A Logic for Desire Based on Causal Inference*

Kaibo Xie[1] and Jialiang Yan[2]

[1] Tsinghua University, Beijing, China
xiekaibozju@gmail.com
[2] Tsinghua University, Beijing, China
yan-jl19@mails.tsinghua.edu.cn

Abstract. Reasoning about desire plays a significant role in philosophy, logic, and artificial intelligence etc. In this paper, we propose an interpretation of "desiring ϕ" based on whether the outcome brought about by "making ϕ true" is preferred by the agent. This interpretation of desire (including conditional desire) is different from the traditional approaches which directly interpret desire as preference. In order to formalize this idea, we construct a desire-causality model by combing the betterness model in preference logic and the causal model in the logic for causal reasoning. We also develop a logic for desire based on this semantics, and an axiomatization for our formal system is given.

1 Introduction

In many fields, ranging from logic to computer science, decision theory and semantics, the need for models to interpret *desire* has been recognized. This is not only because of the unique theoretical status of the concept of desire, but also because at the practical level, desire gives a method for the agent to interact with the world. To meet this needs, desire is usually interpreted from a decision theory perspective to be a choice among options consistent with preference, just like the idea noted in [21][1]. So people tend to describe desire in terms of preference structures in the formal model.

A number of works about modelling desire are based on this tradition. For example, in [4,5], the authors construct a goal-independent model of desire for BDI agents. And in [14,13], the logic of conditional desire is proposed, which is based on utility function on possible worlds. In preference logic, a desire behavior is usually regarded as an outcome of the entanglement of preferences and beliefs. [16,17,25] give various logical approaches to combine preference with belief or other informational attitudes, and study both their static and dynamic structure. The necessity for an independently semantic motivation to analyze desire modalities was proposed in [12]. Then some semantic models were devised,

*Kaibo Xie is supported by the Major Program of the National Social Science Foundation of China (NO. 17ZDA026).

[1] In [21], Stalnaker noted that 'wanting something is preferring it to certain relevant alternatives, the relevant alternatives being those possibilities that the agent believes will be realized if he does not get what he wants.'

e.g., [7,12,27]. In these models, the semantics of desire is related to conditional sentences, so possible worlds in the model are ranked by relative similarity or rationality. In a nutshell, the basic idea in this tradition has been that Desire(ϕ) is formalized by all possible worlds in which the object of desire exists are preferred to all possible worlds in which the object of desire does not exist.

We will take the same perspective and try to model the concept of desire by preference structure. However, in difference to the traditional formal approach mentioned above, when interpreting desire-propositions causality explicitly plays an important role in the framework. We propose that having a desire for a rational agent means that she prefers the outcome of making the object of the desire true brings about. In this way, we can construct a formal model combining causality and preference structure, and then a logic of desire based on causal inference will be proposed. To explain our motivation, let us take an example:

> *Robin was in charge of an important project in the company recently, and the tremendous pressure made him sleepless every night. Robin is a health-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 pills to help him sleep well, even if these pills have some side effects*

In this example, if Robin's desire to take sleeping pills is to be correctly predicted using the traditional approach, then all possible worlds in which taking sleeping pills is true will be preferred to all possible worlds in which sleeping pills are not taken. However, the problem is that if we compare the possible worlds only according to the truth of the proposition 'Robin takes sleeping pills', then there will be some possible worlds that seem to be remote so that we cannot get the correct prediction. For example, there would be a possible world w, in which Robin took sleeping pills but could not sleep, and possible world w', in which Robin did not take sleeping pills but fell asleep. It is clear that w' would be preferred to w, which would lead to a possible world in which Robin did not take sleeping pills but preferred to the possible world in which Robin took sleeping pills. Therefore, we cannot conclude that Robin wants to take sleeping pills. Of course, we can exclude these remote possible worlds according to the principle of context, but the criteria for determining which possible world is remote are not uniform or clear.

In order to make the modification, we need to find a formal method that allows us to explicitly determine the remote possible worlds in the framework and exclude them. We first argue that, as shown above, the challenge is that it is hard to select the possible worlds accurately for comparison when we deal with preference relations by depending only on the object of desire and not considering other propositions. What other propositions should also be taken into account to shrink the range of possible wolds? We suggest the outcome of acting on the object of the desire brings about should be included, because this will enable us to realize the intuitive idea that the preference of ϕ is not necessary condition of the desire of ϕ: if the agent wants ϕ, then he must prefer the outcomes that ϕ will bring when it is realized. In the above example, Robin's desire of taking pills depends on the outcome that this action brings about, i.e., having a good sleep, rather than the action of taking pills as he rejected the side effects of sleeping pills. Then a problem here is how to formalize the relation between an action and its outcome in such examples?

A possible answer to account for Robin's desire is that the desire of taking pills can be seen as a result of comparing the actual world (in which the pills are not taken) and a hypothetical situation in which the pills are taken: Robin desires taking sleeping pills only if the hypothetical situation with the pills taken is 'more preferred' than the actual world[2]. However, how to decide which hypothetical situation should the actual world be compared to? A possible solution is that we can make use of certain similarity ordering as in [15], such that the actual world is only compared to the most similar world(s) in which the pills are taken: Robin desires taking the pills because (i) the most similar world with the pills taken is a possible world in which Robin has a good sleep after he takes the pills (ii) the possible world in which Robin has a good sleep after taking the pills is more preferred than the actual world (it is reasonable to consider a possible world in which Robin takes the pills but still has a bad sleep as a remote possible world.)

However, this approach does not seem to solve the problem when the similarity ordering is controversial. Consider the following example:

> *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.* [3]

Consider two possible worlds: in world w_1, Robin studies in the Netherlands and pays for the tuition; in world w_2, he doesn't study in the Netherlands but still pays for the tuition. If we consider that w_1 is closer to the actual world than w_2, then according to the previous formalization of desire, we will say Robin wants to pay for the tuition which is unreasonable given the situation in the actual world. Therefore we have to refute that w_1 is more similar than w_2 compared with the actual world in this example. However, due to the different standards in the process of refusal, the reasons for rejection will be subjective, resulting in many disputes.

Causality, hence, is needed to account for the examples: in the first example, the pills are desired because taking sleeping pills causes a good sleep; In the second example, though a possible world in which Robin studies in the Netherlands is more likely to be a possible world in which Robin pays for the tuition, Robin's paying the tuition cannot bring about his studying in the Netherlands. So the possible world resulting from "forcing Robin to pay tuition", is a possible world in which Robin pays intuition but does not study in Netherlands. Such a possible

[2] In [12], Heim proposed a way of interpreting desire along this line

[3] The original version of this example comes from discussion with Maria Aloni.

world is apparently worse then the actual world. It explains our intuition that Robin does not desire paying tuition.

Embedding causality into account of desire not only can formalize the relation between an action and its outcome, it also can exclude the remote possible worlds explicitly in the model by deciding which variable's value should be changed and which not. Especially, it shows the similar idea with *Ceteris Paribus* preference (see [28,23,8]), which means the preference under the condition that everything else being equal. When we evaluate an agent's desire, it is important to assume some factors should keep constant. In the second example, Robin doesn't want to pay for the tuition because if we change the fact of Robin's paying to be positive (has paid for the tuition), and keep the fact of Robin's studying constant (studying at home), then a possible world will be obtained where Robin has paid for the tuition but still cannot study in the Netherlands, and which is the worst situation for Robin. Otherwise, if the fact of Robin's studying is also be changed to be positive, we will obtain a better situation than the reality (after paying for the tuition Robin studies in the Netherlands), and hence we will conclude Robin wants to pay for the tuition, which seems a wrong prediction. In this way, causality provide an good predictor of desire behaviors by controlling the value of the event.

In a word, in this paper we argue that wanting ϕ for an agent means the agent has preference to certain outcomes that ϕ brings about. This paper is organized as follows. In Sec.2, we will first review the classical methods of modelling preference and causality. Then, in Sec.3, by combing the idea of these two modelling methods, we construct a new model to capture the concept of desire that we proposed in this paper. A logic will be given in Sec.4, which is built on the logic for causal models. Finally, in Sec.5, we will discuss the conditional desire in term of causality.

2 Modelling methods for preference and causality

As mentioned above, we propose to interpret desire by combining causality and preference structures. So we need a formal account of these two elements. In this section, as a preliminary, we will introduce the approaches to modelling preferences and causality respectively in this section. For preferences, we mainly focus on a static *betterness* model (cf.[26]). And for causality, we will introduce the interventionist approach to causal model (cf. [18]). In the section 3, we will combine these two modelling approaches to construct a novel model for our desire logic.

2.1 Modelling methods for preference

Preference is an important factor in decision making. In the literature, there are two types of preferences that are often discussed: *preference over states of affairs* and *preference over objects.* The former indicates the preference arises from comparisons between outcomes, actions, or situations, etc. And in logical language, states of affairs are represented as propositions which are modeled by sets of possible words. In our proposal, we will focus on this type of preference.

Betterness model Following from the traditional idea, such as [2,9], the preference between propositions can be treated as a binary relation which is represented by a unary modality in the possible worlds model. In [26,24], this relation was reinterpreted as a 'betterness' relation over possible worlds, which inherited the idea from [28]. Based on a betterness relation, a modal preference logic can be obtained, and the processing of preferences in our novel desire logic proposed in this paper will be primarily based on it.

More precisely, a modal language expressing betterness relation is given in [26], which is a modal language that contains a $\langle\leq\rangle$ operator. The formula $\langle\leq\rangle\phi$ can be read as 'ϕ is true in some world that is at least as good as the current world', which express the meaning 'some agent prefers ϕ' in natural language. And $\langle<\rangle$ represents the strict betterness which says 'ϕ is true in some world that is strictly better than the current world'. The model for the language is known as a "betterness model".

Definition 1 (Betterness Model) *A betterness model is a tuple $M = (W, \leq, V)$ where W is a set of possible worlds, $\leq$ is a reflexive and transitive relation (the 'betterness' pre-order) over these worlds, and V is a valuation assigning truth values to proposition letters at worlds.*

For any two possible worlds w_1 and w_2, $w_1 \leq w_2$ means 'w_2 is at least as good as w_1'. In addition, if $w_1 \leq w_2$ but not $w_2 \leq w_1$, then w_2 is strictly better than w_1, written as $w_1 < w_2$.

As we'll see in later sections, we will combine the betterness model with causal model, that is, the pre-order will be embedded into the causal model. In this way, we can rank the assignments in the model. Thus, we can express desire in this way: the transformation between possible worlds caused by causality is a transition to a better world.

Definition 2 (Truth in the Model) *Given a betterness model $M = (W, \leq, V)$, and a world $w \in W$, the truth of preference formulas are defined as follow:*

- *$M, w \models \langle\leq\rangle\phi$ iff for some w' with $w \leq w'$, $M, w' \models \phi$.*

So, in this model, what is preferred is true in a better world, given by the betterness relation.

What we have discussed above are preferences over propositions, which is exactly the type of preference we will focus on in our proposal. There are also preferences between individual objects. Considering the preference over objects naturally leads us to think of properties those objects have. According to [16], a preference over objects can be derived from a more basic source, which is called a *priority base* representing the reasoning underlying preference, and hence it is able to discuss the reason for having a preference. Based on this, the theory of two-level perspective on preference has been proposed. We will not present the this work here, more details can be found in [17,16,25].

To satisfy our needs, we are currently considering only the simplest preference structure, i.e., the partial ordering over possible worlds. So we have presented only a minimal static system of betterness logic here. The logic of desire we propose in this paper can be extended by complicating the structure of preferences, especially by the dynamic preference (see [17,16]), which we will focus on in the future work. Following, we will introduce the classical causal model, and will show that it is natural to embed preference structures into the causal model to interpret desire.

2.2 Modelling methods for causality

In recent years, there is a development in formal approaches to causal reasoning and causal learning, based on the work of structural causal models in [18]. The logic based on causal models, which goes back to [6] and was further developed in [10,19,3]. We will briefly introduce the causal modelling approach on which our study of desire will be built.

Causal models which represent the causal relationships between a finite set of variables. The variables are sorted into the set $\mathcal{U}$ of *exogenous* variables (those whose value is causally independent from the value of every other variable in the system) and the set $\mathcal{V}$ of *endogenous* variables (those whose value is completely determined by the value of other variables in the system). For each variable $X \in \mathcal{U} \cup \mathcal{V}$, $\mathcal{R}(X)$ indicates the possible values that X can take. All this information is provided by a *signature* $\mathcal{S} = \langle \mathcal{U}, \mathcal{V}, \mathcal{R} \rangle$.
A causal model is formally defined as follows.

Definition 3 (Causal model) *A* causal model *is a triple* $\langle \mathcal{S}, \mathcal{F}, \mathcal{A} \rangle$ *where* $\mathcal{S} = \langle \mathcal{U}, \mathcal{V}, \mathcal{R} \rangle$ *and*

- $\mathcal{F} = \{\mathcal{F}_{V_j} \mid V_j \in \mathcal{V}\}$ *assigns, to each endogenous variable* V_j*,* $\mathcal{F}_{V_j}$ *is a mapping from all assignments*[4] *to* $\mathcal{U} \cup \mathcal{V} \setminus \{V_j\}$ *to* $\mathcal{R}(V_j)$*.* $\mathcal{F}$ *is assumed to be recursive*[5]
- $\mathcal{A}$ *is the* ***valuation*** *function, assigning to every* $X \in \mathcal{U} \cup \mathcal{V}$ *a value* $\mathcal{A}(X) \in \mathcal{R}(X)$*. For each* $V_j \in \mathcal{V}$*,* $\mathcal{A}$*'s assignment to* V_j complies with $\mathcal{F}_{V_j}$*, that is:* $\mathcal{A}(V_j) = \mathcal{F}_{V_j}(\mathcal{A}_{V_j})$ *where* $\mathcal{A}_{V_j}$ *is the partial assignment in* $\mathcal{A}$ *to* $\mathcal{U} \cup \mathcal{V} \setminus \{V_j\}$ *(for any* $X \in \mathcal{U} \cup \mathcal{V} \setminus \{V_j\}$*,* $\mathcal{A}(X) = \mathcal{F}_{V_j}(X)$*)*

The intuition behind $\mathcal{F}$ (which is called the set of structural functions) is that: $\mathcal{F}$ is a set of causal rules such that for each endogenous variable V, $\mathcal{F}_V \in \mathcal{F}$ gives the information about how other variables determine the value of V. The intuition behind the requirement that $\mathcal{A}$ has to comply with $\mathcal{F}$ is that the actual assignment should not violate any of the causal rules. The intuition behind the

[4] An assignment a to a set of variables S is a mapping from S to $\bigcup_{X \in S} \mathcal{R}(X)$ such that for each $Z \in S$, $a(Z) \in \mathcal{R}(Z)$

[5] For the definition of recursiveness, see [10]

recursiveness of $\mathcal{F}$ is that we assume that the causal model should contain no causal loops.

In Halpern's formal system for causal reasoning (in the most well-known version), the language describing the causal models is given by extending the propositional language (whose atomic sentences are $X = x$ saying the value of X is x) with counterfactuals $\vec{X} = \vec{x} \;\Box\!\!\rightarrow \phi$ [6] which should be read as a counterfactual conditional: "if the variables in $\vec{X}$ were set to the values $\vec{x}$, respectively, then ϕ would be the case".

Definition 4 (Language $\mathcal{L}_C$) *Formulas φ of the language $\mathcal{L}_C$ based on the signature $\mathcal{S}$ are given by*

$$\varphi ::= Y{=}y \mid \neg\varphi \mid \varphi \wedge \varphi \mid \vec{X}{=}\vec{x} \;\Box\!\!\rightarrow \varphi$$

for $Y \in \mathcal{U} \cup \mathcal{V}$, $y \in \mathcal{R}(Y)$ and $\vec{X}{=}\vec{x}$ an assignment on $\mathcal{S}$.

In Halpern's original version in [10], there are restrictions of this language such that (i) the $\vec{X}$ in the antecedent of a counterfactual $\vec{X} = \vec{x} \;\Box\!\!\rightarrow \phi$ can only be exogenous variables (ii) the consequence of a counterfactual can only be an atomic sentence of the form $Y = y$

In this paper we will work on a language with less restrictions such that: (i) the $\vec{X}$ in the antecedent of a counterfactual $\vec{X} = \vec{x} \;\Box\!\!\rightarrow \phi$ can only be any variable in $\mathcal{U} \cup \mathcal{V}$ (ii) $\Box\!\!\rightarrow$ can be right-nested in the consequence of $\vec{X} = \vec{x} \;\Box\!\!\rightarrow \phi$ and ϕ can contain boolean connectives.

The language that drops the restriction in Halpern's causal language is discussed in [1]. For its semantic interpretation, the truth condition of $\vec{X} = \vec{x} \;\Box\!\!\rightarrow \phi$ in the extended language is slightly different from [10], as it is extended to the cases in which $\vec{X}$ may contain exogenous variables (the Boolean cases are evaluated as usual):

Definition 5 (Counterfactuals) *Let $M = \langle \mathcal{S}, \mathcal{F}, \mathcal{A} \rangle$ be a causal model $\mathcal{S}$; let $\vec{X}{=}\vec{x}$ be an assignment on $\mathcal{S}$. The causal model $M_{\vec{X}=\vec{x}} = \langle \mathcal{F}_{\vec{X}=\vec{x}}, \mathcal{A}^{\mathcal{F}}_{\vec{X}=\vec{x}} \rangle$, resulting from an intervention setting the values of variables in $\vec{X}$ to $\vec{x}$, is such that*

- *$\mathcal{F}_{\vec{X}=\vec{x}}$ 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 unique valuation where (i) the value of each exogenous variable not in $\vec{X}$ is exactly as in $\mathcal{A}$, (ii) the value of each exogenous variable X_i in $\vec{X}$ is the provided x_i, and (iii) the value of each endogenous variable complies with its* new *structural function (that in $\mathcal{F}_{\vec{X}=\vec{x}}$).* [7]

[6] Here we slightly modified the notation in [10] for readability because counterfactual operators may differ in different systems e.g. [3]. In some literature, a counterfactual $\vec{X} = \vec{x} \;\Box\!\!\rightarrow \phi$ will be written as $[\vec{X} = \vec{x}] \;\Box\!\!\rightarrow \phi$. Here we will uniformly use the former way of notation in order to avoid confusion.

[7] Note that, since $\mathcal{F}$ is recursive, the valuation $\mathcal{A}^{\mathcal{F}}_{\vec{X}=\vec{x}}$ is uniquely determined.

The truth condition of a counterfactual in a causal model is defined as:
$\langle \mathcal{S}, \mathcal{F}, \mathcal{A} \rangle \models \vec{X} = \vec{x} \,\Box\!\!\rightarrow \phi$ *iff* $\langle \mathcal{S}, \mathcal{F}_{\vec{X}=\vec{x}}, \mathcal{A}^{\mathcal{F}}_{\vec{X}=\vec{x}} \rangle \models \phi$

[1] provided an axiomatization of $\mathcal{L}_C$ with respect to the causal models. As we will also apply the causal modelling approach in our analysis of desire, our logic will be built on the formal system in [1].

3 modelling desire by combining the two approaches

We will introduce a 'desire-causality model' in this section. The model results from extending a causal model by a pre-order over all possible assignments to all causal variables. The pre-order over all possible assignments to all causal variables aims to reflect the underlying philosophical idea of 'the preference that makes pleasure'[8]. We set that this order is stable and this preference is for its own sake. Then to capture the causal relationship, we will employ the causal model.

Definition 6 (Desire-Causality Model) *Let* $\mathcal{S} = \langle \mathcal{U}, \mathcal{V}, \mathcal{R} \rangle$ *be a signature for a causal model* $\langle \mathcal{S}, \mathcal{F}, \mathcal{A} \rangle$*. A desire-causality model based on* $\mathcal{S}$ *is a tuple* $\langle \mathcal{F}, \mathcal{A}, \leq \rangle$*.*

- $\mathcal{F} = \{f_{V_j} \mid V_j \in \mathcal{V}\}$ *assigns, to each endogenous variable* V_j*, is a mapping from all assignments*[9] *to* $\mathcal{U} \cup \mathcal{V} \setminus \{V_j\}$ *to* $\mathcal{R}(V_j)$*.* $\mathcal{F}$ *is assumed to be recursive*[10]
- $\mathcal{A}$ *is the* ***valuation*** *function, assigning to every* $X \in \mathcal{U} \cup \mathcal{V}$ *a value* $\mathcal{A}(X) \in \mathcal{R}(X)$*. For each* $V_j \in \mathcal{V}$*,* $\mathcal{A}$*'s assignment to* V_j complies with $\mathcal{F}_{V_j}$*, that is:* $\mathcal{A}(V_j) = \mathcal{F}_{V_j}(\mathcal{A}_{V_j})$ *where* $\mathcal{A}_{V_j}$ *is the partial assignment in* $\mathcal{A}$ *to* $\mathcal{U} \cup \mathcal{V} \setminus \{V_j\}$ *(for any* $X \in \mathcal{U} \cup \mathcal{V} \setminus \{V_j\}$*,* $\mathcal{A}(X) = \mathcal{A}_{V_j}(X)$*)*
- $\leq$ *is a pre-order over all possible assignments to* $\mathcal{U} \cup \mathcal{V}$*, if an assignment* $\mathcal{A}_2 \leq \mathcal{A}_1$ *but it is not the case that* $\mathcal{A}_1 \leq \mathcal{A}_2$*, then* $\mathcal{A}_1$ *is strictly preferred to* $\mathcal{A}_2$*, denoted as* $\mathcal{A}_2 < \mathcal{A}_1$*.*

$\leq$ represents agents' preference. As discussed in Sec.2, we employ the classical approach (cf. [11,26,29,16]) of representing preference with the betterness relation, we read $\mathcal{A}_2 \leq \mathcal{A}_1$ as the assignment $\mathcal{A}_1$ is more preferred than $\mathcal{A}_2$.

For simplicity, we will write $X_1, ..., X_n$ as $\vec{X}$ and write a sequence $X_1{=}x_1, ..., X_n = x_n$ as $\vec{X}{=}\vec{x}$. We will also write $X_1{=}x_1 \wedge ... \wedge X_n{=}x_n$ as $\vec{X}{=}\vec{x}$ when the context allows,

Definition 7 (Language $\mathcal{L}_{DC}$) *Formulas* ϕ *of the language* $\mathcal{L}_{DC}$ *based on* $\mathcal{S}$ *are given by*

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

[8] See more philosophical discussions about this in [20,22]

[9] An assignment a to a set of variables S is a mapping from S to $\bigcup_{X \in S} \mathcal{R}(X)$ such that for each $Z \in S$, $a(Z) \in \mathcal{R}(Z)$

[10] For the definition of recursiveness, see [10]

$\mathcal{L}_{DC}$ is the language that extends $\mathcal{L}_C$ with a desire-operator: $D(\vec{X}{=}\vec{x})$ represents the desire that wanting $\vec{X}$ to be X_1,...,X_n to be x_n. Note that we do not consider the iteration of desire in this paper. $(\vec{X}{=}\vec{x}) \Box\!\!\rightarrow \phi$ denotes the counterfactual conditional which means if $X_1{=}x_1$,...,$X_n{=}x_n$ were true, then ϕ would be the case.[11]

Definition 8 (Intervention on desire-causality models) *Let $M = \langle \mathcal{F}, \mathcal{A}, \leq \rangle$ be a DC-model based on $\mathcal{S}$; let $\vec{X}{=}\vec{x}$ be an assignment on $\mathcal{S}$. The DC-model $M_{\vec{X}=\vec{x}} = \langle \mathcal{F}_{\vec{X}=\vec{x}}, \mathcal{A}^{\mathcal{F}}_{\vec{X}=\vec{x}}, \leq\rangle$ is the model that results from an intervention setting the values of variables in $\vec{X}$ to $\vec{x}$, where $\mathcal{F}_{\vec{X}=\vec{x}}$, and $\mathcal{A}^{\mathcal{F}}_{\vec{X}=\vec{x}}$ is defined as in Definition 5.*

In fact, an intervention that sets a variable in X to the value x is an operation that maps a given model M to a new model $M_{X=x}$, in this way, X is cut off from all its causal dependencies and fixed to the value x.

Definition 9 (Semantics of $\mathcal{L}_{DC}$) *Let $\langle \mathcal{F}, \mathcal{A}, \leq\rangle$ be a DC-model based on $\mathcal{S}$, the truth condition of the formulas in $\mathcal{L}$ are given by (boolean cases are defined as usual):*

– $\langle \mathcal{F}, \mathcal{A}, \leq\rangle \models X{=}x$ *iff* $\mathcal{A}(X) = x$
– $\langle \mathcal{F}, \mathcal{A}, \leq\rangle \models (\vec{X}{=}\vec{x}) \Box\!\!\rightarrow \phi$ *iff* $\langle \mathcal{F}_{\vec{X}=\vec{x}}, \mathcal{A}^{\mathcal{F}}_{\vec{X}=\vec{x}}, \leq\rangle \models \phi$
– $\langle \mathcal{F}, \mathcal{A}, \leq\rangle \models D(\vec{X}{=}\vec{x})$ *iff* $\mathcal{A} < \mathcal{A}^{\mathcal{F}}_{\vec{X}=\vec{x}}$

The second item says that the valuations that setting the variables in $\vec{X}$ to $\vec{x}$ are preferred to the one does not doing that setting, which reflects the idea of 'for its own sake'. And the final item defined the desire modality intuitively: wanting $\vec{X}$ to be X_1,...,X_n to be x_n means the possible assignment that sets variables in $\vec{X}$ to be $\vec{x}$ is preferred for its own sake to the real one.

Now let us revisit the example *Robin's desire of taking sleeping pills*. Let $X{=}1$ stand for 'taking sleeping-help pills' and $X{=}0$ stand for its negation; and $Y{=}1$ stands for 'having a good sleep' and $Y{=}0$ stands for its negation. 'Robin wants to take sleeping-help pills' can be formulated as $D(X{=}1)$. This model can be described by Figure 1.

In Figure 1, we can see that for any assignments $\mathcal{A}_i$ and $\mathcal{A}_j$ $(i, j \in \{1, 2, 3\})$ such that $\mathcal{A}_i(Y) = 1$ and $\mathcal{A}_j(Y) = 0$, we have $\mathcal{A}_j < \mathcal{A}_i$. Since $\langle \mathcal{F}_{X=1}, \mathcal{A}^{\mathcal{F}}_{X=1}, \leq\rangle \models X = 1 \wedge Y = 1$, namely $\langle \mathcal{F}, \mathcal{A}^{\mathcal{F}}, \leq\rangle \models (X{=}1) \Box\!\!\rightarrow (X = 1 \wedge Y{=}1)$, the assignment results from forcing $X = 1$ to be true at $\mathcal{A}$ is $\mathcal{A}_2$. Since $\mathcal{A} < \mathcal{A}_2$, we have $\langle \mathcal{F}, \mathcal{A}, \leq\rangle \models D(X = 1)$.

Now we apply the same way of modelling to Robin's desire of studying in the Netherlands. Let $S{=}1$ stands for studying in Netherlands and $S{=}0$ for not; And $T{=}1$ stands for paying tuition fee and $T{=}0$ for not. The corresponding desire-causality model is illustrated in Figure 2.

We can check that $D(T{=}1)$ does not hold at $\mathcal{A}$, which is the actual world. The result fits our intuition that Robin does not desire paying for the tuition.

[11] In our language, $\vec{X} = \vec{x}$ for simplicity, when $\vec{X}$ we will not distinguish $\vec{X} = \vec{x} \Box\!\!\rightarrow \phi$ in the axioms

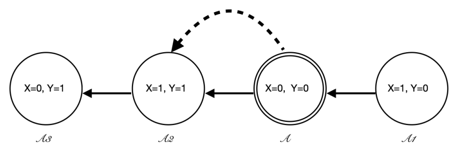

Fig. 1. Robin's desire of taking sleeping pills

The double circle represents the real assignment; the solid arrow is the pre-order $\leq$ and the dashed arrow is the counterfactual transition by setting the value of X to 1.

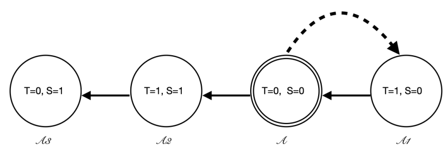

Fig. 2. Robin's desire of tuition

4 The logic for desire-causality models

Our formal system for desire causality models will be built on the logic for causal models. We choose to use the language that allows exogenous variables in the antecedents, so our axiomatization will be based on the system in [1], which consists of $\mathbf{A_1}$ to $\mathbf{A_7}$[12] (including propositional tautology and MP rule) in the Table 1. We will name the axiom system consisting of $\mathbf{A_1}$ to $\mathbf{A_7}$ as L_C.

The axiom system for $\mathcal{L}_{DC}$ (the language extending $\mathcal{L}_C$ with desire operator) with respect to desire causality models is given by extending the L_C with $\mathbf{A_8}$ to $\mathbf{A_{12}}$ and generalization rule. We will name it as L_{DC}. The axioms of L_{DC} is listed as in Table 1.

The soundness of $\mathbf{A_1}$ to $\mathbf{A_6}$ has been proven in [10] [13] and the soundness of $\mathbf{A_7}$ is shown in [1]. The soundness of $\mathbf{A_\neg}$, $\mathbf{A_\wedge}$ and is straightforward, that is: intervention is deterministic, so Boolean operators can be distributed into the consequence. $\mathbf{A_{[][]}}$ means that, if two interventions are applied sequentially, the later intervention will over-wright the previous intervention.

The intuition behind $\mathbf{A_8}$ is that anything that already holds at the actual world will not be desired; $\mathbf{A_9}$ indicates that if a situation is desired conditional

[12] Notice that in $\mathbf{A_6}$, $X \rightsquigarrow V$ is an abbreviation of the following formula:

$$\bigvee_{\substack{\vec{z} \in \mathcal{R}(\mathcal{U} \cup \mathcal{V} \setminus \{X, V\}), \\ \{x_1, x_2\} \subseteq \mathcal{R}(X),\, x_1 \neq x_2, \\ \{v_1, v_2\} \subseteq \mathcal{R}(V),\, v_1 \neq v_2}} \vec{Z}{=}\vec{z}, X{=}x_1 \,\Box\!\!\rightarrow (V{=}v_1) \;\wedge\; \vec{Z}{=}\vec{z}, X{=}x_2 \,\Box\!\!\rightarrow (V{=}v_2),$$

[13] In a slightly different form

P	φ	for φ an instance of a propositional tautology
MP	From φ_1 and $\varphi_1 \to \varphi_2$ infer φ_2	
Generalization	From ϕ infer $\vec{X} = \vec{x} \mathrel{\Box\!\!\to} \phi$	
$\mathbf{A_1}$	$\vec{X}{=}\vec{x} \mathrel{\Box\!\!\to} Y{=}y \to \neg\vec{X}{=}\vec{x} \mathrel{\Box\!\!\to} Y{=}y'$	for $y, y' \in \mathcal{R}(Y)$ with $y \neq y'$
$\mathbf{A_2}$	$\bigvee_{y\in\mathcal{R}(Y)} \vec{X}{=}\vec{x} \mathrel{\Box\!\!\to} Y{=}y$	
$\mathbf{A_3}$	$\big(\vec{X}{=}\vec{x} \mathrel{\Box\!\!\to} (Y{=}y) \wedge \vec{X}{=}\vec{x} \mathrel{\Box\!\!\to} (Z{=}z)\big) \to \vec{X}{=}\vec{x}, Y{=}y \mathrel{\Box\!\!\to} (Z{=}z)$	
$\mathbf{A_4}$	$\vec{X}{=}\vec{x}, Y{=}y \mathrel{\Box\!\!\to} (Y{=}y)$	
$\mathbf{A_5}$	$\big(\vec{X}{=}\vec{x}, Y{=}y \mathrel{\Box\!\!\to} (Z{=}z) \wedge \vec{X}{=}\vec{x}, Z{=}z \mathrel{\Box\!\!\to} (Y{=}y)\big) \to \vec{X}{=}\vec{x} \mathrel{\Box\!\!\to} (Z{=}z)$	for $Y \neq Z$
$\mathbf{A_6}$	$(X_0 \rightsquigarrow X_1 \wedge \cdots \wedge X_{k-1} \rightsquigarrow X_k) \to \neg(X_k \rightsquigarrow X_0)$	
$\mathbf{A_7}$	$U{=}u \leftrightarrow \vec{X}{=}\vec{x} \mathrel{\Box\!\!\to} U{=}u$	for $U \in \mathcal{U}$ with $U \notin \vec{X}$
$\mathbf{A}_\neg$	$\vec{X}{=}\vec{x} \mathrel{\Box\!\!\to} \neg\varphi \leftrightarrow \neg\vec{X}{=}\vec{x} \mathrel{\Box\!\!\to} \varphi$	
$\mathbf{A}_\wedge$	$\vec{X}{=}\vec{x} \mathrel{\Box\!\!\to} (\varphi_1 \wedge \varphi_2) \leftrightarrow (\vec{X}{=}\vec{x} \mathrel{\Box\!\!\to} \varphi_1 \wedge \vec{X}{=}\vec{x} \mathrel{\Box\!\!\to} \varphi_2)$	
$\mathbf{A}_{[][]}$	$\vec{X}{=}\vec{x} \mathrel{\Box\!\!\to} (\vec{Y}{=}\vec{y} \mathrel{\Box\!\!\to} \varphi) \leftrightarrow \vec{X'}{=}\vec{x'}, \vec{Y}{=}\vec{y} \mathrel{\Box\!\!\to} \varphi$	with $\vec{X'}{=}\vec{x'}$ the subassignment of $\vec{X}{=}\vec{x}$ for $\vec{X'} := \vec{X} \setminus \vec{Y}$
$\mathbf{A_8}$	$\vec{X}{=}\vec{x} \to \neg D(\vec{X}{=}\vec{x})$	
$\mathbf{A_9}$	$(\vec{X} = \vec{x} \mathrel{\Box\!\!\to} Z = z) \to (\vec{X} = \vec{x} \mathrel{\Box\!\!\to} D(\vec{Y} = \vec{y})) \leftrightarrow (\vec{X} = \vec{x}, Z = z \mathrel{\Box\!\!\to} D(\vec{Y} = \vec{y}))$	if $\vec{Y} = \vec{y}$ is a full assignment of $\mathcal{U} \cup \mathcal{V}$ and $Z \notin \vec{X}$
$\mathbf{A_{10}}$	$(\vec{X}{=}\vec{x} \mathrel{\Box\!\!\to} \vec{Y}{=}\vec{y}) \to (D(\vec{X}{=}\vec{x}) \leftrightarrow D(\vec{X}{=}\vec{x} \wedge \vec{Y'}{=}\vec{y'}))$	with $\vec{Y'} = \vec{y'}$ the sub-assignment of $\vec{Y}{=}\vec{y}$ for $\vec{Y'} := \vec{Y} \setminus \vec{X}$
$\mathbf{A_{11}}$	$(\vec{X}{=}\vec{x} \mathrel{\Box\!\!\to} D(\vec{Y}{=}\vec{y})) \wedge D(\vec{X}{=}\vec{x}) \to D(\vec{X'}{=}\vec{x'}, \vec{Y}{=}\vec{y})$	with $\vec{X'} = \vec{x'}$ the sub-assignment of $\vec{X}{=}\vec{x}$ for $\vec{X'} := \vec{X} \setminus \vec{Y}$

Table 1. Axioms of L_{DC}

on X, then it is equivalent to say that the situation is also desired conditional on A together with X's consequence. $\mathbf{A_{10}}$ indicates that if X is desired, it is equivalent to say that X together with the consequence of X is also desired. $\mathbf{A_{11}}$ indicates that if X is desired and Y is desired conditional on X, then $X \wedge Y$ must be desired as well (as far as they are disjoint).

Theorem 1 $\mathbf{A_8}$-$\mathbf{A_{11}}$ *is sound with respect to desire causality models.*

Proof. Due to the limitation of space, the proof is put in an external link: Link

Let $\mathbb{C}$ be the set of all maximally L_{DC}-consistent set of $\mathcal{L}_{DC}$-formulas. We will show that for any $\Gamma \in \mathbb{C}$, there is a DC-model M^Γ that models Γ.

Definition 10 (Building the canonical model) *For any $\Gamma \in \mathbb{C}$, the canonical model $M^\Gamma = \langle \mathcal{F}^\Gamma, \mathcal{A}^\Gamma, \leq^\Gamma \rangle$ is defined as below.*

– *Let $\vec{X}$ be a tuple with all variables in $\mathcal{U} \cup \mathcal{V} \setminus V$. For each endogenous variable $V \in \mathcal{V}$, the structural function f_V^Γ is defined in such a way that: for each possible assignment $\vec{X} = \vec{x}$ and any $v \in \mathcal{R}(V)$*

$$f_V^\Gamma(\vec{X} = \vec{x}) = v \quad \textit{if and only if} \quad \vec{X}{=}\vec{x} \mathrel{\Box\!\!\to} V{=}v \in \Gamma$$

– $\leq$ *is defined by: for each full assignments* $\vec{X} = \vec{x_1}$ *and* $\vec{X} = \vec{x_2}$ *where* $\vec{X}$ *are all the variables in* $\mathcal{U} \cup \mathcal{V}$, $\vec{x_1} \leq \vec{x_2}$ *iff* $\vec{X} = \vec{x_1} \,\square\!\!\rightarrow D(\vec{X} = \vec{x_2}) \in \Gamma$ *or* $\vec{x_1} = \vec{x_2}$
– $\mathcal{A}^\Gamma$ *is defined as: for each* $X \in \mathcal{U} \cup \mathcal{V}$ *and any* $x \in \mathcal{R}(X)$, $\mathcal{A}^\Gamma(X) = x$ *iff* $X = x \in \Gamma$

$\mathbf{A_1}$-$\mathbf{A_7}$ *guarantee that: (i) for each variable* $V \in \mathcal{U} \cup \mathcal{V}$ *and any assignment* $\vec{X} = \vec{x}$ *to* $\mathcal{U} \cup \mathcal{V} \setminus V$*, there is exactly one value* $v \in \mathcal{R}(V)$ *such that* $\vec{X}{=}\vec{x} \,\square\!\!\rightarrow V{=}v \in \Gamma$*; (ii) for each variable* $X \in \mathcal{U} \cup \mathcal{V}$ *there is exactly one value* $x \in \mathcal{R}(X)$ *such that* $X = x \in \Gamma$*. Therefore* $\mathcal{F}^\Gamma$ *and* $\mathcal{A}^\Gamma$ *is well defined.*

The definition of $\leq^\Gamma$ *guarantees that* $\leq^\Gamma$ *is reflexive. In addition* $\mathbf{A_8}$ *guarantees that* $\leq^\Gamma$ *is a pre-order.*

Theorem 2 (Truth Lemma for the D-free formulas in $\mathcal{L}_{DC}$) $\vec{X} = \vec{x} \,\square\!\!\rightarrow \vec{Y} = \vec{y} \in \Gamma$ *iff* $\langle \mathcal{F}^\Gamma, \mathcal{A}^\Gamma, \leq^\Gamma \rangle \models \vec{X}{=}\vec{x} \,\square\!\!\rightarrow Y{=}y$

Proof. The truth lemma for the sub-language without desire operator has been proven in [1]: for any $\phi \in \mathcal{L}_C$ (the sub-language of $\mathcal{L}_{DC}$ without the D operator), $\phi \in C$ iff $\mathcal{F}^C, \mathcal{A}^C \models \phi$ where C is maximal consistent set with respect to L_C, where $\mathcal{F}^C$ and $\mathcal{A}^C$ is defined in the same way as in Definition 10. Notice that the truth condition of formulas does not involve preference ordering, we can apply exactly the same strategy of [1] and show that for any formula of the form $[\vec{X} = \vec{x}]\vec{Y} = \vec{y}$, $\vec{X} = \vec{x} \,\square\!\!\rightarrow \vec{Y} = \vec{y} \in \Gamma$ iff $\langle \mathcal{F}^\Gamma_{\vec{X}=\vec{x}}, \mathcal{A}^{\Gamma, \mathcal{F}^\Gamma}_{\vec{X}=\vec{x}}, \leq^\Gamma \rangle \models Y{=}y$ iff $\langle \mathcal{F}^\Gamma, \mathcal{A}^\Gamma, \leq^\Gamma \rangle \models \vec{X}{=}\vec{x} \,\square\!\!\rightarrow Y{=}y$

Theorem 3 (Truth Lemma for $\mathcal{L}_{DC}$) *Let* $\Gamma \in \mathbb{C}$*, for any* $\phi \in \mathcal{L}_{DC}$*,* $\phi \in \Gamma$ *iff* $\langle \mathcal{F}^\Gamma, \mathcal{A}^\Gamma, \leq^\Gamma \rangle \models \phi$

Proof. Due to the limitation of space, the proof is put in an external link: Link

Theorem 4 L_{DC} *is complete with respect to desire causality models.*

5 Discussion

In the daily life, desire can not only be of the form "I desire that ψ", but also be of the form "Conditional on ϕ, I would desire ψ ". In [13,14], the author suggested an approach to model conditional desire only by utility or preference. Instead, in our language based on desire-causality model, the conditional desire can be expressed by causality in such a way: $\vec{X} = \vec{x} \,\square\!\!\rightarrow D(\vec{Y} = \vec{y})$.

In our model, if the condition $\vec{Y} = \vec{y}$ is not a full assignment, the axiom $\mathbf{A_{10}}$ will be not valid. To see this, consider the following counterexample:

> *Robin was seriously ill. He actively cooperated with the doctor's treatment and hoped to recover soon. The doctor gave him two treatment options. One was surgery, which was expensive but would make him heal quickly. Another option is conservative treatment, which costs less but lasts longer. The best course of action for Robin was surgery, so he chose to have it.*

In this case, both ‘surgery’ and ‘conservative treatment’ cause ’Robin’s recovery’. Note that in reality, when faced with both treatment options, Robin opted for surgery. But if surgery is not available, then naturally, Robin wants to be treated conservatively. Because according to our analysis, Robin’s current preference is to cure the disease, which can only be achieved through conservative treatment at this time, so Robin desires conservative treatment. The causal relations are shown as in the Figure 3 below, and the dotted lines represent the causal relations events.

Now we can formulate this example in the DC model: let $X = 1$ and $Y = 1$ stand for ’surgery’ and ’conservative treatment’ respectively, and $X = 0$ and $Y = 0$ stand for their respective negations; and let $Z = 1$ stand for ’Robin’s recovery’, likewise, $Z = 0$ stand for its negation. The formula $D(Y = 1)$ represents the desire of Robin that being treated conservatively. The following Figure 4 depicts this example:

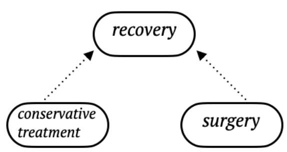

Fig. 3. Causal relations in the example of Robin’s treatment

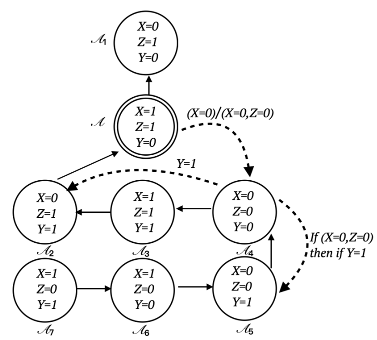

Fig. 4. Robin’s desire of treatment

In Figure 4, The counterfactual transitions represented by the dashed arrows are based on the causal relations described in Figure 3. The transformation *if* $(X = 0, Z = 0)$ *then if* $Y = 0$ says that under the condition that setting the values of both variable X and Z to be 0, we set the value of the variable Y to be 0.

By Figure 3 and 4, it is easy to check that the truth of both $X = 0 \boxright Z = 0$ and $X = 0 \boxright D(Y = 1)$. However, if we set both $X = 0$ and $Z = 0$, then if we set $Y = 0$, we will transform to $\mathcal{A}_5$ rather than $\mathcal{A}_2$, which comes after $\mathcal{A}_4$ in the pre-order. So the formula $(X = 0, Z = 0) \boxright D(Y = 1)$ is not true in the model.

6 Conclusion

In this paper, we propose an interpretation to desire such that an agents’ desire of ϕ is evaluated by how much the outcome of making ϕ true is preferred by the agent. This requires us to take causality into consideration, as "the outcome of making ϕ true" is a concept that involves causal dependency. Therefore we embed preference into causal models to formalize this idea and construct an axiomatized system for this semantics. This system also takes conditional desire into account and is flexible for dynamic extensions in the future.

Acknowledgments. We owe our special thanks to Johan van Benthem, Maria Aloni and Fenrong Liu, for their extensive feedback and precious suggestions to our paper. We benefited greatly from discussions with other members of Tsinghua University–University of Amsterdam Joint Research Centre for Logic. We are also very grateful to the organizers and participants of the 9th Indian Conference on Logic and its Applications, in which we received a lot of inspiring comments to this work. In addition, let us express our sincere gratitude to the anonymous referees of the 8th International Conference on Logic, Rationality and Interaction.

References

1. Fausto Barbero, Katrin Schulz, Sonja Smets, Fernando R Velázquez-Quesada, and Kaibo Xie. Thinking about causation: A causal language with epistemic operators. In *International Workshop on Dynamic Logic*, pages 17–32. Springer, 2020.
2. Craig Boutilier. Conditional logics of normality: a modal approach. *Artificial intelligence*, 68(1):87–154, 1994.
3. Rachael Briggs. Interventionist counterfactuals. *Philosophical studies*, 160(1):139–166, 2012.
4. Didier Dubois, Emiliano Lorini, and Henri Prade. Revising desires–a possibility theory viewpoint. In *Flexible Query Answering Systems 2015*, pages 3–13. Springer, 2016.
5. Didier Dubois, Emiliano Lorini, and Henri Prade. The strength of desires: a logical approach. *Minds and Machines*, 27(1):199–231, 2017.
6. David Galles and Judea Pearl. An axiomatic characterisation of causal counterfactuals. *Foundations of Science*, 1:151–182, 1998.
7. Bart Geurts. Presuppositions and anaphors in attitude contexts. *Linguistics and philosophy*, 21(6):545–601, 1998.
8. Patrick Girard and Marcus Anthony Triplett. Ceteris paribus logic in counterfactual reasoning. *arXiv preprint arXiv:1606.07522*, 2016.
9. Joseph Y Halpern. Defining relative likelihood in partially-ordered preferential structures. *Journal of Artificial Intelligence Research*, 7:1–24, 1997.
10. Joseph Y Halpern. Axiomatizing causal reasoning. *Journal of Artificial Intelligence Research*, 12:317–337, 2000.
11. Sven Ove Hansson. A new semantical approach to the logic of preference. *Erkenntnis*, 31(1):1–42, 1989.
12. Irene Heim. Presupposition projection and the semantics of attitude verbs. *Journal of semantics*, 9(3):183–221, 1992.
13. Jérôme Lang. Conditional desires and utilities: an alternative logical approach to qualitative decision theory. In *ECAI*, pages 318–322. PITMAN, 1996.
14. Jérôme Lang, Leendert Van Der Torre, and Emil Weydert. Hidden uncertainty in the logical representation of desires. In *IJCAI*, pages 685–690, 2003.

15. David Lewis. Counterfactuals and comparative possibility. In *Ifs*, pages 57–85. Springer, 1973.
16. Fenrong Liu. *Reasoning about preference dynamics*, volume 354. Springer Science & Business Media, 2011.
17. Fenrong Liu et al. *Changing for the better: Preference dynamics and agent diversity*. ILLC, 2007.
18. Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.
19. Judea Pearl. Causality: models, reasoning, and inference. *IIE Transactions*, 34(6):583–589, 2002.
20. George F Schueler. *Desire: Its role in practical reason and the explanation of action*. MIT Press, 1995.
21. Robert Stalnaker. *Inquiry*. Cambridge University Press, 1984.
22. Melinda Vadas. Affective and non-affective desire. *Philosophy and Phenomenological Research*, 45(2):273–279, 1984.
23. Johan Van Benthem, Patrick Girard, and Olivier Roy. Everything else being equal: A modal logic for ceteris paribus preferences. *Journal of philosophical logic*, 38(1):83–125, 2009.
24. Johan Van Benthem and Fenrong Liu. Dynamic logic of preference upgrade. *Journal of Applied Non-Classical Logics*, 17(2):157–182, 2007.
25. Johan Van Benthem and Fenrong Liu. Deontic logic and changing preferences. In *Handbook of Philosophical Logic*, pages 1–49. Springer, 2018.
26. Johan Van Benthem, Sieuwert Van Otterloo, and Olivier Roy. Preference logic, conditionals and solution concepts in games. page 61–77, 2006.
27. Elisabeth Villalta. Spanish subjunctive clauses require ordered alternatives. In *Semantics and linguistic theory*, volume 10, pages 239–256, 2000.
28. Georg Henrik Von Wright. The logic of preference. *Philosophy*, 40(151), 1963.
29. Georg Henrik Von Wright. The logic of preference reconsidered. *Theory and Decision*, 3(2):140–169, 1972.