences
  $\vdash \mathrm{WI}_{ab}(\sigma) \wedge \mathrm{WI}_{ab}(\neg\sigma) \supset \bot$

Withholding Information: properties

- One is aware of his/her act.
  $\vdash \mathrm{WI}_{ab}(\sigma) \supset \mathrm{B}_a(\mathrm{WI}_{ab}(\sigma))$
- A self-consistent agent cannot WI from oneself.
  $\vdash \mathrm{WI}_{aa}(\sigma) \wedge \mathrm{Cons}_a(\sigma) \supset \bot$
  where $\mathrm{Cons}_a(\sigma) = \mathrm{B}_a\sigma \supset \neg \mathrm{I}_a\mathrm{B}_a\neg\sigma$
- Lying implies WI.
  $\vdash \mathrm{LIE}_{ab}(\sigma) \supset \mathrm{WI}_{ab}(\neg\sigma)$
  ($\mathrm{LIE}_{ab}(\sigma)$ implies $\mathrm{C}_{ab}\,\sigma$, which implies $\neg\mathrm{C}_{ab}\,\neg\sigma$)
- One cannot BS and WI on the same sentence.
  $\vdash \mathrm{BS}_{ab}(\sigma) \wedge \mathrm{WI}_{ab}(\sigma) \supset \bot$

WI with objectives

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

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

- *a* has an intention to make *b* believe ϕ;
- *a* believes that *b*'s lacking information σ leads *b* to believe ϕ;
- *a* withholds information σ from *b*.

Half-Truth

- Half-truth is a partially true statement intended to deceive or mislead. (Collins English Dictionary)
- A speaker makes a believed-true statement with the intention that a hearer misuses it to reach a wrong conclusion which is desired by the speaker.
- A speaker believes that a hearer has incomplete belief and would make default reasoning to reach the conclusion.
- In this sense, it is also called "indirect lies" or "lying while saying the truth".

Half-Truth

- Def. a, b: agents; δ, σ: sentences

$$HT_{ab}(\sigma, \delta) =^{def} C_{ab}\,\sigma \wedge B_a\sigma \wedge B_a\neg B_b\neg\sigma \wedge I_aB_b\sigma$$
$$\wedge\, B_aB_b\,(\,(\sigma \wedge \neg B_b\neg\delta) \supset \delta\,)$$
$$\wedge\, B_a\neg B_b\neg\delta \wedge B_a\neg\delta$$
$$\wedge\, \neg C_{ab}\neg\delta \wedge I_aB_b\delta$$

Half-Truth

$$HT_{ab}(\sigma, \delta) = C_{ab}\,\sigma \wedge B_a\sigma \wedge B_a\neg B_b\neg\sigma \wedge I_aB_b\sigma$$
$$\wedge\, B_aB_b\,(\,(\sigma \wedge \neg B_b\neg\delta) \supset \delta\,)$$
$$\wedge\, B_a\neg B_b\neg\delta \wedge B_a\neg\delta$$
$$\wedge\, \neg U_{ab}\neg\delta \wedge I_aB_b\delta$$

- *a* communicates a believed-true sentence σ, which *a* believes that *b* considers it possible, with the intention to make *b* believe it;

Half-Truth

$$HT_{ab}(\sigma, \delta) = C_{ab}\,\sigma \wedge B_a\sigma \wedge B_a \neg B_b \neg \sigma \wedge I_a B_b \sigma$$
$$\wedge\, B_a B_b\,((\sigma \wedge \neg B_b \neg \delta) \supset \delta)$$
$$\wedge\, B_a \neg B_b \neg \delta \wedge B_a \neg \delta$$
$$\wedge\, \neg C_{ab} \neg \delta \wedge I_a B_b \delta$$

- *a* communicates a believed-true sentence σ, which *a* believes that *b* considers it possible, with the intention to make *b* believe it;
- *a* believes that *b* uses σ to reach a default conclusion δ;

Half-Truth

$$HT_{ab}(\sigma, \delta) = C_{ab}\,\sigma \wedge B_a\sigma \wedge B_a \neg B_b \neg \sigma \wedge I_a B_b \sigma$$
$$\wedge B_a B_b \left((\sigma \wedge \neg B_b \neg \delta) \supset \delta \right)$$
$$\wedge B_a \neg B_b \neg \delta \wedge B_a \neg \delta$$
$$\wedge \neg C_{ab} \neg \delta \wedge I_a B_b \delta$$

- *a* communicates a believed-true sentence σ, which *a* believes that *b* considers it possible, 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;

Half-Truth

$$HT_{ab}(\sigma, \delta) = C_{ab}\,\sigma \wedge B_a\sigma \wedge B_a \neg B_b \neg \sigma \wedge I_a B_b \sigma$$
$$\wedge B_a B_b\,((\sigma \wedge \neg B_b \neg \delta) \supset \delta)$$
$$\wedge B_a \neg B_b \neg \delta \wedge B_a \neg \delta$$
$$\wedge \neg C_{ab} \neg \delta \wedge I_a B_b \delta$$

- *a* communicates a believed-true sentence σ, which *a* believes that *b* considers it possible, 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;
- *a* does not communicate ¬δ with the intention that *b* believes δ.

Example

- John (a), who wants to marry his girlfriend Mary (b), tells her that he got a permanent job at a company:
  $C_{ab}\,\mathrm{job} \wedge B_a\,\mathrm{job} \wedge B_a \neg B_b \neg\,\mathrm{job} \wedge I_a B_b\,\mathrm{job}$
- John expects that Mary concludes by default that he gets a stable income:
  $B_a B_b\,(\,(\mathrm{job} \wedge \neg B_b \neg \mathrm{stable}) \supset \mathrm{stable}\,)$
- The company is almost bankrupt, and John believes that he would not get a stable income while she disbelieves it:
  $B_a \neg \mathrm{stable} \wedge B_a \neg B_b\, \neg \mathrm{stable}$
- John does not tell Mary this fact and intends to make her believe his stable income:
  $\neg\, C_{ab} \neg \mathrm{stable} \wedge I_a B_b\, \mathrm{stable}$
- In this case, $HT_{ab}(\mathrm{job}, \mathrm{stable})$ holds.

HT: properties

- One cannot do HT on $\bot$ (but can do on $\top$).
  $\vdash \mathrm{HT}_{ab}(\bot, \delta) \supset \bot$
- HT is not insincere.
  $\nvdash \mathrm{HT}_{ab}(\sigma, \delta) \wedge \mathrm{Sinc}(\sigma) \supset \bot$
- HT on combined sentences
  $\vdash \mathrm{HT}_{ab}(\lambda, \delta) \wedge \mathrm{HT}_{ab}(\sigma, \delta) \supset \mathrm{HT}_{ab}(\lambda \vee \sigma, \delta)$
- HT on contrary sentences
  $\vdash \mathrm{HT}_{ab}(\sigma, \delta) \wedge \mathrm{HT}_{ab}(\neg\sigma, \delta) \supset \bot$

HT: properties

- impossible HT
  $\vdash \mathrm{HT}_{ab}(\sigma, \sigma) \supset \bot$
- One is aware of his/her act.
  $\vdash \mathrm{HT}_{ab}(\sigma, \delta) \supset \mathrm{B}_{a}(\mathrm{HT}_{ab}(\sigma, \delta))$
- One cannot HT to oneself (even in the absence of Cons)
  $\vdash \mathrm{HT}_{aa}(\sigma, \delta) \supset \bot$
- Relationship between lie, BS, WI and HT.
  $\vdash \mathrm{LIE}_{ab}(\sigma) \wedge \mathrm{HT}_{ab}(\sigma, \delta) \supset \bot$
  $\vdash \mathrm{BS}_{ab}(\sigma) \wedge \mathrm{HT}_{ab}(\sigma, \delta) \supset \bot$
  $\vdash \mathrm{HT}_{ab}(\sigma, \delta) \supset \mathrm{WI}_{ab}(\neg\delta)$

HT with objectives

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

  $O\text{-}HT_{ab}(\sigma, \delta, \phi) =^{def} I_a B_b \phi \wedge \neg B_a B_b \phi \wedge B_a B_b(\delta \supset \phi) \wedge HT_{ab}(\sigma, \delta)$

- *a* has an intention to make *b* believe ϕ;
- *a* disbelieves that *b* believes ϕ;
- *a* believes that the believed-false sentence δ leads *b* to believe ϕ;
- *a* tells *b* a half-truth sentence σ which would lead *b* to a default conclusion δ.

Grice's maxims for conversation
(Paul Grice: "Studies in the ways of words", 1989)

- The maxim of quality
  - Do not say what you believe to be false.
  - Do not say that for which you lack adequate evidence.

Violated by Lying

Violated by BS

- The maxim of quantity
  - Make your contribution as informative as is required.
  - Do not make your contribution more informative than is required.

Violated by WI and HT

- The maxim of relation: Be relevant.
- The maxim of manner: Avoid obscurity of expression; Avoid ambiguity; Be brief; and Be orderly.

Quantitative guidelines for lying

- A smaller lie would be less sinful than a bigger one from the moral viewpoint.
- From self-interested reasons, a smaller lie would cause less personal discomfort and result in lower criticism or punishment if detected.
- From practical reasons, a bigger lie increases the chance of the lie being detected.

Quantitative maxims for dishonest agents

- Maxim I: Never lie unnecessarily

  λ, σ, ϕ : sentences s.t. $\nvdash \lambda \supset \sigma$.

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

An agent should abstain from uttering the lie $\lambda \wedge \sigma$ in case he/she believes that a simpler lie λ succeeds in persuading a hearer of believing ϕ.

Quantitative maxims for dishonest agents

- Maxim II: Never bullshit unnecessarily

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

- Maxim III: Never withhold info. unnecessarily

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

Remark

- In lying, BS and WI, it is reasonable (and courteous to a hearer) not to lie, BS, and WI more than absolutely necessary (Maxims I －III).
- On the other hand, providing more information in HT increases the knowledge of a hearer, which implies concealing less information for a speaker.
- There seems no reason to prefer a smaller HT that provides less information, so we do not have a maxim mandating it.

Lying vs. BS

- Lies are considered more sinful than BS 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.
- “people do tend to be more tolerant of bullshit than of lies, perhaps because we are less inclined to take the former as a personal affront" .

(H.G. Frankfurt, “On Bullshit”, 2005)

Qualitative maxims for dishonest agents

- Maxim IV: Never lie if you can bullshit your way out of it

λ, σ, ϕ : sentences s.t. $\lambda \not\equiv \sigma$

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

An agent should abstain from lying on σ in case he/she believes that BS λ succeeds in persuading a hearer of believing ϕ.

Lying, BS vs. WI vs. HT

- When both a lie and WI (or both BS and WI) are effective to achieve an objective, WI is preferable to lie or BS because WI introduces no disinformation.

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

Qualitative maxims for dishonest agents

- Maxim V: Never lie nor BS if you can make your way by WI

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

where O-LIE can be replaced by O-BS.

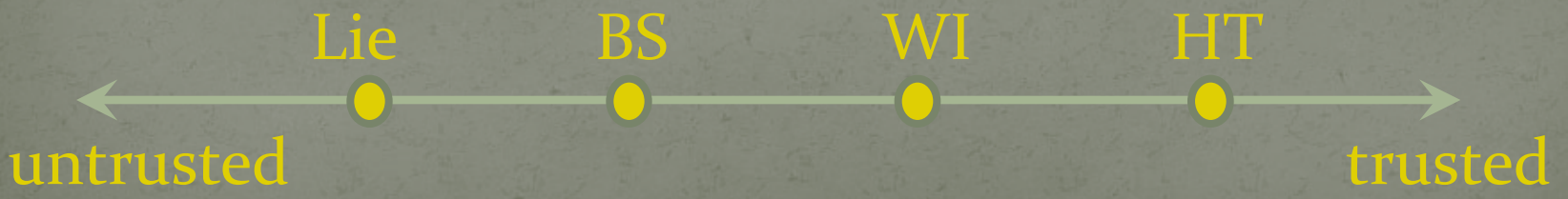
- Maxim VI: Never lie, BS, nor WI if you can make your way by HT

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

where O-LIE can be replaced by O-BS or O-WI.

Utilities of maxims

- In multiagent systems if agents have implemented the dishonesty maxims, then this helps one to reason about the possible dishonesty of other agents, and about the extent to which they can still be trusted.
- For instance, an agent who is caught on HT can perhaps still be trusted not to lie, BS, and WI, but an agent that is caught on lying cannot be trusted at all anymore.



Comparison of Different Categories of Dishonesties

	LIE	(I-)BS	WI	HT
statement	✓	✓		✓
addressee	✓	✓		✓
untruthful	✓			
intention	✓	(✓)	✓	✓
insincere	✓	✓		
valid sentence	✓	✓	✓	
contradictory sentence	✓	✓	✓	✓
contrary sentence	✓	✓	✓	✓
combination	✓		✓	
awareness	✓	✓	✓	✓
self-contradiction†	✓		✓	✓

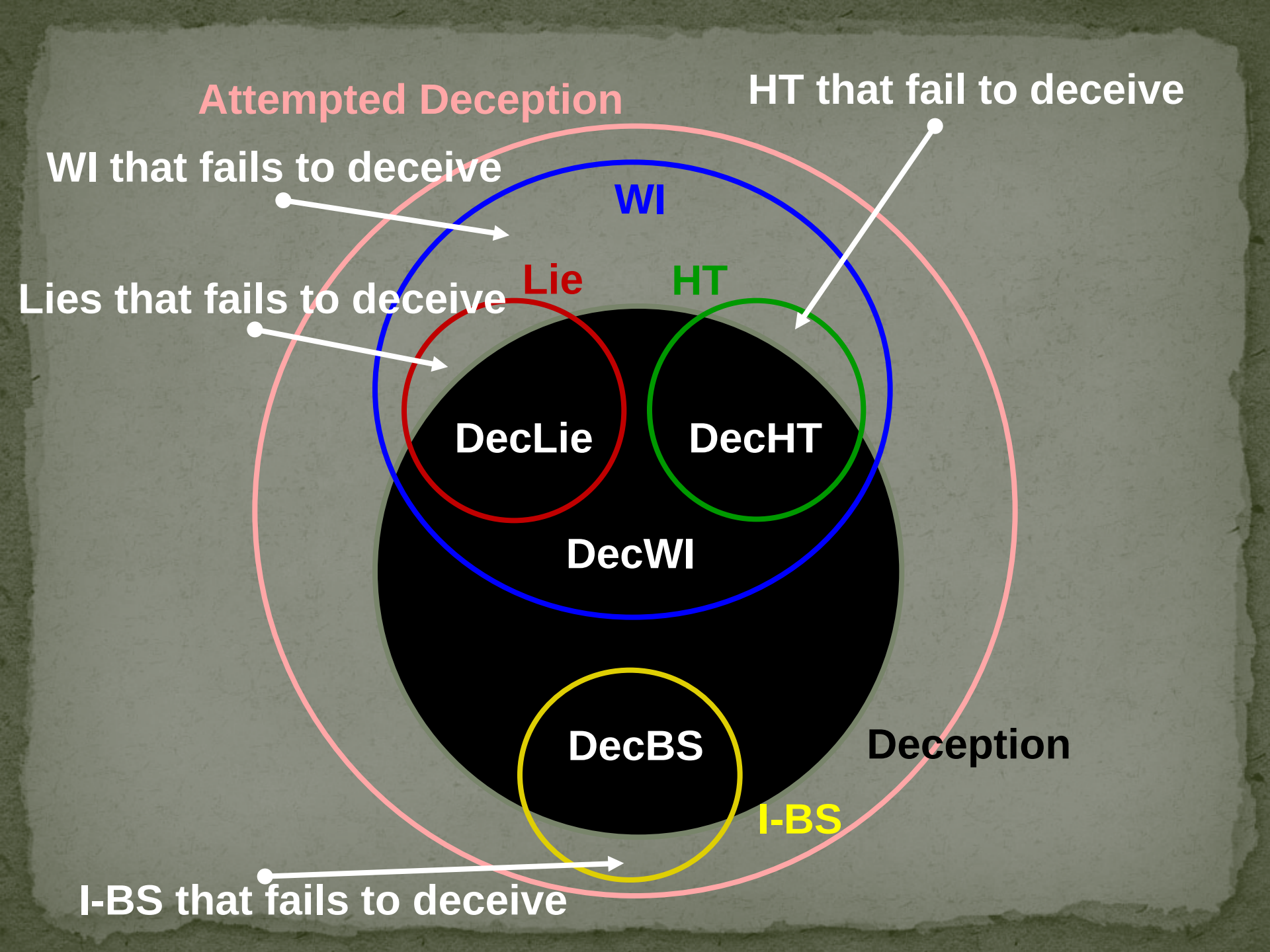
(†) The result holds with (Cons) except HT

Relationship to Deception

- Deception is an act whereby one person causes another person to have a false belief.
- In contrast, whether or not a speaker lies (BS, WI, or HT) depends only on the belief and intention of a speaker, and is independent of the effect of the action.
- Since deception involves a success or an achievement of the act, it is different from lies, BS, WI, and HT. In fact, deception is a perlocutionary act.

Deception

- To formulate deception, we introduce a causal relation "$p \Rightarrow q$" representing "q is a consequence of p" .
- "$p \Rightarrow q$" is true at a possible world iff q is true in every world selected in terms of the world in which p is true. (i.e., $p \Rightarrow q \supset (p \wedge q)$).
- With this extension, we could define different types of deception based on lies, I-BS, WI and HT as follows.
  - $DecLie_{ab}(\sigma) =^{def} LIE_{ab}(\sigma) \Rightarrow B_b\sigma$
  - $DecBS_{ab}(\sigma) =^{def} I\text{-}BS_{ab}(\sigma) \Rightarrow B_b\sigma$
  - $DecWI_{ab}(\sigma) =^{def} WI_{ab}(\sigma) \Rightarrow B_b\neg\sigma$
  - $DecHT_{ab}(\sigma,\delta) =^{def} HT_{ab}(\sigma,\delta) \Rightarrow B_b\delta$


Attempted Deception
HT that fail to deceive
WI that fails to deceive
WI
Lie
HT
Lies that fails to deceive
DecLie
DecHT
DecWI
DecBS
Deception
I-BS
I-BS that fails to deceive

Conclusion

- Formal models of dishonest communication are provided using a relatively simple logical formalization.
- Different categories of dishonesty are compared and formal properties are investigated.
- The dishonesty maxims can be seen as having a normative value, and should ideally be implemented for individual agents in multiagent systems.

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, pp. 286-299, 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 (TRUST), Taipei, Taiwan, May 2011.

- Chiaki Sakama
  Dishonest Reasoning by Abduction.
  IJCAI-2011, pp.1063-1068.
- Chiaki Sakama, Tran Son and Enrico Pontelli
  A Logical Formulation for Negotiation Among Dishonest Agents. IJCAI-2011, pp.1069-1074.
- 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
  Dishonest Arguments in Debate Games.
  4th International Conference on Computational Models of Argument (COMMA), Vienna, Austria, September 2012.
- Chiaki Sakama
  Learning Dishonesty.
  22nd International Conference on Inductive Logic Programming (ILP), Dubrovnik, Croatia, September 2012.

- Papers and slides are available at the author's home page:

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