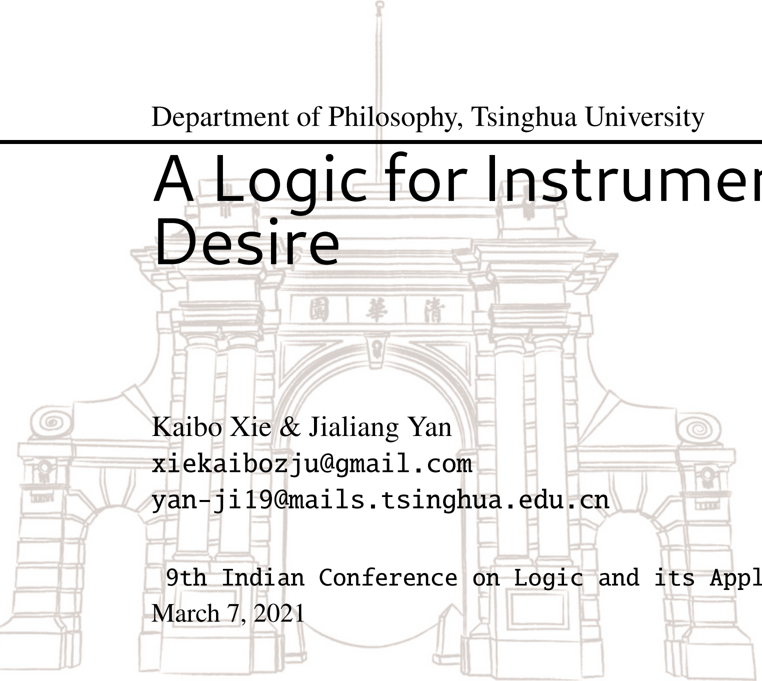

Department of Philosophy, Tsinghua University

A Logic for Instrumental Desire

Kaibo Xie & Jialiang Yan
xiekaibozju@gmail.com
yan-ji19@mails.tsinghua.edu.cn

9th Indian Conference on Logic and its Applications
March 7, 2021

Outline

Introduction

Desire is the primitive component that drives the agent's behavior. Theories of desire have been established in many disciplines. We study the reasoning about desire in order to predict and characterize tendency of the agent's action.

BELLY RUBS
I THINK I WANT BELLY RUBS
HEYBUDDYCOMICS.COM
#070
HEY BUDDY COMICS

Instrumental Desire

In this paper, we give an interpretation of desire from a perspective of instrumentalist, and then propose a logic.

Having a desire, as we explain, if and only if this desire can bring about what the agent most prefers at the moment.
There are two main feature capturing instrumental desire:

- What people desire is not for its own sake.
- Causality should be taken into account .

Example-1

Robin was in charge of an important project in the company recently, and the tremendous pressure made him sleepless every night. Robin is a healthy - conscious person and poor sleep will make him sluggish the next day which bothers Robin most. So having a good sleep is what he needs most at the moment. Therefore, on this sleepless night, Robin wants to take some sleep-helping pills to help him sleep well, even if these pills have some side effects.

Example-1

- Robin wants to take sleeping pills but he does not prefer to doing it. Because he cherished his health, he hated the side effects of pills.

- Robin's priority right now is to get a good night's sleep, and that's exactly what sleeping pills will bring him.

Example-2

> *Robin's company offered him a chance to study in the Netherlands and he attached great importance to it. But his savings are small, so he doesn't want to pay for the tuition. Unfortunately, due to the epidemic, he cannot go abroad. Even after paying the tuition, he can only stay at home and take online lessons.*

(The original version of this example comes from Maria Aloni.)

Example-2

- The similarity approach is not enough to account for this case.
  - w_1: study in the Netherlands; pay for the tuition
  - w_2: not study in the Netherlands; pay for the tuition.
  - If w_1 is closer than w_2 to the actual world, then we will say Robin wants to pay for the tuition which is unreasonable.

- We have to refute that w_1 is more similar than w_2 compared with the actual world.

- Causality is needed here. Robin's paying the tuition cannot bring about his studying in the Netherlands.

The Basic Idea of Modelling

To capture these two features of instrumental desire, we introduce a **desire-causality model** in our paper.

- Following the traditional approach of formalizing preference, e.g. [von Wright, 1972] and [van Benthem & Liu, 2007], we use a total order over all the possible worlds to represent the preference structure of an agent.
- We adopt a causal model which makes use of the interventionist approach to causality from [Halpern, 2000] and [Pearl, 2002].

Causal Modelling

Halpern (2000) and Pearl (2002)

Causal Variables

- $\mathcal{U} = \{U_1, \ldots, U_m\}$ is a set of *exogenous* variables,
- $\mathcal{V} = \{V_1, \ldots, V_n\}$ is a set of *endogenous* variables
- $\mathcal{R}(X)$ is the non-empty range of the variable $X \in \mathcal{U} \cup \mathcal{V}$.

Combining causality and desire

Desire-Causality Model

Let $\mathcal{S} = \langle \mathcal{U}, \mathcal{V}, \mathcal{R} \rangle$. A desire-causality model for $\mathcal{S}$ is a tuple $\langle \mathcal{F}, \mathcal{A}, < \rangle$.

- **Structural Functions** $\mathcal{F} = \{\mathcal{F}_{V_j} \mid V_j \in \mathcal{V}\}$. For each endogenous variable V_j, $\mathcal{F}_{V_j}$ is a mapping from all assignments to $\mathcal{U} \cup \mathcal{V} \setminus \{V_j\}$ to $\mathcal{R}(V_j)$. $\mathcal{F}$ is assumed to be recursive.
- **Valuation Function** $\mathcal{A}$ assigns to every $X \in \mathcal{U} \cup \mathcal{V}$ a value $\mathcal{A}(X) \in \mathcal{R}(X)$. $\mathcal{A}$ has to *comply with* $\mathcal{F}_{V_j}$.
- **Preference Ordering** $<$ is a total order over all possible assignments to $\mathcal{U} \cup \mathcal{V}$.

Language for instrumental desire

Language of Logic ID

Formulas ϕ of the language $\mathcal{L}$ based on $\mathcal{S}$

$$\phi ::= X{=}x \mid \neg\phi \mid \phi \wedge \phi \mid FP(\vec{X} = \vec{x}) \mid D(\vec{X}{=}\vec{x}) \mid (\vec{X}{=}\vec{x}) \mathrel{\Box\!\!\rightarrow} \phi$$

Adopting interventionist approach

The semantics of “desire $X = x$”: after an intervention forcing $X = x$ to be true, the world results from it will be more preferred than the current world.

Intervention on desire-causality models

Let $M = \langle \mathcal{F}, \mathcal{A}, < \rangle$ be a DC-model based on $\mathcal{S}$.
$M_{\vec{X}=\vec{x}} = \langle \mathcal{F}_{\vec{X}=\vec{x}}.\mathcal{A}^{\mathcal{F}}_{\vec{X}=\vec{x}}, < \rangle$ is the DC model resulting from an intervention setting the values of variables in $X_1, \ldots, X_n$ to $x_1, \ldots, x_n$:

- $\mathcal{F}_{\vec{X}=\vec{x}}$ is as $\mathcal{F}$ except that, for each endogenous variable X_i in $\vec{X}$, the function f_{X_i} is replaced by a *constant* function f'_{X_i} that returns the value x_i regardless of the values of all other variables.
- $\mathcal{A}^{\mathcal{F}}_{\vec{X}=\vec{x}}$ is the updated solution to the updated structural equations.

Truth Conditions

Let $\langle \mathcal{F}, \mathcal{A}, < \rangle$ be a DC-model based on $\mathcal{S}$

Truth condition of formulas in $\mathcal{L}$

- $\langle \mathcal{F}, \mathcal{A}, < \rangle \models X{=}x$ iff $\mathcal{A}(X) = x$
- $\langle \mathcal{F}, \mathcal{A}, < \rangle \models FP(\vec{X} = \vec{x})$ iff for any two assignments $\mathcal{A}_1$ and $\mathcal{A}_2$ to $\mathcal{U} \cup \mathcal{V}$ such that $\mathcal{A}_1(\vec{X}) = \vec{x}$ and $\mathcal{A}_2(\vec{X})$ is not $\vec{x}$, $\mathcal{A}_2 < \mathcal{A}_1$
- $\langle \mathcal{F}, \mathcal{A}, < \rangle \models (\vec{X}{=}\vec{x}) \Box\!\!\rightarrow \phi$ iff $\langle \mathcal{F}_{\vec{X}=\vec{x}}, \mathcal{A}^{\mathcal{F}}_{\vec{X}=\vec{x}}, < \rangle \models \phi$
- $\langle \mathcal{F}, \mathcal{A}, < \rangle \models D(\vec{X}{=}\vec{x})$ iff $\mathcal{A}^{\mathcal{F}} < \mathcal{A}^{\mathcal{F}}_{\vec{X}=\vec{x}}$

Two examples revisited

X stands for taking sleep-helping pills. Y stands for having a good sleep.
X is an exogenous variable.
$\mathcal{F}_Y$: the value of Y is equal to the value of X

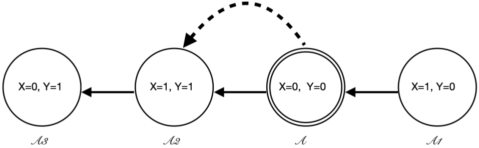

Figure: Robin's desire of pills

- $\mathcal{A}(X) = 0, \mathcal{A}(Y) = 0$
- According to $\mathcal{F}_{X=1}$, $\mathcal{A}_{X=1}(X) = 1$, $\mathcal{A}_{X=1}(Y) = 1$

Two examples revisited

X stands for taking sleep-helping pills. Y stands for having a good sleep.
X is an exogenous variable.
$\mathcal{F}_Y$: the value of Y is equal to the value of X

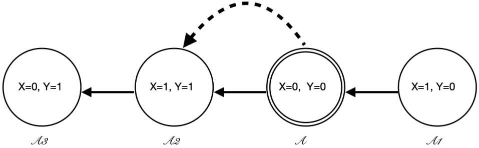

Figure: Robin's desire of pills

- $\mathcal{A}(X) = 0, \mathcal{A}(Y) = 0$
- According to $\mathcal{F}_{X=1}$, $\mathcal{A}_{X=1}(X) = 1, \mathcal{A}_{X=1}(Y) = 1$
- $\mathcal{A} < \mathcal{A}_{X=1}$, therefore $\langle \mathcal{F}, \mathcal{A}, < \rangle \models D(X = 1)$

Two examples revisited

S stands for studying in Netherlands; T stands for paying tuition fee. S is an exogenous variable.

$\mathcal{F}_T$: the value of T is equal to the value of S

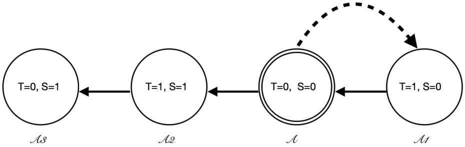

Figure: Robin's desire of tuition

- $\mathcal{A}(S) = 0, \mathcal{A}(T) = 0;$
- According to $\mathcal{F}_{T=1}$, $\mathcal{A}_{T=1}(T) = 1$, $\mathcal{A}_{T=1}(S) = 0$

Two examples revisited

S stands for studying in Netherlands; T stands for paying tuition fee. S is an exogenous variable.

$\mathcal{F}_T$: the value of T is equal to the value of S

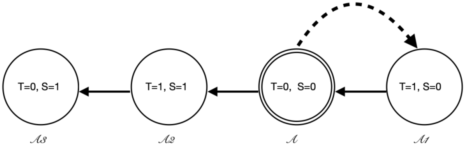

Figure: Robin's desire of tuition

- $\mathcal{A}(S) = 0, \mathcal{A}(T) = 0;$
- According to $\mathcal{F}_{T=1}$, $\mathcal{A}_{T=1}(T) = 1$, $\mathcal{A}_{T=1}(S) = 0$
- $\mathcal{A}_{S=1} < \mathcal{A}$, therefore $\langle \mathcal{F}, \mathcal{A}, < \rangle \models \neg D(T = 1)$

Expected Outcomes

Valid formulas of Logic ID

- $((X{=}x \mathrel{\Box\!\!\rightarrow} Y = y) \wedge PF(Y{=}y)) \rightarrow D(X{=}x)$
- $\vec{X}{=}\vec{x} \mathrel{\Box\!\!\rightarrow} D(\vec{Y}{=}\vec{y}) \rightarrow D(\vec{X}{=}\vec{x} \wedge \vec{Y}{=}\vec{y}))$, if $\vec{X}$ and $\vec{Y}$ are disjoint
- $(\vec{X}{=}\vec{x} \mathrel{\Box\!\!\rightarrow} \vec{Y}{=}\vec{y}) \wedge D(\vec{X}{=}\vec{x}) \rightarrow D(\vec{X}{=}\vec{x} \wedge \vec{Y}{=}\vec{y})$, if $\vec{X}$ and $\vec{Y}$ are disjoint

THANK YOU!