A Formal Account of Deception

Chiaki Sakama

Wakayama University, Japan

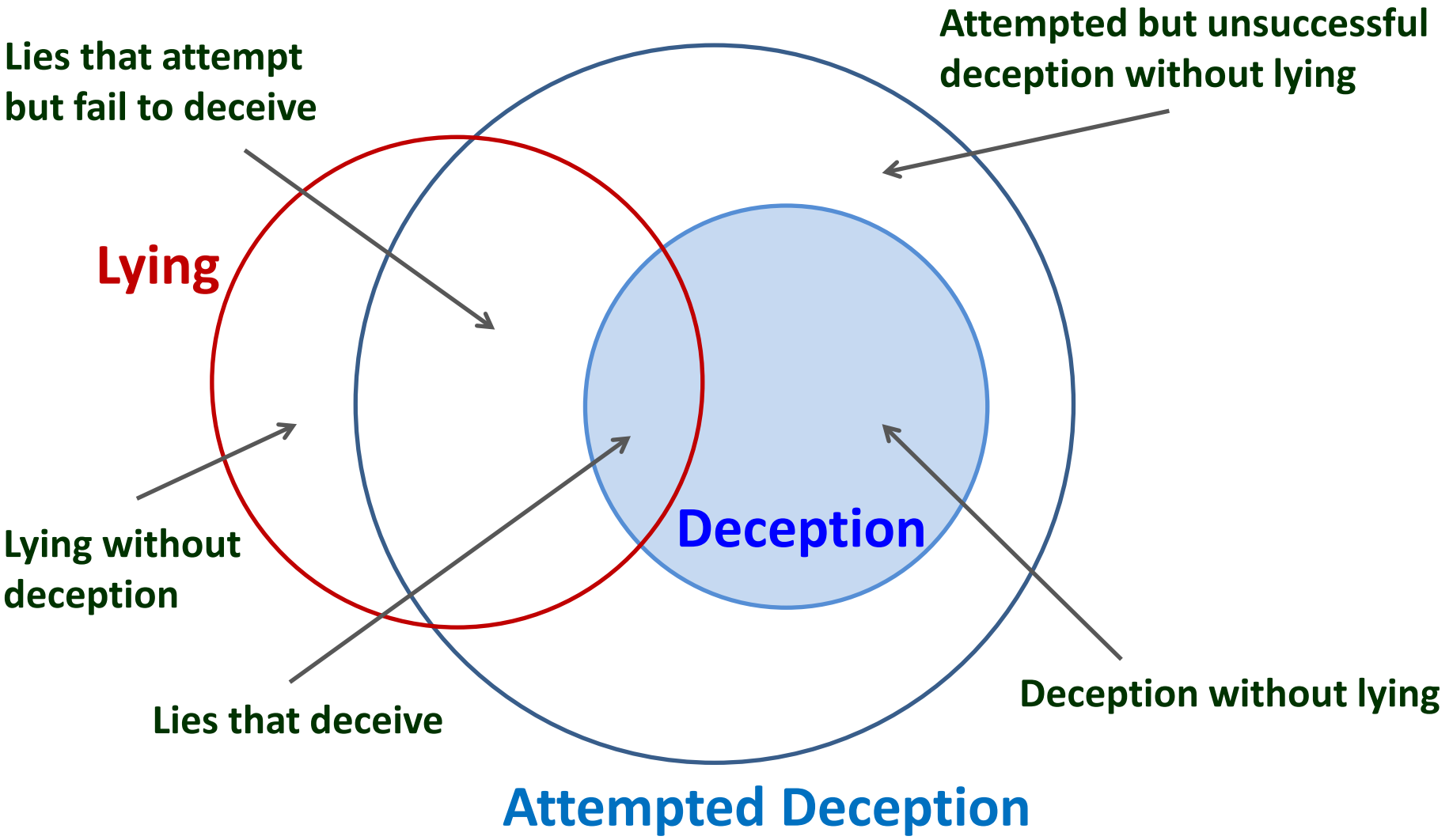
AAAI 2015 Symposium on Deceptive and Counter-Deceptive Machines
Nov. 12-14, 2015, Virginia, USA

What is Deception?

- To cause to believe what is false (OED 1989)
- Most Philosophers agree that
  - **Deception** implies a success of the act.
  - **Deception** involves intention, but not always.
  - **Deception** is a perlocutionary act that produces an effect in the belief state of an addressee.
  - **Self-deception** is an act of deceiving the self.
  - **Deception** is different from lying.

Lying, Deception, Attempted Deception

(T. L. Carson: "*Lying and Deception: Theory and Practice*", OUP, 2010)



Contributions

- Provide a formal account of deception based on **dynamic epistemic logic**.
- 3 different types of deception: **deception by lying**, **deception by bluffing**, and **deception by truth telling**, are formulated.
- 3 different aspects of deception: **intentional deception**, **indirect deception**, and **self-deception**, are analyzed.

Agent Announcement Logic

(H. van Ditmarsch: "Dynamics of Lying", *Synthese* 191, 2013)

- A **truthful agent** *a* believes that a sentence φ is true (represented as $\mathbf{B}_a \varphi$) when it announces φ.
- A **lying agent** *a* believes that a sentence φ is false (represented as $\mathbf{B}_a \neg\varphi$) when it announces φ.
- A **bluffing agent** *a* believes neither φ nor $\neg\varphi$ (represented as $\neg (\mathbf{B}_a \varphi \vee \mathbf{B}_a \neg\varphi)$) when it announces φ.
- A hearer believes a speaker believes an announcement.
- $[\mathbf{Truth}_a \varphi]\psi$: ψ is true after *a*'s **truthful** announcement of φ
- $[\mathbf{Lie}_a \varphi]\psi$: ψ is true after *a*'s **lying** announcement of φ
- $[\mathbf{Bluff}_a \varphi]\psi$: ψ is true after *a*'s **bluffing** announcement of φ

Deception by Lying (DBL)

- a: speaker, b: hearer; φ, ψ: propositional sentences
- $\mathbf{DBL}_{ab}(\varphi, \psi) := \mathbf{B}_a \neg\varphi$ (a believes $\neg\varphi$)
  $\wedge \neg\psi$ (ψ is false)
  $\wedge \mathbf{B}_b (\varphi \supset \psi)$ (b believes $\varphi \supset \psi$)
  $\wedge ([\mathbf{Lie}_a\, \varphi]\mathbf{B}_b\psi \vee ([\mathbf{Lie}_a\, \varphi]\neg\mathbf{B}_b\neg\psi)$
  (b believes the false sentence ψ after a's lying announcement φ, or
  b disbelieves the true sentence $\neg\psi$ after a's lying announcement φ)
- In particular,

  $\mathbf{DBL}_{ab}(\varphi, \varphi) := \mathbf{B}_a \neg\varphi \wedge \neg\varphi$
  $\wedge ([\mathbf{Lie}_a\, \varphi]\mathbf{B}_b\, \varphi \vee ([\mathbf{Lie}_a\, \varphi]\neg\mathbf{B}_b\neg\, \varphi)$

Example

- A salesperson lies that an investment promises a high return with little risk. The lie makes a customer believe that the investment is worth buying.
- The situation is represented by

$$\mathbf{DBL}_{ab}(\varphi, \psi) := \mathbf{B}_a \neg\varphi \wedge \neg\psi \wedge \mathbf{B}_b (\varphi \supset \psi)$$
$$\wedge ([\mathbf{Lie}_a\ \varphi]\mathbf{B}_b\psi \vee ([\mathbf{Lie}_a\ \varphi]\neg\mathbf{B}_b\neg\psi\ )$$

a: salesperson, *b*:customer,
φ =“an investment promises high return”
ψ =“the investment is worth buying”

Deception by Bluffing (DBB)

- a: speaker, b: hearer; φ, ψ: propositional sentences
- $\mathbf{DBB}_{ab}(\varphi, \psi) := \neg (\mathbf{B}_a\varphi \vee \mathbf{B}_a\neg\varphi)$

  (a believes neither φ nor $\neg\varphi$)

  $\wedge \neg\psi$ (ψ is false)

  $\wedge \mathbf{B}_b(\varphi \supset \psi)$ (b believes $\varphi \supset \psi$)

  $\wedge ([\mathbf{Bluff}_a\ \varphi]\mathbf{B}_b\psi \vee ([\mathbf{Bluff}_a\ \varphi]\neg\mathbf{B}_b\neg\psi$)

  (b believes the false sentence ψ after a's bluffing announcement φ, or

  b disbelieves the true sentence $\neg\psi$ after a's bluffing announcement φ)

Deception by Truth-Telling (DBT)

- a: speaker, b: hearer; φ, ψ: propositional sentences
- $\mathbf{DBT}_{ab}(\varphi, \psi) := \mathbf{B}_a\varphi$ (a believes φ)
  $\wedge \neg\psi$ (ψ is false)
  $\wedge \mathbf{B}_b(\varphi \supset \psi)$ (b believes $\varphi \supset \psi$)
  $\wedge ([\mathbf{Truth}_a\ \varphi]\mathbf{B}_b\psi \vee ([\mathbf{Truth}_a\ \varphi]\neg\mathbf{B}_b\neg\psi)$

  (b believes the false sentence ψ after a's truthful announcement φ, or
  b disbelieves the true sentence $\neg\psi$ after a's truthful announcement φ)

Example

- John invites Mary to dinner on the Christmas day. Mary says she has an appointment with another man. John understands Mary has a boyfriend, but Mary has an appointment with her father.
- The situation is represented by

$$\mathbf{DBT}_{ab}(\varphi, \psi) := \mathbf{B}_a\varphi \wedge \neg\psi \wedge \mathbf{B}_b(\varphi \supset \psi) \wedge ([\mathbf{Truth}_a\, \varphi]\mathbf{B}_b\psi \vee ([\mathbf{Truth}_a\, \varphi]\neg\mathbf{B}_b\neg\psi)$$

a: Mary, b: John

φ =“Mary has an appointment with another man”

ψ =“Mary has a boyfriend”

Properties

- **KD45 agent**: having consistent beliefs ($\neg B_a \bot$)
- **DBL**, **DBB**, and **DBT** fail if a KD45 hearer *b* already believes the falsity of an announcement:

  $\vdash B_b \neg\varphi \wedge \mathbf{DBX}_{ab}(\varphi, \varphi) \supset \bot$ where **X**=**L**, **B**, **T**
- **DBL**, **DBB**, and **DBT** fail if a KD45 hearer *b* already believes the falsity of an effect:

  $\vdash B_b \neg\psi \wedge \mathbf{DBX}_{ab}(\varphi, \psi) \supset \bot$

Properties

- **DBL**, **DBB**, and **DBT** fail if a KD45 hearer *b* believes that a KD45 speaker *a* is lying:

$$\vdash \mathrm{B}_b\mathrm{B}_a\neg p \wedge \mathbf{DBX}_{ab}(p, \psi) \supset \bot \quad (p\text{: proposition})$$

- **DBL**, **DBB** and **DBT** fail if a KD45 hearer *b* believes that a speaker *a* is bluffing:

$$\vdash \mathrm{B}_b(\neg\mathrm{B}_a p \wedge \neg\mathrm{B}_a\neg p) \wedge \mathbf{DBX}_{ab}(p, \psi) \supset \bot$$

- It is impossible to make DBB on one's belief:

$$\vdash \mathbf{DBB}_{ab}(\mathrm{B}_a\varphi, \psi) \vee \mathbf{DBB}_{ab}(\neg\mathrm{B}_a\varphi, \psi) \supset \bot$$

Deception by Omission

- A person sells a used car that has some problem in its engine. If (s)he sells the car without informing a customer of the problem, it is **deception by omission**.
- Deception by omission is characterized as **DBT** with announcing a valid sentence (i.e., no informative announcement):

$$\mathbf{DBT}_{ab}(\mathsf{T}, \psi) \equiv \neg\psi \ \wedge \mathbf{B}_b \psi$$

Intention

- Deception involving a speaker's intent to deceive a hearer is distinguished as **intentional deception**.
- a: speaker, b: hearer; φ, ψ: propositional sentences
- $\mathbf{DBL}_{ab}(\varphi, \psi) := \mathbf{B}_a \neg\varphi \wedge \neg\psi \wedge \mathbf{B}_b (\varphi \supset \psi)$
  $\wedge ([\mathbf{Lie}_a \varphi]\mathbf{B}_b\psi \vee ([\mathbf{Lie}_a \varphi]\neg\mathbf{B}_b\neg\psi)$
- $\mathbf{I\text{-}DBL}_{ab}(\varphi, \psi) := \mathbf{B}_a \neg\psi \wedge \mathbf{B}_a\mathbf{B}_b (\varphi \supset \psi) \wedge \mathbf{DBL}_{ab}(\varphi, \psi)$
- $\mathbf{I\text{-}DBB}_{ab}(\varphi, \psi) := \mathbf{B}_a \neg\psi \wedge \mathbf{B}_a\mathbf{B}_b (\varphi \supset \psi) \wedge \mathbf{DBB}_{ab}(\varphi, \psi)$
- $\mathbf{I\text{-}DBT}_{ab}(\varphi, \psi) := \mathbf{B}_a \neg\psi \wedge \mathbf{B}_a\mathbf{B}_b (\varphi \supset \psi) \wedge \mathbf{DBT}_{ab}(\varphi, \psi)$
- A speaker makes an announcement φ expecting that it will cause the hearer's believing a false sentence ψ.

Properties

- If a speaker deceives a hearer while believing his/her deceptive act, then it is intentional deception:

  $\vdash \mathbf{DBX}_{ab}(\varphi, \psi) \wedge \mathrm{B}_a(\mathbf{DBX}_{ab}(\varphi, \psi)) \supset \mathbf{I\text{-}DBX}_{ab}(\varphi, \psi)$

  where **X=L**, **B**, **T**

- By putting $\varphi \equiv \psi$,

  $\mathbf{I\text{-}DBL}_{ab}(\varphi, \psi) := \mathbf{B}_a \neg\psi \wedge \mathbf{B}_a\mathbf{B}_b (\varphi \supset \psi) \wedge \mathbf{DBL}_{ab}(\varphi, \psi)$

  becomes $\mathbf{I\text{-}DBL}_{ab}(\varphi, \varphi) \equiv \mathbf{DBL}_{ab}(\varphi, \varphi)$

- A liar always intends to deceive a hearer wrt the sentence being announced.

Properties

- By contrast,
  $\vdash$ **I-DBB**$_{ab}(\varphi, \varphi) \supset \bot$
  $\vdash$ **I-DBT**$_{ab}(\varphi, \varphi) \supset \bot$ for any KD45 speaker *a*
- **DBB** or **DBT** can be intentional only if a hearer comes to believe a false sentence that is different from the sentence announced by a speaker.
- **I-DBB** or **I-DBT** requires advanced techniques as a speaker has to select an announcement that is different from the false fact which the speaker wants a hearer to believe.
- Very young children do not have advanced technique of deception, then most deception by them is of the type **(I-)DBL**$_{ab}(\varphi, \varphi)$ that is the most simple form of deception.

Indirect Deception

- John visits a clinic for a medical check. He is diagnosed as having a serious cancer. A doctor does not inform the patient of this fact. John has no symptom giving him any reason to believe the fact. He tells his wife that the result of a medical test is normal.
- The situation is represented by

$$\mathbf{DBL}_{ab}(\varphi, \varphi) \wedge \mathbf{DBT}_{bc}(\varphi, \varphi)$$

  a: doctor, b: John, c: wife, φ =“normal”
- In this case, a doctor **indirectly deceives** John’s wife by lying.

Indirect Deception

- *a*, *b*, *c*: agent; φ, ψ , λ: propositional sentences

  IN-DBL$_{ab}(\varphi, \lambda)$:= (I-)**DBL**$_{ab}(\varphi, \psi) \wedge$ **DBT**$_{bc}(\psi, \lambda)$

  IN-DBB$_{ab}(\varphi, \lambda)$:= (I-)**DBB**$_{ab}(\varphi, \psi) \wedge$ **DBT**$_{bc}(\psi, \lambda)$

  IN-DBT$_{ab}(\varphi, \lambda)$:= (I-)**DBT**$_{ab}(\varphi, \psi) \wedge$ **DBT**$_{bc}(\psi, \lambda)$
- An agent *a* may have intention to deceive *b*, while an agent *b* does not have intention to deceive *c*.
- An indirect deception could be chained like:
  (I-)**DBX**$_{ab}(\varphi, \psi_1) \wedge$ **DBT**$_{bc}(\psi_1, \psi_2) \wedge$ **DBT**$_{cd}(\psi_2, \psi_3) \wedge$...
  where **X=L**, **B**, **T.**

Self-Deception

- Self-deception in **DBL** involves a mental state of an agent who has contradictory belief wrt a false fact:

  $\vdash \mathbf{DBL}_{aa}(p, p) \equiv \mathbf{B}_a(p \wedge \neg p) \wedge \neg p$ (p: proposition)

- $\mathbf{DBL}_{aa}(p, p)$ is impossible for KD45 agents.
- By contrast, $\mathbf{DBL}_{aa}(\varphi, \psi)$ implies

  $\mathbf{B}_a \neg\varphi \wedge \neg\psi \wedge \mathbf{B}_a(\varphi \supset \psi) \wedge (\mathbf{B}_a\psi \vee \neg\mathbf{B}_a\neg\psi)$

  which is consistent for a sentence $\varphi \not\equiv \psi$

- A speaker *a* believes the falsity of φ while believes the effect ψ or its possibility that would be obtained if φ were the case. (**counterfactual inference**)

Properties

- Self-deception by **DBB** or **DBT** is possible on an announced sentence even by a KD45 agent:

  $\vdash \mathbf{DBB}_{aa}(p, p) \supset \neg(\mathrm{B}_a p \vee \mathrm{B}_a \neg p) \wedge \neg p$

  $\vdash \mathbf{DBT}_{aa}(p, p) \supset \mathrm{B}_a p \wedge \neg p$

- Any KD45 agent cannot intentionally deceive oneself by **DBL**, **DBB** or **DBT**:

  $\vdash \mathbf{I\text{-}DBX}_{aa}(\varphi, \psi) \supset \bot$ where **X=L, B, T**

- Self-deception is possible by agents with consistent belief only when it is done unconsciously.

``*… self-deception occurs when the conscious mind is kept in dark*" (R. Trivers: ``*The Folly of Fools*", 2011).

Indirect Self-Deception

- There is a meeting 3 months ahead, say, on March 31. Mary is unwilling to attend it and she deliberately records the wrong date, April 1st, on her online assistant. Mary is very busy and has completely forgotten the actual date. On April 1st, her online assistant informs her of the meeting ...
- The situation is represented by

  $\mathbf{IN\text{-}DBL}_{aa}(\varphi, \varphi) = \mathbf{I\text{-}DBL}_{ab}(\varphi, \varphi) \wedge \mathbf{DBT}_{ba}(\varphi, \varphi)$

  a: Mary, b: online-assistant,

  φ =“Meeting on April 1st”

Indirect Self-Deception

- Recall that self-deception on a lying sentence is impossible for KD45 agents:
  $\vdash \mathbf{DBL}_{aa}(p, p) \equiv \mathbf{B}_a(p \wedge \neg p) \wedge \neg p$
- Interestingly, however, KD45 agents can deceive oneself by using **indirect DBL** even on a lying sentence.
  $\mathbf{IN\text{-}DBL}_{aa}(\varphi, \varphi) = \mathbf{I\text{-}DBL}_{ab}(\varphi, \varphi) \wedge \mathbf{DBT}_{ba}(\varphi, \varphi)$
- Generally, indirect self-deception is represented by
  $\mathbf{IN\text{-}DBX}_{aa}(\varphi, \lambda) = (\mathbf{I\text{-}})\mathbf{DBX}_{ab}(\varphi, \psi) \wedge \mathbf{DBT}_{ba}(\psi, \lambda)$
  where **X=L**, **B**, **T**
- As such, self-deception does not always involve contradiction if it is done indirectly.