Deception in Epistemic Causal Logic

Chiaki Sakama

Wakayama University Japan

August 19, 2021

What is deception?

To cause to believe what is false (OED 1989)

What kind of logic is needed for formal account of deception?

To cause to believe what is false (OED 1989)

- a logic needs to express causality between a deceptive act and its effect.

What kind of logic is needed for formal account of deception?

To cause to believe what is false (OED 1989)

- a logic needs to express causality between a deceptive act and its effect.
- a logic needs to express belief states of agents.

What kind of logic is needed for formal account of deception?

To cause to believe what is false (OED 1989)

- a logic needs to express causality between a deceptive act and its effect.
- a logic needs to express belief states of agents.

For this purpose, we use

epistemic causal logic

$=$ causal logic [Giunchiglia, et al., AIJ 2004]

$+$ belief modality

Epistemic Causal Logic: Definitions

causal rule

A causal rule is of the form:

$$\phi \Rightarrow \psi \qquad (\phi, \psi : \text{propositional formula})$$

meaning "ψ is caused if ϕ is true."
A (causal) theory is a finite set of causal rules.

model

Given a theory T and an interpretation I, define

$$T^I = \{\, \psi \mid (\phi \Rightarrow \psi) \in T \text{ for some } \phi \text{ and } I \models \phi \,\}.$$

I is a model of T if I is the unique model of T^I.
If every model of T satisfies a formula F, written $T \models F$.
If T has no model, written $T \models \bot$.

Epistemic Causal Logic: Axioms

$U^t_{ab}\phi$: an agent a utters a sentence ϕ to an agent b at time t.
$B^t_a\phi$: an agent a believes a sentence ϕ at time t

Axioms for utterance and beliefs

(axioms of utterance): $U^t_{ab}\phi \Rightarrow U^t_{ab}\phi$ and $\neg U^t_{ab}\phi \Rightarrow \neg U^t_{ab}\phi$.

(axioms of belief): $B^t_a\phi \Rightarrow B^t_a\phi$ and $\neg B^t_a\phi \Rightarrow \neg B^t_a\phi$.

$B^t_a\phi \equiv B^t_a\psi$ if $\phi \equiv \psi$.

$B^t_a(\phi \wedge \psi) \equiv B^t_a\phi \wedge B^t_a\psi$.

(axioms of inertia): $B^t_a\phi \wedge B^{t+1}_a\phi \Rightarrow B^{t+1}_a\phi$

$\neg B^t_a\phi \wedge \neg B^{t+1}_a\phi \Rightarrow \neg B^{t+1}_a\phi$.

(axiom of truth): $B^t_a\top$ for any t.

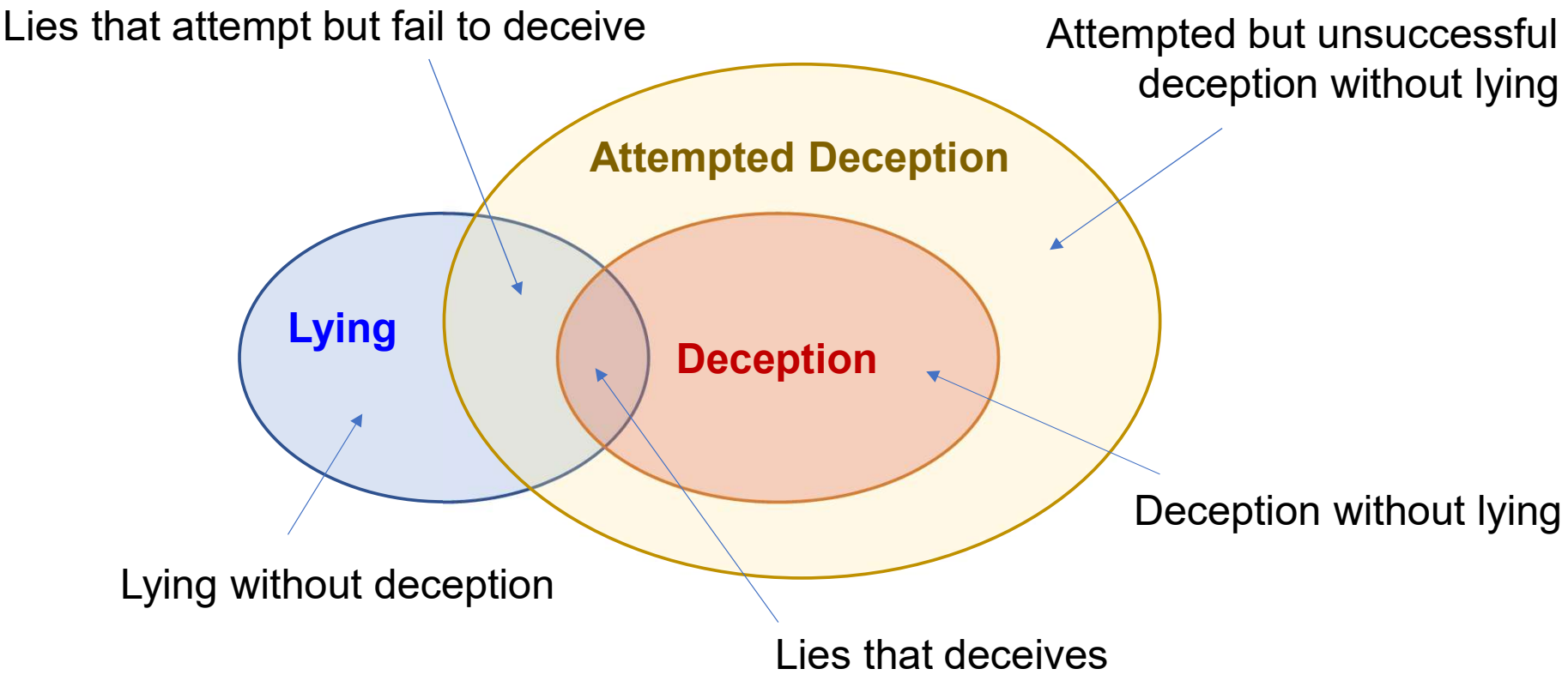
(axiom of rationality): $\neg B^t_a\bot$ for any t if a is *rational*.

(axiom of credibility): $U^t_{ab}\phi \Rightarrow B^{t+1}_b\phi$ if b is *credulous*.

(axiom of reflection): $U^t_{ab}\phi \Rightarrow B^{t+1}_b B^t_a\phi$ if b is *reflective*.

Lying, deception and attempted deception

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



Deception by Lying

Lying (a: speaker, b: hearer, ϕ: sentence)

$$\mathrm{LIE}_{ab}^{t}(\phi) \stackrel{def}{=} B_a^t \neg\phi \wedge U_{ab}^t \phi$$

(a lies to b if a utters a believed-false sentence ϕ to b at t)

Deception by Lying

$$\mathrm{DBL}_{ab}^{t+1}(\phi) \stackrel{def}{=} \neg\phi \wedge (\,\mathrm{LIE}_{ab}^{t}(\phi) \Rightarrow B_b^{t+1}\phi\,).$$

(a lies to b at t on a false sentence ϕ, which causes b's believing ϕ at the next time step $t+1$)

Note

- In lying, a speaker a believes ϕ but the actual falsity of ϕ is not required.
- In DBL, the actual falsity of ϕ is required.

Properties of DBL (1)

DBL does not happen if a sentence ϕ is true

$$\phi \wedge \mathrm{DBL}_{ab}^{t+1}(\phi) \models \bot$$

Lying on a false sentence succeeds to deceive if a hearer is credulous, i.e., $U_{ab}^{t}\phi \Rightarrow B_{b}^{t+1}\phi$

$$\neg\phi \wedge \mathrm{LIE}_{ab}^{t}(\phi) \models B_{b}^{t+1}\phi$$

DBL on the valid sentence always fails

$$\mathrm{DBL}_{ab}^{t+1}(\top) \models \bot$$

DBL on the contradictory sentence fails if a hearer is rational, i.e., $\neg B_{b}^{t}\bot$

$$\mathrm{LIE}_{ab}^{t}(\bot) \wedge \mathrm{DBL}_{ab}^{t+1}(\bot) \models \bot \quad \text{if } b \text{ is rational.}$$

Properties of DBL (2)

DBL fails if a rational hearer believes the contrary

$B_b^t \neg\phi \wedge \mathrm{LIE}_{ab}^t(\phi) \wedge \mathrm{DBL}_{ab}^{t+1}(\phi) \models \bot$ if b is rational.

If a rational hearer is credulous, DBL succeeds even if the hearer believes the contrary

$\neg\phi \wedge B_b^t \neg\phi \wedge \mathrm{LIE}_{ab}^t(\phi) \models B_b^{t+1}\phi$
if b is credulous and rational.

A hearer b does not believe that a speaker a is lying if b is rational, reflective (i.e., $U_{ab}^t\phi \Rightarrow B_b^{t+1}B_a^t\phi$) and believes that a is also rational.

$\mathrm{LIE}_{ab}^t(\phi) \wedge B_b^{t+1}(\mathrm{LIE}_{ab}^t(\phi)) \models \bot$ if b is rational, reflective, and believes that a speaker a is rational.

Deception without Lying

Deception by Bluffing

$$\mathrm{DBB}^{t+1}_{ab}(\phi) \stackrel{def}{=} \neg\phi \wedge (\,\mathrm{BLUF}^{t}_{ab}(\phi) \Rightarrow B^{t+1}_{b}\phi\,)$$

where $\mathrm{BLUF}^{t}_{ab}(\phi) \stackrel{def}{=} \neg B^{t}_{a}\phi \wedge \neg B^{t}_{a}\neg\phi \wedge U^{t}_{ab}\phi$

Deception by Truthful Telling

$$\mathrm{DBT}^{t+1}_{ab}(\phi) \stackrel{def}{=} \neg\phi \wedge (\,\mathrm{TRT}^{t}_{ab}(\phi) \Rightarrow B^{t+1}_{b}\phi\,)$$

where $\mathrm{TRT}^{t}_{ab}(\phi) \stackrel{def}{=} B^{t}_{a}\phi \wedge U^{t}_{ab}\phi$

Deception by Omission (or withholding information)

$$\mathrm{DBO}^{t+1}_{ab}(\phi) \stackrel{def}{=} \phi \wedge (\,\mathrm{WI}^{t}_{ab}(\phi) \Rightarrow \neg B^{t+1}_{b}\phi\,)$$

where $\mathrm{WI}^{t}_{ab}(\phi) \stackrel{def}{=} B^{t}_{a}\phi \wedge \neg U^{t}_{ab}\phi$

Intentional Deception

Intentional DBL and DBB

- $\text{I-DBL}_{ab}^{t+1}(\phi) \stackrel{def}{=} \neg\phi \wedge (\,\text{LIE}_{ab}^{t}(\phi) \wedge B_a^t B_b^{t+1}\phi \Rightarrow B_b^{t+1}\phi\,)$
- $\text{I-DBB}_{ab}^{t+1}(\phi) \stackrel{def}{=} \neg\phi \wedge (\,\text{BLUF}_{ab}^{t}(\phi) \wedge B_a^t B_b^{t+1}\phi \Rightarrow B_b^{t+1}\phi\,)$

By $B_a^t B_b^{t+1}\phi$, a speaker a believes that a hearer b will believe the false sentence ϕ in the next time step.

Intentional Deception

Intentional DBL and DBB

- $\text{I-DBL}_{ab}^{t+1}(\phi) \stackrel{def}{=} \neg\phi \wedge (\,\text{LIE}_{ab}^{t}(\phi) \wedge B_a^t B_b^{t+1}\phi \Rightarrow B_b^{t+1}\phi\,)$
- $\text{I-DBB}_{ab}^{t+1}(\phi) \stackrel{def}{=} \neg\phi \wedge (\,\text{BLUF}_{ab}^{t}(\phi) \wedge B_a^t B_b^{t+1}\phi \Rightarrow B_b^{t+1}\phi\,)$

Intentional DBT

$$\text{I-DBT}_{ab}^{t+1}(\phi,\psi) \stackrel{def}{=} \neg\psi \wedge (\,\text{TRT}_{ab}^{t}(\phi) \wedge B_a^t(B_b^{t+1}\phi \supset B_b^{t+1}\psi) \wedge B_a^t\neg\psi \Rightarrow B_b^{t+1}\psi\,)$$

A speaker a truthfully tells ϕ while a believes that a hearer b's believing ϕ leads to b's believing another false sentence ψ in the next time step.

Intentional Deception

Intentional DBL and DBB

- $\text{I-DBL}_{ab}^{t+1}(\phi) \stackrel{def}{=} \neg\phi \wedge (\,\text{LIE}_{ab}^{t}(\phi) \wedge B_a^t B_b^{t+1}\phi \Rightarrow B_b^{t+1}\phi\,)$
- $\text{I-DBB}_{ab}^{t+1}(\phi) \stackrel{def}{=} \neg\phi \wedge (\,\text{BLUF}_{ab}^{t}(\phi) \wedge B_a^t B_b^{t+1}\phi \Rightarrow B_b^{t+1}\phi\,)$

Intentional DBT

$$\text{I-DBT}_{ab}^{t+1}(\phi,\psi) \stackrel{def}{=} \neg\psi \wedge (\,\text{TRT}_{ab}^{t}(\phi) \wedge B_a^t(B_b^{t+1}\phi \supset B_b^{t+1}\psi) \wedge B_a^t\neg\psi \Rightarrow B_b^{t+1}\psi\,)$$

Intentional DBO

$$\text{I-DBO}_{ab}^{t+1}(\phi) \stackrel{def}{=} \phi \wedge (\,\text{WI}_{ab}^{t}(\phi) \wedge B_a^t\neg B_b^t\phi \Rightarrow \neg B_b^{t+1}\phi\,)$$

A speaker a withholds ϕ while believing b's ignorance of ϕ, which causes b's disbelieving ϕ (or prevents b from believing ϕ) in the next time step.

Indirect Deception

- $\text{IN-DBL}_{ac}(\phi) \stackrel{def}{=} \text{(I-)DBL}_{ab}^{t+1}(\phi) \wedge \text{DBT}_{bc}^{t+2}(\phi)$
  (a's lying on ϕ results in b's believing a false sentence ϕ, and then b's truthful telling on ϕ results in c's believing ϕ)
- $\text{IN-DBB}_{ac}(\phi) \stackrel{def}{=} \text{(I-)DBB}_{ab}^{t+1}(\phi) \wedge \text{DBT}_{bc}^{t+2}(\phi)$
- $\text{IN-DBT}_{ac}(\phi) \stackrel{def}{=} \text{DBT}_{ab}^{t+1}(\phi) \wedge \text{DBT}_{bc}^{t+2}(\phi)$
- $\text{IN-DBO}_{ac}(\phi) \stackrel{def}{=} \text{(I-)DBO}_{ab}^{t+1}(\phi) \wedge \neg U_{bc}^{t+1}\phi \Rightarrow \neg B_{c}^{t+2}\phi$
  (a's withholding ϕ results in b's disbelieving a true sentence ϕ. Then b does not inform c of ϕ, which results in c's disbelieving ϕ)
- $\text{IN-I-DBT}_{ac}(\phi, \psi) \stackrel{def}{=} \text{I-DBT}_{ab}^{t+1}(\phi, \psi) \wedge \text{DBT}_{bc}^{t+2}(\psi)$

Self-Deception

Self-deception by lying produces contradictory belief

$$\mathrm{LIE}^t_{aa}(\phi) \wedge \text{(I-)}\mathrm{DBL}^{t+1}_{aa}(\phi) \models B^{t+1}_a \bot$$

$B^t_a \neg\phi$ in $\mathrm{LIE}^t_{aa}(\phi)$ implies $B^{t+1}_a \neg\phi$ by the axioms of inertia. $\mathrm{DBL}^{t+1}_{aa}(\phi)$ implies $B^{t+1}_a \phi$. Then, $B^{t+1}_a \neg\phi \wedge B^{t+1}_a \phi \equiv B^{t+1}_a \bot$

If a rational agent is credulous, self-DBL does not involve contradictory belief

$$\mathrm{LIE}^t_{aa}(\phi) \wedge \text{(I-)}\mathrm{DBL}^{t+1}_{aa}(\phi) \not\models B^{t+1}_a \bot$$

if a is credulous and rational

A credulous agent revises its belief from $B^t_a \neg\phi$ to $B^{t+1}_a \phi$. $B^{t+1}_a \phi$ implies $\neg B^{t+1}_a \neg\phi$ by the axiom of rationality. Then the axioms of inertia do not produce $B^{t+1}_a \neg\phi$ from $B^t_a \neg\phi$.

Self-Deception

Self-deception by bluffing does not produce contradictory belief

$$\mathrm{BLUF}^{t}_{aa}(\phi) \wedge \text{(I-)}\mathrm{DBB}^{t+1}_{aa}(\phi) \models \neg B^{t}_{a}\phi \wedge B^{t+1}_{a}\phi$$

Self-deception by omission does not produce contradictory belief

$$\mathrm{WI}^{t}_{aa}(\phi) \wedge \text{(I-)}\mathrm{DBO}^{t+1}_{aa}(\phi) \models B^{t}_{a}\phi \wedge \neg B^{t+1}_{a}\phi$$

(a person, who believes something true but does not refer to it, will *forget* it.)

Self-deception by truthful-telling does not contradict while inconsistency arises if accompanied by intention

- $\mathrm{TRT}^{t}_{aa}(\phi) \wedge \mathrm{DBT}^{t+1}_{aa}(\phi) \not\models B^{t+1}_{a}\bot$
- $\mathrm{TRT}^{t}_{aa}(\phi) \wedge \text{I-}\mathrm{DBT}^{t+1}_{aa}(\phi, \psi) \models B^{t+1}_{a}\bot$

Summary

- Different types of deception are formulated using epistemic causal logic.
- From the computational perspective, a causal theory handled in this study is translated into a logic program under the answer set semantics.
- The current framework is extended to handle more complicated cases by taking *a theory of mind* into consideration.