Semantics & Pragmatics Volume 13, Article 16: 1–50, 2020
https://doi.org/10.3765/sp.13.16

Intervention in deontic reasoning*

WooJin Chung
Institut Jean Nicod,
Département d'études cognitives,
ENS, EHESS, CNRS, PSL University

Submitted 2019-04-09 / First decision 2019-11-08 / Revision received 2020-03-04 / Second decision 2020-07-31 / Revision received 2020-08-17 / Accepted 2020-11-06 / Published 2020-11-09 / Final typesetting 2022-06-03

Abstract The 'if p, ought p' problem, famously known as Zvolenszky's puzzle (Zvolenszky 2002), questions whether possible world semantics can assign proper truth conditions to sentences of the form 'if p, ought p'. This paper suggests that it is not a problem of possible worlds semantics of modality, but rather, the 'if p, ought p' problem reveals the counterfactual nature of deontic modals which otherwise would have gone unnoticed. I propose that a counterfactual-based formulation of deontic necessity that implements intervention, jointly with the assumption that indicative conditionals facilitate backtracking, offers a principled solution to the 'if p, ought p' problem. I also present empirical evidence in favor of an interventionist approach to counterfactuals as opposed to similarity-based theories, at least in the domain of deontic reasoning.

Keywords: Zvolenszky's puzzle, modality, counterfactual, causal networks

1 Introduction

Suppose that Britney Spears has a contract with Pepsi, requiring that she does not drink non-Pepsi cola in public. In this situation, (1) is intuitively false.

* I gratefully acknowledge comments and suggestions from Sam Alxatib, Nadine Bade, Chris Barker, Justin Bledin, Lucas Champollion, Simon Charlow, Paul Egre, Semoon Hoe, Magdalena Kaufmann, Stefan Kaufmann, Songhee Kim, Jeremy Kuhn, Daniel Lassiter, Alda Mari, Salvador Mascarenhas, Seungho Nam, Paolo Santorio, Philippe Schlenker, Benjamin Spector, Anna Szabolcsi, and three anonymous referees. The work was funded by *Agence Nationale de la Recherche* grants ANR-17-EURE-0017 (*Frontiers of Cognition*) and ANR-18-CE28-0008 (*Between Language and Reasoning*).

(1) If Britney Spears drinks Coca-Cola in public, then she must drink Coca-Cola in public.

Zvolenszky claims that an unadorned version of Kratzer's modal semantics incorrectly predicts that the above example is true. In fact, any sentence of the form 'if *p*, must/should/ought *p*' comes out as true. Here is a brief summary of the problem: given the standard view of conditionals (Kratzer 1991a) and modality (Kratzer 1981b, 1991b), an *if* conditional restricts the quantificational domain of an independently-supplied modal operator. In (1), the antecedent of the *if* conditional restricts the quantificational domain of *must* in the consequent. Consequently, every world quantified over by *must* would be a world in which Britney Spears drinks Coca-Cola in public, and the prejacent 'she drinks Coca-Cola in public' would be trivially true in those worlds. The problem plagues the theory of deontic conditionals, and it has previously been acknowledged in the literature on deontic logic (van Fraassen 1972, Spohn 1975, Jackson 1985) and also in the linguistics literature (Frank 1997, Condoravdi 2002, Arregui 2011).

A prominent response to this problem is to invoke *the double modalization strategy*, where we add a covert epistemic necessity operator and let the *if*-clause restrict its domain rather than the overt deontic *must* (Kratzer 2012). In every epistemically accessible world in which Britney drinks Coca-Cola in public, is she obliged to drink Coca-Cola in public? Perhaps not.

Do we have a satisfying solution? Zvolenszky claims that there is a yet more dire issue which she calls *the flipside problem*. The following set of examples are all instances of the 'if *p*, ought *p*' construction, but unlike the Coca-Cola example in (1), they are intuitively true. For instance, considering that the Dalai Lama is extremely mild mannered, he does not, and should not get angry without a good reason. So if the Dalai Lama is angry, then he should be angry. Likewise, considering Yogi Bear's indolent disposition, if Yogi Bear works then he has to work. Zvolenszky argues that possible worlds semantics fails to correctly predict the truth conditions of the 'if *p*, ought *p*' sentences because the flipside examples are identical in structure to the Coca-Cola example.

(2) The flipside problem (Zvolenszky 2002)

a. If the Dalai Lama is angry, then he should be angry.
b. If Yogi Bear works then he has to work/is obliged to work.
c. If Bart Simpson listens to Bartók, then he must/is obliged to do so.

Zvolenszky's argument is not conclusive because not only the structure but also the world of evaluation affects the truth value of a sentence; the worlds at which 'she must drink Coca-Cola in public' is valued are different from the worlds at which 'he should be angry' is valued. One can respond to Zvolenszky in the following way: considering that the Dalai Lama only gets angry for a good reason, if he is angry then there must have been a good reason for him to be angry. Given such a circumstance, 'he should be angry' is intuitively true. By contrast, we do not infer that Britney had a good reason to drink Coca-Cola in public when she in fact did so. Thus, we are led to conclude that 'she must drink Coca-Cola in public' is false. Note also that the Dalai Lama example would be intuitively false if he was a normal mortal who sometimes gets angry for less than admirable reasons. We would not infer in this context that he had a good reason to be angry and thus judge the sentence as false. Zvolenszky overlooked the possibility that epistemic modals facilitate abductive inference, the result of which crucially affects the interpretation of deontic conditionals.

So far so good, but the devil is in the details. Even after granting that an additional layer of epistemic modality is required for the interpretation of 'if p, ought p', we cannot achieve explanatory adequacy for the following reason: one needs to strictly ignore certain causal dependencies in interpreting the deontic modal, despite the fact that such dependencies are evidently relevant. The standard account does not offer a good explanation of why the dependencies have to be ignored and why there is no reading in which they are considered as relevant.

My take on this issue is that the double modalization strategy is not the source of the problem. Rather, the 'if p, ought p' problem reveals the counterfactual nature of deontic modals that would otherwise have gone unnoticed. In this paper, I present a solution to the problem which maintains the view that the *if*-clause in 'if p, ought p' restricts a covert epistemic necessity operator. The innovation is in the semantics of deontic modals: I propose that the interpretation of deontic modals requires counterfactual reasoning, in the sense that the set of relevant worlds for the interpretation of deontic modals is restricted to the counterfactual prejacent-worlds and the counterfactual alternative-to-the-prejacent-worlds. Furthermore, I propose that counterfactual selection is based on manipulation of causal networks models (Pearl 2000), as opposed to similarity of worlds. It involves *intervention* — an operation that prevents lumping of causally relevant propositions. The interplay of abductive reasoning — due to the covert epistemic modal — and

causal reasoning — due to the counterfactual semantics inherent to *ought* — is the key to assigning proper truth conditions to 'if p, ought p' sentences.

This paper is organized as follows: Section 2 presents the standard account of modality due to Angelika Kratzer, and Section 3 introduces the technical details of Zvolenszky's arguments as well as the double modalization strategy. Section 4 takes a closer look at the double modalization strategy and offers a step-by-step analysis of how the *if*-clause in 'if p, ought p' influences the construction of the deontic accessibility relation. Section 5 points out an issue that concerns the construction of the deontic accessibility relation elaborated in Section 4. I also present a widely known overgeneration problem. Section 6 proposes a counterfactual-based semantics of *ought*, on which my analysis of the 'if p, ought p' problem is based. Section 7 fleshes out the analysis, and Section 8 presents empirical arguments in favor of an interventionist approach to counterfactuals. Section 9 compares the proposed analysis with Carr's (2014) theory of deontic conditionals.

2 Preliminary: the standard account of modality

Before we proceed to the main discussion, I will briefly introduce the Kratzerian view of deontic necessity (Kratzer 1981b, 1991b). Kratzer proposes that modal expressions are interpreted with respect to two conversational backgrounds that jointly determine the accessible worlds: a modal base and an ordering source (typically labeled as f and g, respectively). The modal base takes a world and returns the set of propositions that are relevant to the evaluation of the modal, and its intersection yields a set of relevant worlds. I will refer to such a set as *the modal background*. The ordering source takes a world and returns the set of propositions that is used to construct a preorder, which imposes an ordering on the modal background in the following way: $u \leq_{g(w)} v$ (informally read as 'u is at least as good as v') iff the set of propositions in $g(w)$ that are true in v is a subset of the set of propositions in $g(w)$ that are true in u.

(3) Preorder $\leq_{g(w)}$ with respect to $g(w)$

For all $u, v \in W$, $u \leq_{g(w)} v$
iff $\{p : p \in g(w) \wedge p(v) = 1\} \subseteq \{p : p \in g(w) \wedge p(u) = 1\}$

Given the set of relevant worlds and the preorder, the definition of modal necessity in (4) amounts to saying that for every sequence of relevant worlds

in the modal background ordered with respect to the preorder, there is a point at which the prejacent p is true in all worlds that are at least as good.

(4) Modal necessity (Kratzer 1991b)[1]
A proposition p is a necessity in a world w with respect to a modal base f and an ordering source g if and only if the following condition is satisfied:

> For all $u \in \cap f(w)$ there is a $v \in \cap f(w)$ such that $v \leq_{g(w)} u$ and for all $z \in \cap f(w)$: if $z \leq_{g(w)} v$, then $z \in p$.

We can simplify the above formulation by making the Limit Assumption (Kaufmann 2017), the assumption that we can always identify *the best worlds* with respect to $\leq_{g(w)}$. The best worlds can be represented with the BEST operator (Portner 2009); it takes a modal background (a set of worlds) determined by the modal base and an ordering source (a set of propositions), and returns the maximal subset of the former such that no relevant world is strictly better than any member of this maximal subset.

(5) The BEST operator (Portner 2009)
The BEST operator takes a set of worlds M and a set of propositions O (notation: $\text{BEST}(M)(O)$), and returns the set of M-worlds that are best-ranked according to $\leq_O$.

Under the Limit Assumption, modal necessity reduces to the necessary truth of the prejacent in the best worlds. The flavor of the modal is determined by the choice of the modal base and the ordering source.

(6) Modal necessity under the Limit Assumption
$[\![\text{must}_{Kratzer}]\!]^w = \lambda p.\forall w' \in \text{BEST}(\cap f(w))(g(w)) : p(w') = 1$,
where f is a modal base and g is an ordering source

The deontic interpretation of *must* involves *a circumstantial modal base* and *a deontic ordering source*. The circumstantial modal base takes a world of evaluation and returns the set of propositions that corresponds to the relevant circumstances of the world. Intersecting the propositions yields a cir-

1 This version of modal necessity does not distinguish strong necessity modals such as *must* from weak necessity ones, such as *ought* or *should*. While many have reported that the two types of necessity modals differ in various aspects, I will not be concerned with the differences in this paper since the issues at hand affect the interpretation of both types.

cumstantial modal background, the members of which are circumstantially accessible worlds. The deontic ordering source takes a world of evaluation and returns the set of propositions that characterizes the ideals of the world. The BEST operator returns the set of circumstantially accessible worlds that are best-ranked with respect to the deontic ordering source. The resulting set of worlds serves as the domain of quantification. I will call such worlds *the deontically best worlds*. Deontic necessity asserts the necessary truth of the prejacent in the deontically best worlds.

The following section is dedicated to introducing the technical details of Zvolenszky's (2002) criticisms as well as a prominent response to the criticisms, namely the double modalization strategy. At first glance, the prominent response seems to dispel Zvolenszky's worries. However, in sections to follow, I will point out yet another issue, particularly due to the underspecification of how the modal background of a deontic modal should be determined.

3 Zvolenszky's puzzle

Let us return to the Coca-Cola example in (1), repeated below as (7), and take a closer look at the problem.

(7) If Britney Spears drinks Coca-Cola in public, then she must drink Coca-Cola in public.

A widely accepted view of conditionals is the restrictor analysis (Kratzer 1991a). In this view, conditionals restrict a given domain of quantification. This means that in (7), the conditional restricts the circumstantial modal base supplied to *must*. Combining the deontic interpretation of (6) with the restrictor view yields the semantics in (8). The formula reads as follows: *"In every deontically best world in which Britney Spears drinks Coca-Cola in public, the prejacent* 'she drinks Coca-Cola in public' *is true."*

(8) $[\![(7)]\!]^w = \forall w' \in \underbrace{\text{BEST}(\cap(c(w) \cup \{\textbf{coke}\})(d(w))}_{\textit{deontically best } \textbf{coke}\textit{-worlds of } w} : \textbf{coke}(w') = 1,$

where c is a circumstantial modal base, and d is a deontic ordering source

This is trivially true because the quantificational domain of $must_{Kratzer}$ is restricted to only those worlds where Britney Spears drinks Coca-Cola in public.

The problem arises because the conditional is restricting the modal domain in such a way that the antecedent proposition is guaranteed to be true in all of the modal worlds.

A possible solution to this problem is to invoke the double modalization strategy (Geurts 2004). Specifically, conditionals are not necessarily pure restrictors but can introduce their own necessity modal operator — an epistemic necessity operator for indicative conditionals. The antecedent of the conditional restricts the modal base of this covert modal operator. Just like deontic necessity, epistemic necessity can be formulated in terms of the BEST operator but it involves different conversational backgrounds: *an epistemic modal base* and *a stereotypical ordering source*. The epistemic modal base determines the epistemically accessible worlds, and the stereotypical ordering source is used to pick out the worlds that are in accordance with the normal course of events. I will refer to the selected worlds as *the epistemically best worlds*.

I will use 'if$_{ind}$' as a shorthand for a covert epistemic necessity modal restricted by an *if*-clause. For simplicity, I will treat it as an operator with the following semantics:

(9) Indicative conditionals
$[\![\text{if}_{ind}]\!]^w = \lambda p \lambda q. \forall w' \in \underbrace{\text{BEST}(\cap(e(w) \cup \{p\}))(s(w))}_{\textit{epistemically best } p\textit{-worlds of } w} : q(w') = 1,$

where e is an epistemic modal base and s is a stereotypical ordering source

The standard account coupled with the double modalization strategy, hereafter referred to as *the prominent response*, yields the truth conditions in (10) for the Coca-Cola example in (7). The formula reads as follows: *"In every epistemically best world in which Britney Spears drinks Coca-Cola in public, 'she must drink Coca-Cola in public' is true."*

(10) Double modal reading
$[\![(7)]\!]^w = \forall w' \in \underbrace{\text{BEST}(\cap(e(w) \cup \{\textbf{coke}\}))(s(w))}_{\textit{epistemically best } \textbf{coke}\textit{-worlds of } w} :$
$\forall w'' \in \underbrace{\text{BEST}(\cap c(w'))(d(w'))}_{\textit{deontically best worlds of } w'} : \textbf{coke}(w'') = 1,$

where c is a circumstantial modal base, d is a deontic ordering source, e is an epistemic modal base, and s is a stereotypical ordering source

There are two BEST operators involved in (10). One is introduced by the indicative conditional and the other is supplied by the deontic *must*. We can now interpret (10) in the following way:

(11) A step-by-step interpretation of (10)
   a. Identify the epistemically best worlds of w in which Britney Spears drinks Coca-Cola in public.
   b. For each of the epistemically best worlds w', locate the deontically best worlds of *w'*.
   c. Assert that the prejacent 'she drinks Coca-Cola in public' is true in those deontically best worlds.

Zvolenszky argues that the prominent response crucially relies on the assumption that the antecedent of the *if* conditional, **coke**, is *circumstantially irrelevant* for the interpretation of the deontic modal (i.e., it needs to be the case that $\textbf{coke} \notin c(w')$). Otherwise, the sentence would again be rendered trivially true. Zvolenszky claims that this type of approach predicts that the antecedent proposition p in 'if p, ought p' is *never relevant* to the interpretation of *ought* in the consequent. She argues that there are cases in which we do have to consider the antecedent proposition in evaluating the deontic modal in the consequent. The flipside examples in (2), repeated below as (12), are such cases.

(12)
   a. If the Dalai Lama is angry, then he should be angry.
   b. If Yogi Bear works then he has to work.
   c. If Bart Simpson listens to Bartók, then he must do so.

Consider the Dalai Lama example in (12a). There is a sense in which the sentence is true, but not in a trivial sense: considering the fact that the Dalai Lama is extremely mild mannered, if he is angry, then he has a good reason to be angry, and in fact in such a situation he should be angry. Clearly, the sentence is not trivially true. So the single modal reading, provided in (13), is not the right way to go.

(13) Single modal reading

$$[\![(12a)]\!]^w = \forall w' \in \underbrace{\text{BEST}(\cap(c(w) \cup \{\textbf{angry}\}))(d(w))}_{\textit{deontically best } \textbf{angry}\textit{-worlds of } w} : \textbf{angry}(w') = 1,$$

where c is a circumstantial modal base, and d is a deontic ordering source

Zvolenszky dismisses the prominent response because from her point of view (which I will oppose later in this section), we cannot conclude that the deontically best worlds of w' are necessarily **angry**-worlds unless the conditional antecedent (i.e., **angry**) is a relevant circumstance at w'. Zvolenszky's claim is that the role of double modalization in analyzing 'if p, ought p' is to render the conditional antecedent irrelevant for the interpretation of the deontic modal in the consequent, and therefore the prominent response predicts the Dalai Lama example to be false.

(14) Double modal reading

$$[\![(12a)]\!]^w = \forall w' \in \underbrace{\text{BEST}(\cap(e(w) \cup \{\textbf{angry}\}))(s(w))}_{\textit{epistemically best } \textbf{angry}\textit{-worlds of } w} : \forall w'' \in \underbrace{\text{BEST}(\cap c(w'))(d(w'))}_{\textit{deontically best worlds of } w'} : \textbf{angry}(w'') = 1,$$

where c is a circumstantial modal base, d is a deontic ordering source, e is an epistemic modal base, and s is a stereotypical ordering source

Zvolenszky also considers a possible alternative to the double modalization strategy: Frank's (1997) *context reduction* — which Zvolenszky refers to as *expansion* — expands the modal background to contain both the prejacent-worlds and the negation-of-the-prejacent-worlds (see also Condoravdi 2002 and Werner 2006 for related discussion). Context reduction is originally formulated in a DRT framework, and below is Zvolenszky's restatement of context reduction in premise semantics:

(15) Zvolenszky's restatement of context reduction in premise semantics
For any sentence p, world w, ordering source O, and modal bases M and $M^{p/\neg p}$, 'it must be that p' is true in w relative to M and O iff p is true in all the worlds closest (by O) to w within $M^{p/\neg p}$, where $M^{p/\neg p}$ is the result of expanding M with respect to both p and its negation; this amounts to removing restrictions on M, if any, with respect to p or p's negation; $M^{p/\neg p}$ thus leaves p open.

Zvolenszky argues that context reduction is subject to the very same problem as the double modalization strategy because the conditional antecedent p in 'if p, ought p' is effectively disregarded in the interpretation of *ought*.[2] Zvolenszky also criticizes context reduction for being an ad hoc strat-

2 As an anonymous reviewer notes, this is not entirely precise. Contra Zvolenszky's argument, it is not the case that context reduction is ignoring p since it is being used to expand M with

egy, which only exists to circumvent the 'if p, ought p' problem. Zvolenszky concludes that possible worlds semantics of modals inevitably fails because it cannot make systematic predictions for the 'if p, ought p' examples.

I am not fully convinced that the double modalization strategy fails. As noted earlier, the deontic modals in the two examples are evaluated at different worlds and I suggest that this is responsible for the diverging truth values. Particularly, considering that the Dalai Lama is extremely mild mannered, the epistemically best worlds in which the Dalai Lama is angry are worlds where he has a good reason to be angry. Given that he has a good reason, he should be angry.

Although I agree that the 'if p, ought p' examples involve an additional layer of epistemic necessity, I claim that the standard modal semantics still needs to rely on ad hoc stipulations to derive the desired reading. In what follows, I will first argue that in the salient context where the Dalai Lama only gets angry for a good reason, the correct way to construct a deontic accessibility relation is to restrict the modal base of *should* to worlds in which whatever justifies the Dalai Lama's anger has taken place. I will then point out that constructing such a deontic accessibility relation requires making unnatural assumptions about what counts as circumstantially relevant for the interpretation of the deontic modal.

4 Indicative conditionals and the composition of deontic accessibility relations

In Kratzer's system, the deontically best worlds are jointly identified by a circumstantial modal base and a deontic ordering source. So in order to construct a deontic accessibility relation that renders 'he should be angry' true, we can manipulate the modal base or the ordering source.[3] There are three possible strategies, which are listed below:

respect to both p and its negation. Rather, what is lacking in context reduction is how exactly expansion is carried out. My proposal in the following sections can be understood as fleshing out the exact process of context reduction: it requires counterfactual reasoning, particularly involving intervention.

3 An anonymous reviewer points out that the modal base of the outer modal restricted by the *if*-clause needn't always be epistemic. I agree that this is a possibility, but as far as the composition of the deontic accessibility relation is concerned, the outer modal — whether its modal base is epistemic or not — can only manipulate the deontic accessibility relation in the ways specified in (16).

(16) 'The Dalai Lama should be angry' is rendered true in w' if…

a. Reproduction of single modal reading:
angry $\in c(w')$
"The Dalai Lama being angry is circumstantially relevant."

b. Idealizing unconditional anger:
angry $\in d(w')$
"The Dalai Lama being angry is an ideal."

c. Idealizing conditional anger:
trigger $\in c(w')$ and **trigger** $\rightarrow$ **angry** $\in d(w')$
"It is an ideal that the Dalai Lama is angry in the presence of a sufficiently reprehensible event (i.e., **trigger**), and the fact that such an event occurred is circumstantially relevant."

I am in favor of the third strategy (16c), but let us take a look at each of the available options. The first strategy (16a), which is to render **angry** circumstantially relevant, should be immediately rejected because it yields trivial truth; if all relevant worlds were **angry**-worlds, the prediction of the double modal reading would be identical to the single modal reading.

The second strategy (16b) dictates that the Dalai Lama being angry is an ideal of w'. This amounts to saying that one of the ideals in w' is that the Dalai Lama is angry *unconditionally*, and he should be angry even without any justification. I doubt that this is the right characterization of the worlds identified from the hypothetical assumption that the Dalai Lama is angry. What aspect of epistemic reasoning makes us shift the ideals in a way that the Dalai Lama being unconditionally angry is a virtue?

The last strategy (16c) suggests that the indicative conditional affects the composition of the modal base. It is not the case that the Dalai Lama should be angry no matter what, but his anger is justified given certain circumstances. As Zvolenszky notes, we take into account that "the Dalai Lama does not get angry unless he has a very good reason for doing so". So upon interpreting the indicative conditional of (12a), we reason that if the Dalai Lama is angry, there must have been a very good reason for him to be angry. Whatever that reason would be, when identifying the set of the epistemically best **angry**-worlds, we locate ourselves in a world in which a sufficiently reprehensible event that justifies the Dalai Lama's anger occurred. I will represent the occurrence of such an event as **trigger**. Should the Dalai Lama be angry in a world where **trigger** is true? Intuitively yes. Thus, if the Dalai Lama is angry, he should be angry.

The abductive inference from **angry** to **trigger** is an instance of what has been referred to as *backtracking* in the literature on conditionals (Lewis 1979, Pearl 2000, Bennett 2003, Arregui 2005, Kaufmann 2013, Lassiter 2017, among many others). While it is debatable whether counterfactuals validate such an inference, the availability of backtracking in indicative conditionals is less controversial.

By contrast, Britney Spears is not obliged to drink Coca-Cola no matter what, and in fact, no contextually salient circumstance justifies her Coca-Cola-drinking. For this reason, in the epistemically best worlds in which Britney Spears drinks Coca-Cola in public, she is not obliged to do so. Note also that if we suppose that Britney Spears is extremely disciplined and she never breaches a contract without a good reason, we would judge 'if Britney Spears drinks Coca-Cola in public, then she must drink Coca-Cola in public' differently. To summarize, although the Dalai Lama example and the Coca-Cola example are identical in structure, the two crucially differ in whether the indicative conditional identifies a circumstance that justifies the Dalai Lama's anger or Britney's Coca-Cola-drinking. For the Coca-Cola example, we cannot identify such a circumstance.

Does the consideration of backtracking of indicative conditionals clear away the doubts on the prominent response? There are two issues that need to be addressed. First, the prominent response overgenerates, as the single modal reading of the 'if p, ought p' examples is not available. Second, the construction of the circumstantial modal base that would derive the desired deontic accessibility relation requires one to strictly ignore certain causal dependencies that are salient in the context.

5 Two issues with the prominent response

5.1 The prominent response overgenerates

It is well-known that the prominent response overgenerates, as acknowledged in Kratzer (2012). Specifically, the single modal reading, which would render both the Coca-Cola example and the Dalai Lama example trivially true, does not arise; recall that we do not judge the Coca-Cola example as true given Britney's contract with Pepsi, or the Dalai Lama example as trivially true. So the proponents of this line of thought would have to stipulate that the single modal reading (i.e., (8) and (13)) is categorically unavailable for sentences of the from 'if p, ought p', while the most natural reading of other

deontic conditionals is the single modal reading. The double modalization strategy is invoked because the single modal reading yields trival truth and that needs to be avoided.

It would be preferable to have a theory that does not resort to such a stipulation. Moreover, there is no reason in principle to avoid a trivial interpretation. Kratzer (2012) notes that she has no problem perceiving a trivial reading of (17), although a non-trivial reading of the sentence is simultaneously available. It then is puzzling that a single modal construal is categorically unavailable for the 'if p, ought p' examples.

(17) I could not possibly work more than I do.

I will show that the semantics I propose in Section 6 blocks the single modal reading of 'if p, ought p' because it yields an inconsistent modal base. The very same semantics also does not require us to make unnatural assumptions about what counts as circumstantially relevant, which is the topic of the following section.

5.2 The prominent response is forced to ignore certain factual dependencies

Let us take a closer look at how the prominent response builds the deontic accessibility relation. For the Dalai Lama example, the indicative conditional quantifies over the epistemically best **angry**-worlds, which are **trigger**-worlds. The indicative conditional affects the deontic accessibility relation in a way that **trigger** is circumstantially relevant for the interpretation of *should* embedded in the conditional. Granted that **trigger** is circumstantially relevant, we must ask ourselves whether **angry** — the proposition that led us to infer that **trigger** is true — is a relevant circumstance. Proponents of the prominent response would definitely not want to say that it is circumstantially relevant, because if it were, they would predict that Zvolenszky's examples are all trivially true. For the Dalai Lama example, if the modal background of *should* were confined to **angry**-worlds, the deontically best worlds would all be **angry**-worlds irrespective of one's ideals. Thus, 'if the Dalai Lama is angry, he should be **angry**' would be trivially true.

To summarize, the preferred setup is that **trigger** is circumstantially relevant but **angry** is not (i.e., **trigger** $\in c(w')$, **angry** $\notin c(w')$). But what justifies hand-tailoring the circumstantial modal base in this manner? What underlies the abductive inference from **angry** to **trigger** is that given the Dalai Lama's

mild disposition, he would be angry if **trigger** was true, but he would not be if **trigger** was false. Despite the fact that the truth value of **trigger** determines the truth value of **angry**, we are required to assume that the former is circumstantially relevant but the latter is not. The dependency between **trigger** and **angry** is ignored.

It is important to note that the way in which the prominent response picks up relevant facts cannot be explained in terms of the flexibility of deontic modals in deciding what are the relevant circumstances (see section 8.2 for details). In other words, we cannot say that the Dalai Lama example is true because *should* has an option to ignore the salient dependency between **trigger** and **angry**. The problem is that ignoring such a dependency *is the only way to interpret the sentence*. There is no interpretation of the sentence where the dependency between **trigger** and **angry** is circumstantially relevant. Had the dependency been relevant, the sentence would have been either trivially true or false depending on whether **trigger** is circumstantially relevant or not: If **trigger** were considered as relevant, it would also bring in **angry** and this would yield trivial truth. On the other hand, if **trigger** were regarded as irrelevant, there would be no relevant circumstance that justifies the Dalai Lama's anger. But none of the two readings are available, and the only interpretation we get is that the Dalai Lama is rightfully angry in the epistemically best **angry**-worlds. Thus, we are required to pick up facts in a very specific way that we disregard the salient dependency between **trigger** and **angry**, although this information played a key role in identifying the epistemically best **angry**-worlds. Why is it the case that the dependency between **trigger** and **angry** is *always* ignored?

5.3 If certain facts always come and go together, Death will always be there for you

A similar point has been made by Carr (2014). Consider the following scenario originally due to Gibbard & Harper (1978). Given that Death will make a very good prediction of your whereabouts, (18a) and (18b) are both true.

(18) Self-frustrating *ought* (Carr 2014)
If you are in the same city as Death tomorrow, then you'll die. Death has planned to be wherever he predicts you'll be, and he's very reliable in such predictions. Your options are to stay in Damascus or to go to Aleppo. But, as you know, if you stay in Damascus, then that's

excellent evidence that Death will already be there. Similarly for going to Aleppo.

a. If you go to Aleppo, you ought not to go to Aleppo (because Death will be there).
b. If you stay in Damascus, you ought not to stay in Damascus (because Death will be there).

The above examples are of the form 'if p, ought $\neg p$' unlike Zvolenszky's examples, but they raise the same issue. First, the *if*-clause cannot directly restrict *ought* because it yields contradiction. Carr argues that the double modalization strategy is not without a problem either: upon interpreting 'If you go to Aleppo' (henceforth **you-alp**), we infer that Death is in Aleppo (henceforth **death-alp**) in the epistemically best worlds where you go to Aleppo. Carr notes that if Death were in Aleppo in every circumstantially relevant world but your location were undetermined (i.e., you could either be in Aleppo or Damascus), the prominent response would predict 'if you go to Aleppo, you ought not to go to Aleppo' as true. However, Carr claims that such a quantificational domain can only be constructed by ad hoc stipulation, and it might not even be a coherent configuration of background assumptions.

> "What is the justification for holding Death's location fixed throughout the modal background? Presumably the information that Death will be in Aleppo is an inference from (a) the information from the antecedent that you will go to Aleppo and (b) the background information that Death will be where you are. But while the modal background reflects the conclusion from these two premises, it doesn't reflect the premises themselves: the modal background doesn't reflect premise (b). It allows that you might be in a different place from Death, since it includes both **Aa**-worlds and **Ad**-worlds.[4]
>
> In other words, speakers must assume *in the very same breath* that Death must be in the same place as you and that he might not be. How is it possible for context to fix the sets of relevant circumstances in this way? It's not just that the stipulation of

4 **A**, **a**, and **d** refer to 'Death is in Aleppo', 'you are in Aleppo', and 'you are in Damascus', respectively.

> the modal background is ad hoc: *it's not clear that this is even a coherent configuration of background assumptions.*"
>
> [Carr 2014: 574–575]

Carr's argumentation is similar in structure to mine concerning the Dalai Lama example. I argued that if **trigger** is circumstantially relevant for the interpretation of *should* in the epistemically best **angry**-worlds, then **angry** has to be circumstantially relevant as well because the indicative *if*-clause 'if the Dalai Lama is angry' identifies **trigger** based on the assumption that the truth value of **trigger** determines the truth value of **angry**. Carr claims that upon interpreting the indicative *if*-clause 'if you go to Aleppo', we infer that the epistemically best **you-alp**-worlds are **death-alp**-worlds based on the assumption that Death will be where you are. So in the epistemically best **you-alp**-worlds, if **you-alp** is circumstantially irrelevant for the interpretation of *ought*, then **death-alp** has to be circumstantially irrelevant as well. Carr's point is that we cannot cherry-pick and fix Death's location while moving around your location, because the two propositions are tightly correlated.

Carr concludes that Kratzer's modal semantics in its current form cannot make correct predictions for the self-frustrating *ought* examples. She proposes that the deontic ordering source has to be information sensitive, in the sense that the ordering between worlds can change given more information. While I agree that the self-frustrating *ought* scenario calls for an amendment to Kratzer's modal semantics, I will take a different path that justifies the construction of what Carr claims to be an ad hoc modal background.

The research objective is to motivate the construction of a modal background where (i) **trigger** is necessarily true but **angry** an open possibility for the Dalai Lama example, and (ii) **death-alp** is necessarily true but **you-alp** is an open possibility for the self-frustrating *ought* example. My idea is to use counterfactual semantics to construct the modal background. Some theories of counterfactuals guarantee that the dependency between **angry** (**death-alp**) and **trigger** (**you-alp**) is ignored upon making a counterfactual assumption about the truth of **angry**.

6 Proposal

In this section, I offer a counterfactual-based semantics of *ought* that permits an alteration of causal dependencies between facts. I will first provide a rough sketch of the proposal which abstracts over the details of counterfactual semantics, and work out the full version in section 6.2. I will tentatively use a selection function f and represent the counterfactual p-worlds of w as $f(w, p)$.

I suggest that the modal background of *ought* is constructed by (i) entertaining the counterfactual prejacent-worlds and each of the counterfactual alternative-to-the-prejacent-worlds and (ii) performing a union operation on the set of entertained counterfactual worlds. I will refer to the outcome of the union operation as the counterfactual testing ground of the prejacent:

(19) Counterfactual testing ground
The counterfactual testing ground of p at w is defined as follows:

$$\mathbf{cfg}_w(p) = \cup\{f(w, r) \mid r \in Alt(p)\},$$

where $f(w, p)$ selects the counterfactual p-worlds of w and $Alt(p)$ is the set of alternatives to p (which includes p)

I propose that 'ought p' conveys that there exists a counterfactual p-world which is a good world, and for each $r \in Alt(p)$ such that $r \neq p$ is it the case that no counterfactual r-world is a good world. The set of good worlds is defined as the set of deontically best worlds in the counterfactual testing ground.[5]

5 Not only the semantics in (20) makes it explicit that reasoning with counterfactuals affects the construction of the modal background, but also there is cross-linguistic evidence in favor of the existential-based formulation. In Korean, 'ought p' effectively translates to 'only if p, good', as exemplified below:

(i) John-un sakwa-lul mek-eya toy-n-ta.
John-TOP apple-ACC eat-only.if GOOD-PRES-DECL
'(Lit.) Only if John (were to) eat an apple, good.'
'John ought to eat an apple.'

Chung (2019) shows that the Korean monomorphemic *-(e)ya* 'only if' is ambiguous between an existential reading (i.e., the exhaustifier negates existential quantification) and a universal reading (i.e., the exhaustifier negates universal quantification) and that the existential reading is the default interpretation. Accordingly, Chung proposes that the compositional semantics of the above construction involves existential quantification over each counter-

(20) The semantics of *ought* (preliminary)

$$[\![\text{ought } p]\!]^w = \exists w' \in f(w,p) : \underbrace{\text{BEST}(\mathbf{cfg}_w(p))(d(w))(w') = 1}_{\textit{good worlds of } w}$$
$$\wedge\ \forall r \in Alt(p) \text{ s.t. } r \neq p :$$
$$\neg \exists w' \in f(w,r) : \underbrace{\text{BEST}(\mathbf{cfg}_w(p))(d(w))(w') = 1}_{\textit{good worlds of } w}$$

The remainder of Section 6 is dedicated to precisely defining the counterfactual selection function f, which is used to construct the counterfactual testing ground. In what follows, I introduce a notational variant of Santorio's (2019) filtering semantics which implements Pearl-style causal networks (Pearl 2000) within premise semantics. Theories based on causal networks hypothesize that counterfactuals directly encode causal dependencies, and constructing the quantificational domain of a counterfactual does not require paying maximal attention to the facts of the world, contra the standard similarity-based accounts. Also, these theories typically implement *intervention*, an operation that removes certain causal dependencies between facts. I will use filtering semantics to formally define the selection function f.

I would like to note that my analysis of the 'if p, ought p' problem does not crucially hinge on the endorsement of filtering semantics. Premise semantics can replicate the results by imposing a particular restriction on the totally realistic ordering source. Despite the fact that some versions of premise semantics are well-suited for the current purpose, I will use filtering semantics for the demonstration because there are independent reasons to opt for filtering semantics. Section 8 elaborates on this topic.

6.1 An interventionist view of counterfactuals

Inspired by the pioneering work of Pearl (2000), there have been various attempts among linguists and philosophers to directly encode causal relations into possible worlds semantics of counterfactuals (Hiddleston 2005, Schulz 2011, Briggs 2012, Kaufmann 2013, Santorio 2019, Ciardelli, Zhang & Champollion 2018). While the implementations vary, they all agree in two aspects. First, the theories share the assumption that the truth of a counterfactual is determined by how a set of causally relevant variables (formally represented as partitions of worlds) are linked and what their values are. Second, they all

factual alternative-worlds unless the context facilitates the universal reading of *-(e)ya* 'only if'.

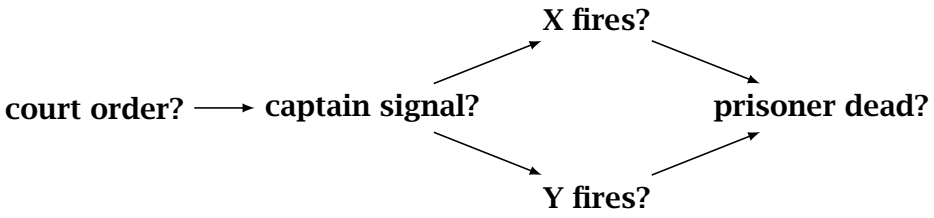
implement *intervention*, an operation that removes certain links between the causal variables.

I will first illustrate how the system works via a well-known firing squad example from Pearl (2000). The description of the causal dependencies is cited from Santorio (2019).

(21) The firing squad
A firing squad is positioned to execute a prisoner. The squad is waiting for a court order. The court issuing the execution order will result in the captain sending a signal to the two members of the squad, X and Y, who will fire and kill the prisoner. The court not issuing the order will result in the captain not sending the signal, the two riflemen not shooting, and the prisoner remaining alive.

The scenario manifests that (i) the court order causally affects the captain sending a signal, (ii) the captain's signal causally affects firing of the squad members X and Y, and (iii) X and Y's firing both causally affects whether the prisoner dies or remains alive. The directionality of the causal dependencies can be represented using a directed graph, as in (22).

(22) Causal dependencies in the firing squad scenario

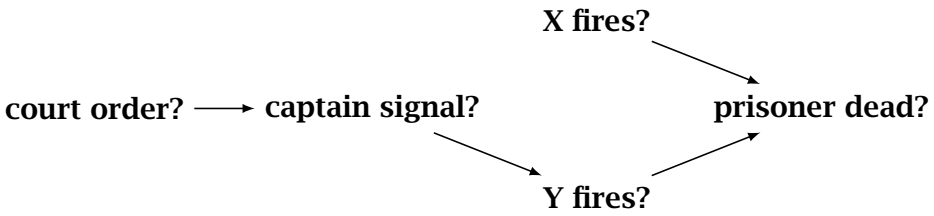


The value of each *endogenous* causal variable — a variable that is causally downstream from some causal variable — is deterministically computed from the causal graph and the values of *exogenous* causal variables — variables that are not causally downstream from others. In the above graph, there is only one exogenous causal variable, **court order?**, and all others are endogenous causal variables. Let us suppose that the court did not issue the order and fix the value of **court order?**. Then the captain did not send a signal, no squad member fired, and the prisoner is alive. Now, consider the example in (23). How does the counterfactual antecedent affect the causal graph, and how do causal networks-based theories interpret the counterfactual?

(23) If X had fired, the prisoner would have died.

Upon supposing that X had fired, we ignore the causal dependency between **captain signal?** and **X fires?**. In the graph representation, this amounts to removing the edge between the two corresponding nodes. The resulting causal graph is given in (24).

(24) Intervention on 'X fires?'



The operation just illustrated, which removes the causal dependency between the node that questions the truth of the counterfactual antecedent and the node that has a direct causal influence on it, is dubbed *intervention*. After intervening on **X fires?**, we reason whether the consequent is true in the revised setting: The court still did not issue the order, so the captain did not send a signal. Y did not fire accordingly, but since we are supposing that X fired, the prisoner is dead.

One of the notable differences between causal networks-based theories of counterfactuals and Kratzer's original premise semantics for counterfactuals is that the former does not consider every fact of the evaluation world in identifying the domain of quantification. Instead, only the values of causally relevant variables and how they are causally linked matter. In identifying the quantificational domain of the counterfactual in the firing squad example, we only check whether the court issued the execution order, whether the captain sent a signal, whether the squad members fired, and how these facts are dependent on each other.

It is noteworthy to mention that intervening on a causal variable only resets the values of causal variables that are *causally downstream* from the intervened variable, and others remain intact. In the firing squad scenario, only **prisoner dead?** is causally downstream from **X fires?**, so the values of **court order?**, **captain signal?**, and **Y fires?** remain intact. As a heads-up, Section 7 demonstrates that this aspect of intervention is what allows us to ignore the problematic factual dependencies in the Dalai Lama example and the self-frustrating *ought* examples.

There are a number of implementations of intervention within premise semantics, including Kaufmann 2013, Santorio 2019, and Ciardelli, Zhang & Champollion 2018. Among them, I introduce a notational variant of Santorio's filtering semantics, which I adopt in the remainder of this paper. The version I present effectively produces the same outcome but separately manages the value of the intervened variable and other information.

Santorio presents two versions of filtering semantics, the basic version and the refined version. The two versions differ primarily in whether they can make the right prediction for counterfactuals with disjunctive antecedents. Keeping the filtering mechanism intact, the refined version offers a systematic way of deciding which causal variable to intervene on. Since none of the problematic examples under consideration involves disjunctive antecedents, I will introduce the notational variant of the basic version for brevity.

In filtering semantics, the information about causal dependencies is encoded into the ordering source — call it *a causal ordering source* — and the directionality in causal dependencies is captured by representing members of the causal ordering source as pairs of a question denotation (partition of worlds) and a proposition, rather than as propositions. The question denotation represents the possible values of a single causal variable, and the other member of the pair, the proposition, describes the conditions under which the values are obtained. For example, the causal dependency between **court order?** and **captain signal?** is represented as in (25).

(25) Filtering semantics representation of '**court order?** → **captain signal?**'
⟨{**signal**, ¬**signal**}, **signal** ↔ **court-order**⟩

Exogenous variables, the values of which are required to compute the values of endogenous variables, are also members of the causal ordering source. The question denotation of an exogenous variable represents its possible values, and the other member of the pair, the proposition, is the value of the variable in the evaluation world. For example, **court order?** in the firing squad scenario is represented as follows, assuming that the court did not issue an order:

(26) Filtering semantics representation of **court order?**
⟨{**court-order**, ¬**court-order**}, ¬**court-order**⟩

The causal ordering source consists of the information about causal dependencies and the values of exogenous variables. I implement intervention

as a 'filtering operation', in the sense that any information representing a direct causal influence from one causal variable to the one that questions the truth of the counterfactual antecedent is removed from the causal ordering source. Given the causal ordering source where its members are represented as a question-proposition pair, this amounts to removing a pair containing a question that is settled by the counterfactual antecedent.[6]

As an illustration, consider again the example in (23). The causal ordering source before filtering is given in (27a). Among its members, the question of ⟨{**x-fire**, ¬**x-fire**}, **x-fire** ↔ **signal**⟩ is settled by the counterfactual antecedent **x-fire**, so it is filtered out. The filtered pair will be struck out as in (27b).

(27) a. Causal ordering source
⟨{**court-order**, ¬**court-order**}, ¬**court-order**⟩
⟨{**signal**, ¬**signal**}, **signal** ↔ **court-order**⟩
⟨{**x-fire**, ¬**x-fire**}, **x-fire** ↔ **signal**⟩
⟨{**y-fire**, ¬**y-fire**}, **y-fire** ↔ **signal**⟩
⟨{**dead**, ¬**dead**}, **dead** ↔ **x-fire** ∨ **y-fire**⟩

b. Causal ordering source filtered for **x-fire**
⟨{**court-order**, ¬**court-order**}, ¬**court-order**⟩
⟨{**signal**, ¬**signal**}, **signal** ↔ **court-order**⟩
~~⟨{**x-fire**, ¬**x-fire**}, **x-fire** ↔ **signal**⟩~~

6 This is where Santorio's (2019) formulation differs from mine. Instead of removing the target pair, Santorio replaces the proposition member of the target pair with the counterfactual antecedent. Concerning (23), this means that we replace the proposition **x-fire** ↔ **signal** in ⟨{**x-fire**, ¬**x-fire**}, **x-fire** ↔ **signal**⟩ with **x-fire**, which produces the pair ⟨{**x-fire**, ¬**x-fire**}, **x-fire**⟩. This version of filtering effectively removes target causal dependencies by rendering them uninformative: the contribution of ⟨{**x-fire**, ¬**x-fire**}, **x-fire**⟩ to the construction of the quantificational domain is **x-fire**, which is no different from the contribution of the counterfactual antecedent, which is again **x-fire**. Since the quantificational domain of a counterfactual is jointly determined by the counterfactual antecedent and the causal ordering source, the modified causal dependency is uninformative.
Another minor difference between Santorio's (2019) formulation and mine is that Santorio first adds the antecedent proposition to the causal ordering source, and then filters the outcome for the antecedent. On the other hand, I first filter the causal ordering source for the antecedent, and then add the antecedent to construct the domain of quantification. While the two versions effectively produce the same result, my formulation distinguishes the contribution of the antecedent from other causally relevant information (i.e., values of exogenous variables and causal dependencies), and maintains the standard modal base vs. ordering source distinction. See also Ciardelli, Zhang & Champollion 2018 for a similar idea.

$\langle\{\textbf{y-fire}, \neg\textbf{y-fire}\}, \textbf{y-fire} \leftrightarrow \textbf{signal}\rangle$
$\langle\{\textbf{dead}, \neg\textbf{dead}\}, \textbf{dead} \leftrightarrow \textbf{x-fire} \vee \textbf{y-fire}\rangle$

Given the filtered causal ordering source and the antecedent proposition, the quantificational domain of a counterfactual is constructed as follows: We first construct a set of propositions by iterating the causal ordering source (a set of pairs) and collecting the proposition member of each pair. Call it *the proposition set* of the causal ordering source. We add the antecedent proposition to the proposition set, and lastly, intersect the outcome to construct the domain of quantification. The set consists of antecedent-worlds that abide by the causal laws, given the values of exogenous variables.

Hereafter, I will use the term *the causal background* of p to refer to the intersection of the proposition set of a causal ordering source filtered for p. The causal background of p can be understood as the summary of the backgrounded facts that contribute to the truth condition of a given counterfactual. Jointly with the antecedent p, it determines the domain of quantification. Also, I will call the members of this quantificational domain *the causally relevant p-worlds (of the world of evaluation)*.

(28) Causal background
The causal background of p in a world w (notation: $\mathbf{bg}_w(p)$) is the intersection of the proposition set of a causal ordering source filtered for p in w.

(29) Causally relevant worlds
The causally relevant p-worlds of w are worlds that are members of the set resulting from the intersection of p with the causal background of p in w (i.e., $p \wedge \mathbf{bg}_w(p)$).

As an illustration, given the causal ordering source filtered for **x-fire** in (27b), the causal background of **x-fire** is the intersection of $\{\neg\textbf{court-order}, \textbf{signal} \leftrightarrow \textbf{court-order}, \textbf{y-fire} \leftrightarrow \textbf{signal}, \textbf{dead} \leftrightarrow \textbf{x-fire} \vee \textbf{y-fire}\}$. This causal background, in conjunction with the antecedent **x-fire**, determines the quantificational domain of a counterfactual.

6.2 The semantics of *ought*

At the beginning of Section 6, I offered a preliminary semantics of *ought* which abstracted over how the quantificational domain of a counterfactual is constructed. In what follows, I will fill in the missing parts of the analysis.

Let me first briefly summarize the differences between similarity-based theories of counterfactuals (with the Limit Assumption) and filtering semantics. Since the standard account of modality introduced earlier subsumes premise semantics for counterfactuals, I will use Kratzer's premise semantics for comparison.[7] Premise semantics and filtering semantics differ in the extent to which they pay attention to facts and dependencies between facts of the world of evaluation: the former pays maximal attention to facts and dependencies between facts, whereas the latter only takes into account causally relevant facts and causal dependencies. Due to the ways in which the two theories diverge, they differ in how the domain of quantification is constructed, and this amounts to saying that the selection function f selects the causally relevant p-worlds in filtering semantics whereas it selects the closest p-worlds in premise semantics.

(30) Comparison of the domain of quantification

a. Premise semantics
$f(w, p) = \textsc{best}(p)(t(w))$,
where t is a totally realistic ordering source

b. Filtering semantics
$f(w, p) = p \wedge \mathbf{bg}_w(p)$ (i.e., causally relevant p-worlds of w)

As far as the 'if p, ought p' problem is concerned, both theories are in principle capable of handling the problem: Filtering semantics explicitly encodes intervention, which alters the causal dependencies between the antecedent and its causes. Premise semantics by default does not guarantee that the dependencies between the antecedent and its causes are broken upon hypothesizing that the antecedent is true, but this can be achieved by constraining the totally realistic ordering source in a particular way that it is faithful to Lewis's (1979) idea that earlier affairs are overdetermined by later ones. However, there are empirical reasons to opt for filtering semantics. While a detailed discussion is deferred to Section 8, the table in (31) is a summary of comparison between the two theories.

7 Although filtering semantics is a version of premise semantics, I will reserve the term 'premise semantics' for Kratzer's premise semantics (Kratzer 1981b, Kratzer 1981a).

(31) Comparison of empirical coverage

	Filtering semantics	Premise semantics
Zvolenszky's puzzle	✓	✓
Modified love triangle	✓	✗
Flexible in picking up facts	✓	✗

For concreteness and consistency, I will use filtering semantics in all of my analyses. A fully worked-out semantics of *ought* can be given by replacing the selection function f in the preliminary semantics with (30b).

(32) The semantics of *ought* (final)

a. Auxiliary: counterfactual testing ground of p in w

$$\mathbf{cfg}_w(p) = \cup\{r \wedge \mathbf{bg}_w(p) \mid r \in Alt(p)\}$$

b. $[\![\text{ought } p]\!]^w$

$$= \exists w' \in (p \wedge \mathbf{bg}_w(p)) : \underbrace{\text{BEST}(\mathbf{cfg}_w(p))(d(w))}_{\textit{good worlds of } w}(w') = 1$$
$$\wedge \forall r \in Alt(p) \text{ s.t. } r \neq p :$$
$$\neg \exists w' \in (r \wedge \mathbf{bg}_w(r)) : \underbrace{\text{BEST}(\mathbf{cfg}_w(p))(d(w))}_{\textit{good worlds of } w}(w') = 1$$

Note that in the above formula, the causal ordering source is not only filtered for the antecedent p, but also for each of the alternatives r to p. This implies that if the antecedent and its alternatives are not parts of a single causal variable (i.e., they are not members of a single partition), one needs to intervene on more than one causal variable. While there is no problem with doing so, I suggest that intervention can be confined to a single variable as far as the interpretation of *ought* is concerned. Typically, *ought* statements are concerned with the deontic status of the prejacent and its negation. For example, when interpreting 'Britney Spears must drink Coca-Cola in public', we are not concerned with any alternative actions Britney can take, with the exception of her not drinking Coca-Cola in public. Likewise, when interpreting 'the Dalai Lama should be angry', we are not concerned with any alternative mental state of the Dalai Lama, with the exception of him not being angry. In these cases, the contextually salient set of alternatives is $\{p, \neg p\}$, where p is the prejacent. The alternative set constitutes a single causal variable **p?**, which asks whether p is true or false.

Some uses of *ought* are concerned with the evaluation of the possible courses of actions an agent can take given a salient set of choices (Williams

1981), and in such cases the alternative set does not consist of p and $\neg p$. The self-frustrating *ought* scenario is an illustrative case; we deliberate on the agent's decision to go to Aleppo or to stay in Damascus. I argue that it is natural to pack the set of the agent's choices into a single causal variable, as they are mutually exclusive. In the self-frustrating *ought* scenario, there is no world in which you go to Aleppo and simultaneously stay in Damascus, so the two propositions are mutually exclusive. The other requirement for partitionhood, such that the propositions need to be jointly exhaustive, can be satisfied by restricting the domain of the partitions (thus the domain of casual variables) to only those worlds in which at least one of the alternatives is true. In the self-frustrating *ought* scenario, this would mean that we are only concerned with the worlds where you are either in Aleppo or in Damascus.

There are a number of theories of deontic modality that are closely related to the suggested semantics. My proposal can find its ancestor in Jackson's (1985) *actualism*, in the sense that counterfactual reasoning is involved in the interpretation of *ought*. The version of Jackson's semantics that adopts the Limit Assumption is as follows: 'ought p' is true w.r.t. an alternative set $Alt(p)$ if and only if every world that might be actual were p true (i.e., the closest p-worlds from Jackson's perspective) is better than every world that might be actual were r true (i.e., the closest r-worlds), for all $r \in Alt(p)$ such that $r \neq p$. My proposal crucially differs from Jackson's in that I endorse an interventionist approach to counterfactuals whereas Jackson adopts a similarity-based framework (Stalnaker 1968, Lewis 1973, Kratzer 1981a).

More recently, Arregui (2011) proposes that the interpretation of deontic modals involves counterfactual-style revisions because one needs to pay maximal attention to facts and dependencies between facts. Due to the interventionist approach I endorse, my semantics differs from Arregui in that I only pay attention to causally relevant facts and dependencies and allow severing of certain causal dependencies. In Arregui's system, law-like dependencies are never violated. This view was supported by examples that concern *non-causal* dependencies between facts. What I show in this paper is that some dependencies need to be ignored, particularly the ones that involve causal reasoning.

A similar approach has been taken in the analysis of desire verbs (Heim 1992; see also Villalta 2008, Ogihara 2014, Rubinstein 2017, and von Fintel & Iatridou 2017). In order to capture the non-monotonic characteristics of *want*, Heim proposes that 'α wants p' is true if and only if every doxastically acces-

sible world w' from the world of evaluation is such that every closest p-world from w' is more desirable than any closest $\neg p$-world from w'. Heim's analysis has an additional layer of quantification due to the doxastic accessibility relation, and just like Jackson, adopts a similarity-based framework.

Before concluding this section, I would like to note that there is a simpler way of formulating the proposed semantics if we grant the assumption that the modal prejacent and its alternatives are mutually exclusive.[8] Instead of introducing existential quantifiers, the formula in (33) universally quantifies over the deontically best worlds within the counterfactual testing ground and asserts that the prejacent is true in the quantified worlds.[9] It minimally differs from the standard account in that the modal background is defined in terms of the counterfactual testing ground of the prejacent.

(33) Alternative formulation of the proposal
$$[\![\text{ought } p]\!]^w = \forall w' \in \underbrace{\textsc{best}(\mathbf{cfg}_w(p))(d(w))}_{\textit{good worlds of } w} : \; p(w') = 1$$

For reasons of familiarity, I will use the alternative formulation in (33) in my analyses except when it becomes crucial to provide a full-fledged conditional semantics.

8 Recall that I am assuming that the set of alternatives either consists of (i) the prejacent and its negation or (ii) mutually exclusive choices.

9 Proof of equivalence: (i) To prove that the original formula implies the alternative formula, suppose that the latter is false, that is, there exists a deontically best world which is a $\neg p$-world. Granted that p and its alternatives are mutually exclusive (hence all alternative-to-p-worlds are $\neg p$-worlds), the supposition is in contradiction with the original formula because the latter implies that no $\neg p$-world is a deontically best world. (ii) To prove that the alternative formula implies the original formula, suppose that the latter is false, that is, no causally relevant p-world is deontically best or there exists a causally relevant alternative-to-p-world which is deontically best. First, if no causally relevant p-world is deontically best then it follows that the alternative formula is false because by definition, every p-world in the modal background is a causally relevant p-world and the set of deontically best worlds is a subset of the modal background; the alternative formula asserts that the deontically best worlds are all p-worlds, but if no causally relevant p-world is deontically best then no p-world can be deontically best. Second, if there exists a causally relevant alternative-to-p-world which is deontically best, then it follows that the alternative formula is false due to the assumption that all alternative-to-p-worlds are $\neg p$-worlds.

7 Analysis

Let us revisit Zvolenszky's and Carr's concerns. Zvolenszky criticizes Frank's expansion strategy for being ad hoc, which solely exists to circumvent the 'if p, ought p' problem. She also argues that it cannot make systematic predictions for the Coca-Cola example and the Dalai Lama example. In the proposed semantics, it is not stipulated that whatever information that entails or contradicts the prejacent is removed from the set of circumstantially relevant facts. Rather, expansion of the modal background is merely a side effect of quantifying over the counterfactual prejacent-worlds and each of the counterfactual alternative-to-the-prejacent-worlds. Since the prejacent and its alternatives are mutually exclusive, the modal background of *ought* — which I define as the counterfactual testing ground of the prejacent — does not force the truth of the prejacent or any of its alternatives.

Zvolenszky's second criticism is related to Carr's worry, since it boils down to the issue that certain facts that come and go together cannot be separated without ignoring the salient dependency between them. For instance, the Dalai Lama's anger is tightly correlated with the occurrence of a sufficiently reprehensible event, and Death is always waiting for you at the target destination, regardless of the choice you make. In what follows, I show that this is no longer an issue because we are able to ignore the link between the correlated propositions by making a counterfactual assumption.

In what follows, I flesh out the technical details of my analysis. As in the prominent response, I will assume that the 'if p, ought p' examples require invoking the double modalization strategy. However, the counterfactual semantics built into the deontic modal ensures that the dependency between the prejacent and its cause is ignored. Moreover, it offers an explanation of why the examples do not have a single modal construal and prevents overgeneration.

7.1 Why Death is still in Aleppo and the Dalai Lama is rightfully angry

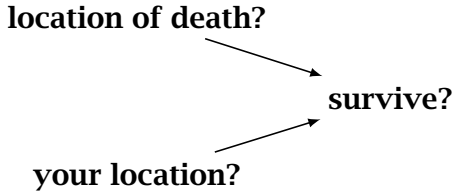
In the self-frustrating *ought* scenario repeated below as (34), Carr's concern is that upon interpreting the consequent of 'If you go to Aleppo, you ought not to go to Aleppo', we have to disregard the hypothetical assumption 'you go to Aleppo' but keep 'Death is in Aleppo', which was inferred from the disregarded hypothetical assumption. Given the tight correlation between the two propositions, Carr doubts that this is a coherent configuration of background assumptions.

(34) a. If you go to Aleppo, you ought not to go to Aleppo (because Death will be there).
b. If you stay in Damascus, you ought not to stay in Damascus (because Death will be there).

In my analysis, what Carr regards as an incoherent configuration of background assumptions is justified by the way in which the modal background is constructed. In fact, it is the only configuration derived from my analysis. I will assume that the indicative conditional in 'If you go to Aleppo, you ought not to go to Aleppo' takes us from the world of evaluation w to worlds, say w', where Death and you are both in Aleppo. This can be implemented within premise semantics by requiring that the stereotypical ordering source $s(w)$ contains the information that '**death-alp** → **you-alp**' and '**death-dms** → **you-dms**'. In each w', the counterfactual testing ground of the prejacent is constructed, and the deontically best worlds are selected from the worlds in the counterfactual testing ground.

The relevant causal graph for the self-frustrating *ought* scenario is depicted in (35). Despite the fact that you being in Aleppo (Damascus) is a good indication of Death being in Aleppo (Damascus), your whereabouts are not causally dependent on Death's whereabouts. Thus, **location of death?** and **your location?** are exogenous variables. The two variables affect the endogenous variable **survive?**, which captures the intuition that you will be die whenever you and Death end up in the same location. We are interested in how 'ought ¬**you-alp**', which is equivalent to 'ought **you-dms**' in the given context, is valued in every epistemically best **you-alp**-world. In each of these worlds, Death is in Aleppo, so the value of **location of death?** is **death-alp**. The causal ordering source before intervention is given in (36).

(35) Causal dependencies in the self-frustrating *ought* scenario



(36) Causal ordering source
⟨{**death-alp**, **death-dms**}, **death-alp**⟩
⟨{**you-alp**, **you-dms**}, **you-alp**⟩
⟨{**survive**, ¬**survive**}, **survive** ↔ $X \neq Y$ in **you-X** ∧ **death-Y**⟩

Intervening on the variable **your location?** does not affect the causal graph because **your location?** is already an exogenous variable. The corresponding filtered causal ordering source is given in (37).

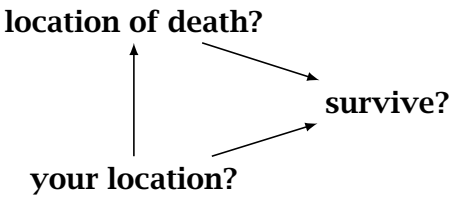
(37) Causal ordering source filtered for **you-dms**
⟨{**death-alp**, **death-dms**}, **death-alp**⟩
~~⟨{**you-alp**, **you-dms**}, **you-alp**⟩~~
⟨{**survive**, ¬**survive**}, **survive** ↔ $X \neq Y$ in **you-X** ∧ **death-Y**⟩

The counterfactual testing ground of **you-dms** is the union of the causally relevant **you-dms**-worlds and the causally relevant **you-alp**-worlds. Assuming that the two alternatives jointly exhaustify the domain, this set consists of worlds in which (i) Death is in Aleppo, (ii) you are either in Aleppo or Damascus, and (iii) the causal law concerning your survival is abided by. Given that Death's location is fixed to Aleppo, the best worlds in the counterfactual testing ground are ones in which you are in Damascus (you would survive).

(38) $[\![\text{if}_{ind}$ **you-alp**, ought ¬**you-alp**$]\!]^w = [\![\text{if}_{ind}$ **you-alp**, ought **you-dms**$]\!]^w$
$= \forall w' \in \underbrace{\text{BEST}(\cap(e(w) \cup \{\textbf{you-alp}\}))(s(w))}_{\textit{epistemically best } \textbf{you-alp}\textit{-worlds of } w} :$
$\forall w'' \in \underbrace{\text{BEST}(\textbf{cfg}_{w'}(\textbf{you-dms}))(d(w'))}_{\textit{good worlds of w'}} :\ \textbf{you-dms}(w'') = 1$

To further test the predictions, let us consider a slightly different scenario, where Death does not predict your whereabouts and makes a move, but rather keeps track of your location and makes a move based on the observation. In this scenario, it is fairly difficult to judge 'if you go to Aleppo, you ought not to go to Aleppo' as true. This is exactly what the proposed semantics predicts. As depicted in the following causal graph, **your location?** causally influences **location of death?**:

(39) Death keeps track of your whereabouts



(40) Corresponding causal ordering source
⟨{**death-alp**, **death-dms**},
you-alp → **death-alp** ∧ **you-dms** → **death-dms**⟩
⟨{**you-alp**, **you-dms**}, **you-alp**⟩
⟨{**survive**, ¬**survive**}, **survive** ↔ $X \neq Y$ in **you-X** ∧ **death-Y**⟩

Intervening on **your location?** does not change the causal graph, as the variable is already exogenous. If we set the value of **your location?** to **you-alp**, the value of **location of death?** is set to **death-alp**. And if we set the value of **your location?** to **you-dms**, the value of **location of death?** is set to **death-dms**. Therefore, the causally relevant **you-alp**-worlds are **death-alp**-worlds, and the causally relevant **you-dms**-worlds are **death-dms**-worlds.

(41) Causal ordering source filtered for **your location?**
⟨{**death-alp**, **death-dms**},
you-alp → **death-alp** ∧ **you-dms** → **death-dms**⟩
~~⟨{**you-alp**, **you-dms**}, **you-alp**⟩~~
⟨{**survive**, ¬**survive**}, **survive** ↔ $X \neq Y$ in **you-X** ∧ **death-Y**⟩

The counterfactual testing ground of **you-dms** consists of worlds in which you and Death are both in Aleppo and worlds where you and Death are both in Damascus. As far as your survival is concerned, the two sets of worlds are tied. Consequently, (38) would be false in this scenario because the best worlds are not necessarily **you-dms**-worlds.

The proposed semantics also makes the right prediction for the Dalai Lama example repeated below as (42). The intuition behind this example is that the Dalai Lama should be angry in the presence of a sufficiently reprehensible event (it would be bad if the Dalai Lama were not angry). My analysis captures this intuition.

(42) If the Dalai Lama is angry, then he should be angry.

An occurrence of a sufficiently reprehensible event (i.e., **trigger**) induces the Dalai Lama's anger, so we can posit the causal graph in (43).

(43) Causal dependencies in the Dalai Lama example

trigger? ⟶ **angry?**

I will assume that the stereotypical ordering source $s(w)$ supplied to the indicative conditional contains '**trigger** → **angry**' and '¬**trigger** → ¬**angry**',

and the indicative conditional takes us to the worlds, say w', where **angry** and **trigger** are both true. We can set the value of the exogenous variable **trigger?** to **trigger**. The corresponding causal ordering source is given in (44).

(44) Causal ordering source
⟨{**trigger**, ¬**trigger**}, **trigger**⟩
⟨{**angry**, ¬**angry**}, **trigger** ↔ **angry**⟩

We intervene on **angry?**, removing the causal link between **trigger?** and **angry?** as in (45). So the causal ordering source filtered for **angry** solely consists of the value of the exogenous variable **trigger?**. The causal background of **angry** is thus identical to **trigger**.

(45) Causal dependencies after intervening on **angry?**

trigger? **angry?**

(46) Causal ordering source filtered for **angry**
⟨{**trigger**, ¬**trigger**}, **trigger**⟩
~~⟨{**angry**, ¬**angry**}, **trigger** ↔ **angry**⟩~~

(47) $[\![\text{if}_{ind}\ \textbf{angry}, \text{should}\ \textbf{angry}]\!]^w$
$= \forall w' \in \underbrace{\text{BEST}(\cap(e(w) \cup \{\textbf{angry}\}))(s(w))}_{\textit{epistemically best } \textbf{angry}\textit{-worlds of } w} :$
$\forall w'' \in \underbrace{\text{BEST}(\textbf{cfg}_{w'}(\textbf{angry}))(d(w'))}_{\textit{good worlds of } w'} :\ \textbf{angry}(w'') = 1$

The counterfactual testing ground of **angry** is the union of the causally relevant **angry**-worlds and the causally relevant ¬**angry**-worlds, both of which are **trigger**-worlds. Given that **trigger** is true in the worlds under consideration, the best worlds are all **angry**-worlds.

While the proposed semantics renders the Dalai Lama example and the self-frustrating *ought* examples true, the Coca-Cola example, repeated below as (48), is predicted as false. Unlike in the Dalai Lama example, no causal variable influences Britney's decision to drink Coca-Cola in public, and the causal graph in (49) illustrates this.

(48) If Britney Spears drinks Coca-Cola in public, then she must drink Coca-Cola in public.

(49) No causal dependency in the Coca-Cola example

coke?

When identifying the set of epistemically best **coke**-worlds, no proposition that is relevant to the evaluation of *must* is identified. This contrasts the Dalai Lama example and the self-frustrating *ought* examples, where **trigger** and **death-aleppo** are inferred via backtracking, respectively. The causal ordering source contains one member, which is the value of the exogenous variable **coke?**. Intervening on **coke?** filters out itself, and the filtered causal ordering source is the empty set.

(50) Causal ordering source
$\langle \{\textbf{coke}, \neg\textbf{coke}\}, \textbf{coke}\rangle$

(51) Causal ordering source filtered for **coke** (empty set)
~~$\langle \{\textbf{coke}, \neg\textbf{coke}\}, \textbf{coke}\rangle$~~

The upshot is that the counterfactual testing ground of **coke** is the union of the **coke**-worlds and the ¬**coke**-worlds, which is the set of all worlds. This is equivalent to having no restrictions on the modal base in constructing the deontically best worlds. In the very best worlds, Britney does not drink Coca-Cola in public because that would be a breach of contract with Pepsi.

(52) $[\![\text{if}_{ind}\ \textbf{coke}, \text{must}\ \textbf{coke}]\!]^w$
$= \forall w' \in \underbrace{\text{BEST}(\cap(e(w) \cup \{\textbf{coke}\}))(s(w))}_{\textit{epistemically best } \textbf{coke}\textit{-worlds of } w} :$
$\forall w'' \in \underbrace{\text{BEST}(\textbf{cfg}_{w'}(\textbf{coke}))(d(w'))}_{\textit{good worlds of } w'} :\ \textbf{coke}(w'') = 1$

7.2 Avoiding overgeneration

As noted in section 5.1, the standard account coupled with the double modalization strategy has trouble explaining why the single modal reading is not available for sentences of the form 'if p, ought p'. For instance, why is the single modal reading of the Coca-Cola example — which yields trivial truth — categorically unavailable? In order to explain this phenomenon, I will use the existential quantification-based formulation in (32) which explicitly manifests that counterfactual semantics is built into the semantics of *ought*.

I argue that the single modal reading is not available for such examples because letting the *if*-clause directly restrict the deontic modal gives rise to an inconsistent modal base, which triggers *ex falso quodlibet* which is counterintuitive. In the proposed semantics, 'ought p' is inherently counterfactual. It quantifies over the counterfactual p-worlds and the counterfac-

tual alternative-to-p-worlds, and checks which quantified set contains the best worlds among them. So if an *if*-clause were to directly restrict the deontic modal, it would be restricting the modal base of the counterfactuals. Consider the example in (53), the single modal reading of which is provided in (54). Assuming that the contextually salient set of alternatives is {**convene**, ¬**convene**}, the interpretation of 'ought **convene**' involves two counterfactuals that respectively quantify over the causally relevant **convene**-worlds and the causally relevant ¬**convene**-worlds. In the single modal reading, **murder** restricts the modal base of the two counterfactuals, and accordingly the modal background of *ought* solely consists of **murder**-worlds. The deontically best worlds in the counterfactual testing ground are all **convene**-worlds, and the sentence is predicted as true.

(53) If a murder occurs, the jurors ought to convene.

(54) $[\![$if **murder**, ought **convene**$]\!]^w$

$$= \exists w' \in (\underbrace{\overbrace{\boxed{\textbf{murder}}}^{\textit{restrictor}} \wedge \textbf{convene}}_{\textit{modal base}} \wedge \overbrace{\boxed{\textbf{bg}_w(\textbf{convene})}}^{\textit{filtered ordering source}}) : \underbrace{\textsc{best}(\textbf{cfg}_w(\textbf{convene}))(d(w))(w') = 1}_{\textit{good worlds of } w}$$

$$\underbrace{\textit{causally relevant } \textbf{convene}\textit{-worlds of } w}$$

$$\wedge \neg \exists w' \in (\underbrace{\overbrace{\boxed{\textbf{murder}}}^{\textit{restrictor}} \wedge \neg\textbf{convene}}_{\textit{modal base}} \wedge \overbrace{\boxed{\textbf{bg}_w(\neg\textbf{convene})}}^{\textit{filtered ordering source}}) : \underbrace{\textsc{best}(\textbf{cfg}_w(\textbf{convene}))(d(w))(w') = 1}_{\textit{good worlds of } w}$$

$$\underbrace{\textit{causally relevant } \neg\textbf{convene}\textit{-worlds of } w}$$

Let us return to the overgeneration problem. As noted earlier, the alternative set is typically $\{p, \neg p\}$. In such cases, the *if*-clause in 'if p, ought p' would be directly restricting the counterfactual operator that takes $\neg p$ as its antecedent, and the modal base — which should have consisted of consistent propositions — would contain both p (due to the restrictor) and $\neg p$ (due to the modal prejacent). The conjunction of p and $\neg p$ is a contradiction, and any valuation with respect to a contradiction is an instance of *ex falso quodlibet* which is counterintuitive. Thus the single modal reading not only yields trivial truth but also gives rise to a counterintuitive inference, and I suggest that the latter is what prevents one from readily perceiving a trivial reading. Intuitively, for the Coca-Cola example, the single modal reading supposes

that Britney Spears drank Coca-Cola in public *and* did not drink Coca-Cola in public, which is contradictory.

(55) Single modal reading of (48)

$$[\![(48)]\!]^w = \exists w' \in \Big(\overbrace{\big(\underbrace{\boxed{\textbf{coke}}}_{\textit{antecedent}} \wedge \underbrace{\boxed{\textbf{coke}}}_{\textit{restrictor}} \big)}^{\textit{modal base (redundant restrictor)}} \wedge \overbrace{\boxed{\textbf{bg}_w(\textbf{coke})}}^{\textit{filtered ordering source}} \Big) : \underbrace{\textsc{best}(\textbf{cfg}_w(\textbf{coke}))(d(w))(w') = 1}_{\textit{good worlds of } w}$$

$$\wedge \neg \exists w' \in \Big(\overbrace{\big(\underbrace{\boxed{\neg\textbf{coke}}}_{\textit{antecedent}} \wedge \underbrace{\boxed{\textbf{coke}}}_{\textit{restrictor}} \big)}^{\textit{inconsistent modal base!}} \wedge \overbrace{\boxed{\textbf{bg}_w(\neg\textbf{coke})}}^{\textit{filtered ordering source}} \Big) : \underbrace{\textsc{best}(\textbf{cfg}_w(\textbf{coke}))(d(w))(w') = 1}_{\textit{good worlds of } w}$$

For cases in which the alternative set characterizes the set of salient choices an agent can take, the members of the alternative set are mutually exclusive. Therefore, had 'if p' restricted a counterfactual operator that takes any of p's alternatives as its antecedent, it would again yield an inconsistent modal base. So either way, the single modal reading 'if p, ought p' is marked.

8 Interventionist approach or similarity-based semantics?

In the previous section, I fleshed out the analysis of the 'if p, ought p' examples using Santorio's filtering semantics which implements intervention. While interventionist approaches make it explicit that the dependency between the counterfactual antecedent and its cause needs to be broken, it cannot be hastily concluded that similarity-based accounts (Stalnaker 1968, Lewis 1973, Kratzer 1981a) fail to make the same prediction. Specifically, it could be the case that defining the selection function f in (20) (see also (30a)) in terms of Kratzer's premise semantics can handle the problematic examples; as long as the theory guarantees that the dependencies between the antecedent and its causes are ignored, there would be no difference in predictions concerning the 'if p, ought p' examples under discussion. While an unadorned version of premise semantics does not guarantee this, it seems possible to reproduce the results by constraining the totally realistic ordering source $t(w')$ in a way that the induced ordering $\leq_{t(w')}$ faithfully implements Lewis's (1979) notion of similarity. Lewis claims that earlier affairs are

overdetermined by later ones, in the sense that whatever happened earlier leaves many traces in the world and it typically takes removing more than one trace (i.e., happening of a miracle) to manipulate the earlier affair. What this means for the Dalai Lama example is that in an epistemically best **angry**-world w' where **trigger** and **angry** are both true, the closest $\neg$**angry**-worlds of w' are worlds where **trigger** is true. This is because the occurrence of a sufficiently reprehensible event precedes the change of the Dalai Lama's mental state; manipulating the value of **trigger** requires eliminating all of its traces, and eliminating each trace makes a world more distant from w'. For the self-frustrating *ought* example, given that you and Death are both in Aleppo in an epistemically best **you-alp**-world w', the closest **you-dms**-worlds of w' are **death-alp**-worlds because the change in Death's whereabouts (or Death's prediction of your whereabouts) precedes the change in yours. In both cases, the dependency between the counterfactual antecedent and its cause happens to be ignored.

In what follows, I present two pieces of evidence in favor of an interventionist approach, at least when the question is about which theory of counterfactuals to build into the semantics of *ought*.

8.1 Modified love triangle

There is a case in which incorporating filtering semantics into the semantics of *ought* makes better predictions. The following scenario is originally due to Santorio (2019), but it additionally includes the assessments of the possible outcomes from Andy, Billy, and Charlie's perspectives.

> Andy, Billy, and Charlie are in a love triangle. Billy is pursuing Andy; Charlie is pursuing Billy; and Andy is pursuing Charlie. Each of them is very annoyed by their suitor and wants to avoid them. There's a party going on and all three were invited. Each of them keeps track of whether the person they like is going. Each of them wants an occasion to spend time with their beloved and without their suitor. Having an occasion of this kind would be sufficient for each of them to go. But under the present circumstances, they are in a deadlock and none of them would make it to the party. For Andy's sake, it would be best if he and Charlie were at the party; it would be disastrous

if all three of them were at the party, or if only Bill and Charlie were there. Similarly for Billy and Charlie.

In this situation, (56) is false but (57) is true.

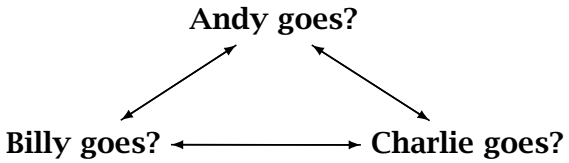
(56) For Andy's sake, he ought to be at the party.

(57) For Billy's sake, Andy ought to be at the party.

Let me first note that even the standard account has to consider the possible consequences of Andy being at the party to make proper deontic judgments in the above scenario. Otherwise, (56) would be predicted as true because for Andy's sake, the worlds in which Andy and Charlie (but not Billy) are at the party are ideal, and it follows that every best world is a world where Andy is at the party. In view of the standard account, the modal background of *ought* somehow needs to be restricted to ensure that no world in which only Andy and Charlie are at the party is included in the set. Had such worlds been present in the modal background, the best worlds would exclusively consist of those worlds and (56) would be rendered true. The way in which the proposed semantics configures the modal background via counterfactual reasoning can be understood as a means to correctly restrict the modal background.

So which theory of counterfactuals better suits this purpose? The semantics of *ought* proposed in (32) makes the right predictions. According to Santorio, the corresponding causal graph and the causal ordering source are as follows:

(58) Causal dependencies in the modified love triangle scenario

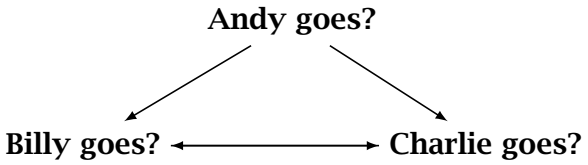


(59) Causal ordering source
$\langle \{\mathbf{andy}, \neg\mathbf{andy}\}, \mathbf{andy} \leftrightarrow (\mathbf{charlie} \wedge \neg\mathbf{billy})\rangle$
$\langle \{\mathbf{billy}, \neg\mathbf{billy}\}, \mathbf{billy} \leftrightarrow (\mathbf{andy} \wedge \neg\mathbf{charlie})\rangle$
$\langle \{\mathbf{charlie}, \neg\mathbf{charlie}\}, \mathbf{charlie} \leftrightarrow (\mathbf{billy} \wedge \neg\mathbf{andy})\rangle$

Recall that the counterfactual testing ground of **andy** is defined as the union of the counterfactual **andy**-worlds and the counterfactual ¬**andy**-

worlds. Let us first examine how the counterfactual **andy**-worlds look like in filtering semantics. We intervene on **Andy goes?**, and the resulting causal graph and the causal ordering source are given as follows:

(60) Causal dependencies after intervening on **Andy goes?**



(61) Causal ordering source filtered for either **andy** or ¬**andy**
~~⟨{**andy**, ¬**andy**}, **andy** ↔ (**charlie** ∧ ¬**billy**)⟩~~
⟨{**billy**, ¬**billy**}, **billy** ↔ (**andy** ∧ ¬**charlie**)⟩
⟨{**charlie**, ¬**charlie**}, **charlie** ↔ (**billy** ∧ ¬**andy**)⟩

The analysis of (56) is given in (62b). The causally relevant **andy**-worlds, which are derived from intersecting **andy**, **billy** ↔ (**andy** ∧ ¬**charlie**), and **charlie** ↔ (**billy** ∧ ¬**andy**), are worlds in which **billy** is true and **charlie** is false. Intervening on **Andy goes?** but adding ¬**andy** instead of **andy** to the hypothetical stock of assumptions gives us the causally relevant ¬**andy**-worlds, which are worlds where both **billy** and **charlie** are false. So the counterfactual testing ground of **andy** is the union of (i) the worlds in which Andy is at the party, Billy is at the party, but Charlie is not at the party and (ii) the worlds in which none of the three are at the party. For Andy's sake, being at the party does no good to him; in fact, it could even be better to stay home. The best worlds in the counterfactual testing ground of **andy** are all ¬**andy**-worlds, so the analysis in (62b) predicts (56) to be false.

(62) a. Counterfactual testing ground of **andy** in w
$\mathbf{cfg}_w(\mathbf{andy})$
$= \cup\{(\mathbf{andy} \wedge \mathbf{billy} \wedge \neg\mathbf{charlie}), (\neg\mathbf{andy} \wedge \neg\mathbf{billy} \wedge \neg\mathbf{charlie})\}$

b. $[\![\text{ought } \mathbf{andy}]\!]^w = \forall w' \in \underbrace{\text{BEST}(\mathbf{cfg}_w(\mathbf{andy}))(d(w))}_{\textit{good worlds of } w}\colon\ \mathbf{andy}(w') = 1$

By contrast, (57) is predicted to be true. We again intervene on **Andy goes?** and identify the same counterfactual testing ground of **andy**. However, the deontic ordering source $d(w)$ idealizes different worlds. For Billy's sake, the best worlds are ones in which he and Andy are at the party, with the exclusion of Charlie. Therefore, every deontically best world in the counterfactual testing ground is an **andy**-world.

Adopting a similarity-based account of counterfactuals leads to different predictions. Santorio points out that in a non-backtracking interpretation, similarity-based theories predict that the closest worlds where Andy is at the party are worlds where Billy and Charlie are both at the party. This is due to the validation of the inference rule *loop*, which is illustrated below:[10]

(63) Loop
andy $\Box\!\!\rightarrow$ **billy**
billy $\Box\!\!\rightarrow$ **charlie**
charlie $\Box\!\!\rightarrow$ **andy**
andy $\Box\!\!\rightarrow$ **charlie**

Thus, the counterfactual testing ground of **andy** would consist of worlds in which (i) Andy, Billy, and Charlie are all at the party and (ii) none of them are at the party. The proposed counterfactual-based semantics of *ought* coupled with any similarity-based theory of counterfactuals would convey that no world in the latter is a best world. From any of the three rivals in love's perspective, the latter worlds evidently outrank the former ones because it would be disastrous to have all three of them at the party. As for (56), the similarity-based semantics of *ought* correctly predicts that the sentence is false. However, the problem is that it also predicts that (57) is false, contrary to intuition.

Santorio (p.c.) notes that a backtracking interpretation is available to some speakers, where the counterfactual worlds in which Andy is at the party are worlds where Billy or Charlie is at the party. While there is no doubt about the availability of such a reading, it seems that this cannot be how the counterfactuals built into *ought* can be interpreted in (56) because that would predict that the sentence is true. We would be comparing (i) the worlds in which Andy is at the party and Billy or Charlie is at the party to (ii) the worlds where none of them are at the party. For Andy's sake, the best worlds are where he and Billy are at the party, and such worlds are in the former set. It follows that none of the latter worlds are best worlds, and (56) is predicted to be true, contrary to the fact.

I would like to emphasize that it is not my intention to argue that interventionist approaches are superior to similarity-based ones in general. Rather, my point is that a theory that implements intervention — Santorio's filtering semantics in particular — better suits my counterfactual-based se-

10 The readers are referred to Santorio 2019 and Halpern 2013 for formal proofs.

mantics of *ought*, as it makes the right predictions for both the 'if p, ought p' examples and the modified love triangle examples.

On a side note, I would like to point out that Jackson's (1985) semantics of *ought* makes the wrong prediction for (57). The version of his semantics that adopts the Limit Assumption states that 'ought p' is true w.r.t. an alternative set $Alt(p)$ if and only if every closest p-world is better than every closest r-world, for all $r \in Alt(p)$ such that $r \neq p$. Concerning (57), he would compare (i) the worlds in which either Andy, Billy, and Charlie are all at the party (non-backtracking reading) or Andy is at the party and Billy or Charlie is at the party (backtracking reading) to (ii) the worlds where none of them are at the party. Whether one opts for the backtracking reading or the non-backtracking reading, it is not the case that every world in the former is better than any world in the latter: it would be disastrous if all three of them attended the party (non-backtracking reading), and it would also be horrifying to Billy if only Andy and Charlie attended the party (some worlds quantified over by a backtracking counterfactual).

Lastly, I would like to note that the 'if p, ought p' examples and the modified love triangle examples are not the only cases in which we have to speculate about the possible consequences of bringing about the prejacent. Consider the famous Professor Procrastinate scenario (Jackson & Pargetter 1986) depicted below:

> Professor Procrastinate receives an invitation to review a book. He is the best person to do the review, and the best thing that can happen is that he says yes, and then writes the review. However, suppose it is further the case that were Procrastinate to say yes, he would not in fact get around to writing the review because he would keep on putting the task off. Thus, what would in fact happen were he to say yes is that he would not write the review. Moreover, this is the worst that can happen. It would lead to the book not being reviewed at all.

(64) Procrastinate ought to accept.

(65) Procrastinate ought to accept and write.

In the above scenario, (64) is false but (65) is true. Von Fintel (2012) claims that (64) is false because "we are led to believe that the ideal course of action (accepting and writing) is not available, which then, under those realistic constraints, makes us assent to *Procrastinate ought not to accept*". In other

words, when assessing whether or not Procrastinate ought to accept, we are required to speculate about how the world would be if Procrastinate accepted the review request.

8.2 Interventionist approaches are more flexible in picking up facts

An anonymous reviewer points out that an *ought* statement that is not embedded under a conditional is quite flexible in picking up circumstantially relevant facts of the world of evaluation. As shown in (66), we can judge 'the Dalai Lama should be angry' true, given a trigger. However, we may want to talk about our ideals as in (67), and deliver that there should not have been a trigger in the first place.

(66) a. There was a lot of police brutality at the Peace March.
b. The Dalai Lama should be angry (and he is).

(67) a. There was a lot of police brutality at the Peace March, government propaganda and repression. The Dalai Lama is very angry.
b. We live in such terrible times! The Dalai Lama should not be angry…Those things should not be happening…What is the world coming to?

The standard account can predict that 'the Dalai Lama should not be angry' in (67b) is true by assuming that **trigger** is circumstantially irrelevant. If **trigger** is ignored, the deontically best worlds will be worlds where the Dalai Lama is not angry because we do not idealize unconditional anger.[11] The

11 As an anonymous reviewer notes, such a flexibility suddenly disappears when *ought* is embedded in a conditional. Example (ib) cannot be read as 'if the Dalai Lama is angry, something terrible that justifies his anger has happened but that should not have happened and he should not be angry'. Under the assumption that the epistemically best **angry**-worlds are **trigger**-worlds, (ib) cannot be interpreted in the same way as (67b), where **trigger** is regarded as circumstantially irrelevant.

(i) a. If the Dalai Lama is angry, he should be angry.
b. #If the Dalai Lama is angry, he should not be angry.

My speculation is that a conditional that embeds *ought* restricts the deontic accessibility relation in a way that whatever has been inferred from the conditional cannot be disregarded. In the case of the Dalai Lama example, it would be **trigger**, and the circumstantially relevant worlds need to be **trigger**-worlds. After all, conditionals are typically used to discuss how the worlds are like in the relevant antecedent-worlds, so it is not too far-fetched to assume that such a restriction is imposed.

anonymous reviewer suggests that such a flexibility in picking up relevant facts is a preferred property of deontic modals.

I agree that there are reasons for wanting to keep the flexibility. However, similarity-based theories of counterfactuals cannot offer room for such a flexibility, because the theories require that counterfactuals pay maximal attention to facts and dependencies between facts of the world of evaluation. There is no way to leave out **trigger** in identifying the domain of quantification.

On the other hand, interventionist approaches do not necessarily impose such a requirement. In fact, Ciardelli, Zhang & Champollion (2018) conduct an experiment that shows that it is possible to drop contextually salient information in assessing a counterfactual. The participants were provided with the following scenario and were asked to give truth value judgments for the sentences in (68):

> Imagine a long hallway with a light in the middle and with two switches, one at each end. One switch is called switch A and the other one is called switch B. The light is on whenever both switches are in the same position (both up or both down); otherwise, the light is off. Right now, switch A and switch B are both up, and the light is on. But things could be different…
>
> [Ciardelli, Zhang & Champollion 2018: 578]

(68) a. If switch A was down, the light would be off.
b. If switch B was down, the light would be off.
c. If switch A or switch B was down, the light would be off.

Ciardelli et al. report that approximately two thirds of the participants judged the sentences as true. However, a third of the participants judged them as indeterminate. Ciardelli et al. suggest that the minority judgment is due to a less salient reading of counterfactuals where only the counterfactual antecedent and the causal laws are considered, ignoring the current state of the system. As a consequence, the participants take into consideration all possible positions of the switches that are compatible with the antecedent and the causal law. Since the antecedent and the causal law do not fix the state of the light, the sentences are judged as indeterminate.

Once we grant that there is a possible reading of counterfactuals which only takes into account the antecedent and the causal laws, the proposed

semantics can make sense of (67b). We need not take into account **trigger** in constructing the causal background of **angry** in w (i.e., $\mathbf{bg}_w(p)$) because it is not a causal law but rather a value of a causal variable. Similarly for the causal background of $\neg$**angry** in w. The upshot is that the counterfactual testing ground of **angry** (i.e., $\mathbf{cfg}_w(\mathbf{angry})$) is not confined to **trigger**-worlds, so within the counterfactual testing ground the deontically best worlds are ones where the Dalai Lama is not angry (and **trigger** is false). Therefore, 'the Dalai Lama should not be angry' is predicted as true.

What I have shown is that adopting an interventionist approach to counterfactuals in fleshing out the semantics of *ought* (compare (32) to (20)) provides enough flexibility to explain why (67b) is true. What I have not shown is whether the counterfactual semantics is flexible to the extent that it can pay attention to the values of some causal variables while disregarding others. My understanding of Kratzer's modal semantics is that it does not impose a restriction on picking up relevant facts of the world. So in order to match the standard account in this respect, the counterfactual semantics would have to allow paying partial attention to the values of the causal variables. Ciardelli, Zhang & Champollion's (2018) experimental results are not well-suited to test whether counterfactuals can be interpreted this way, and I am afraid that carrying out relevant experiments is not within the scope of this paper.

Independently of whether counterfactuals can pay partial attention to the values of the causal variables, I am not certain at the moment whether there is evidence for preferring such degree of flexibility in deontic reasoning, or whether it better come with certain restrictions (e.g., either consider the value of all causal variables or disregard them altogether). Either way, it seems clear that interventionist approaches have an edge over the similarity-based ones, at least when considering which theory to adopt in fleshing out a counterfactual-based theory of deontic modality.

9 Comparison with Carr (2014)

Concerning the self-frustrating *ought* examples, Carr (2014) points out that there is no natural way to tear apart the propositions 'you are in Aleppo' and 'Death is in Aleppo', because one is inferred from the other. Carr invokes causal decision theory, in which the causal expected utility of acts are compared. If you go to Aleppo, Death is already there, so the decision problem is defined as follows:

(69) Decision problem, given that Death is in Aleppo

	death-alp	**death-dms**
you-alp	−100	0
you-dms	0	−100

Under the assumption that Death is a perfect predictor of your whereabouts, the causal expected utility of the decisions to be compared are calculated in (70). For brevity, I will respectively shorten **you-alp** and **you-dms** to **a** and **d**, and **death-aleppo** and **death-damascus** to **A** and **D**.

(70) a. $\mathrm{EU}(\mathbf{d} \mid \mathbf{a})$
$= \Pr(\mathbf{d} \mathrel{\square\!\!\rightarrow} \mathbf{A} \mid \mathbf{a})\mathrm{U}(\mathbf{Ad}) + \Pr(\mathbf{d} \mathrel{\square\!\!\rightarrow} \mathbf{D} \mid \mathbf{a})\mathrm{U}(\mathbf{Dd})$
$= \Pr(\mathbf{d} \mathrel{\square\!\!\rightarrow} \mathbf{A})\mathrm{U}(\mathbf{Ad}) + \Pr(\mathbf{d} \mathrel{\square\!\!\rightarrow} \mathbf{D})\mathrm{U}(\mathbf{Dd}) = 0$

b. $\mathrm{EU}(\mathbf{a} \mid \mathbf{a})$
$= \Pr(\mathbf{a} \mathrel{\square\!\!\rightarrow} \mathbf{A} \mid \mathbf{a})\mathrm{U}(\mathbf{Aa}) + \Pr(\mathbf{a} \mathrel{\square\!\!\rightarrow} \mathbf{D} \mid \mathbf{a})\mathrm{U}(\mathbf{Da})$
$= \Pr(\mathbf{a} \mathrel{\square\!\!\rightarrow} \mathbf{A})\mathrm{U}(\mathbf{Aa}) + \Pr(\mathbf{a} \mathrel{\square\!\!\rightarrow} \mathbf{D})\mathrm{U}(\mathbf{Da}) = -100$

What is crucial here is the equivalence of certain conditional probabilities in causal decision theory. For instance, $\Pr(\mathbf{d} \mathrel{\square\!\!\rightarrow} \mathbf{A} \mid \mathbf{a}) = \Pr(\mathbf{d} \mathrel{\square\!\!\rightarrow} \mathbf{A})$ and $\Pr(\mathbf{d} \mathrel{\square\!\!\rightarrow} \mathbf{D} \mid \mathbf{a}) = \Pr(\mathbf{d} \mathrel{\square\!\!\rightarrow} \mathbf{D})$. The equivalence holds because the location of Death is not causally affected by your location.

Carr also develops a premise semantics-based account and argues that the ordering source supplied to *ought* needs to be information sensitive. In order to implement information sensitivity, the ordering source is defined as a function that takes a newly introduced *deontic information parameter* i, and returns an ordering updated with the information i.[12] Carr's *domain* function can be understood as an extension of the BEST operator which additionally takes the deontic information parameter i into account.

(71) $\mathit{domain}(w, f, g, i)$ is the set of worlds in the modal background f ranked highest by the ordering $g(i)$.[13]

12 Given how Carr defines the domain of quantification (example (71)), my understanding is that all ordering sources, including the stereotypical ordering source supplied to an epistemic modal, takes the deontic information parameter. I speculate that the stereotypical ordering source is a constant function that is unaffected by the deontic information parameter.

13 The deontic ordering source seems to return an ordering on worlds in Carr's formulation. Precisely speaking, what needs to be returned in premise semantics is a set of propositions corresponding to the ideals of the world. As for the self-frustrating *ought* exam-

(72) 'ought p' is true at $\langle w, f, g, i\rangle$ iff $\forall w' \in domain(w, f, g, i) : \ p(w') = 1$,
where f is a modal base, g is an ordering source, and i is a deontic information parameter

Carr's interpretation of 'if p, ought p' is provided in (73). The *if*-clause updates the epistemic modal base e and the deontic information parameter i, but not the circumstantial modal base supplied to *ought*. The deontic ordering source updated with the information p (notation: $d(i + p)$) is supplied to *ought*, and the best worlds with respect to $d(i + p)$ are selected. In the self-frustrating *ought* scenario, if you are in Aleppo, the deontic ordering source updated with **you-alp** ranks **you-dms**-worlds higher than **you-alp**-worlds. Consequently, the deontically best worlds are **you-dms**-worlds.

(73) 'if p, ought p' is true at $\langle w, e, c, s, d, i\rangle$ iff
$\forall w' \in domain(w, e + p, s, i + p) :$
$\forall w'' \in domain(w', c, d, i + p) : \ p(w'') = 1$,
where e is an epistemic modal base, s is a stereotypical ordering source, c is a circumstantial modal base, d is a deontic ordering source, and i is a deontic information parameter

My understanding of Carr's premise semantics-based account is that it offers an alternative way of capturing conditional obligation, that is, one without restricting the circumstantial modal base. In Carr's analysis, the indicative conditional first quantifies over the epistemically best **you-alp**-worlds, say w'. At this point, the additional expressive power due to positing an information-sensitive ordering source gives Carr two options to make 'you ought not to go to Aleppo' true in w'. The first option is to consider **death-alp** as circumstantially relevant in w'. However, Carr claims that doing so would also bring in **you-alp** because the former was inferred from the latter, and the resulting semantics would be trivial. This leads Carr to take the second option, which is to assume that **death-alp** is not circumstantially relevant in w' but the very same information updates the deontic information parameter so that the ideals of w' are conditioned on Death's presence in Aleppo.[14]

ples, I suppose that the relevant ordering can be derived by defining $d(i +$ **you-alp**$)$ as $d(i) \cup \{$**you-dms**$\}$, and $d(i +$ **you-dms**$)$ as $d(i) \cup \{$**you-alp**$\}$, respectively.

14 Technically speaking, **you-alp** updates the deontic information parameter rather than **death-alp** in Carr's analysis. However, what eventually makes the agent prefer not going to Aleppo

The semantics proposed in this paper is more in the spirit of the standard account in that conditional obligation is captured by restricting the modal base. What Carr claimed to be an incoherent configuration of background assumptions is made coherent via counterfactual reasoning, which allows us to ignore the dependency between **you-alp** and **death-alp**. There is also an advatange to adopt the counterfactual-based semantics: it offers an explanation of why a single modal construal of 'if p, ought p' is not available. Carr's analysis assumes that the single modal construal is available but is ignored because it yields trivial interpretation, but the proposed semantics categorically blocks such a reading because it causes a clash between the counterfactual antecedent and the restrictor (see section 7.2).[15]

As for Carr's causal decision theory-based account, it is difficult to make a comparison with my account due to the quantitative nature of the former. However, I would like to note that the proposed semantics captures the same intuition as causal decision theory: the counterfactual assumption that you are in Aleppo (or Damascus) does not affect Death's whereabouts. Causal decision theory reflects this intuition by equating the two probabilities, Pr(**d** $\square\!\!\rightarrow$ **A** | **a**) and Pr(**d** $\square\!\!\rightarrow$ **A**). On the other hand, the proposed semantics ignores the dependency between **you-alp** (**a**) and **death-alp** (**A**) and fixes Death's location to Aleppo when constructing the circumstantial modal base of *ought*. For this reason, despite the lack of quantitative reasoning, the underlying mechanism of the proposed semantics is more similar to causal decision theory than to Carr's premise semantics-based solution.

is the information that Death is in Aleppo, so **death-alp** is effectively updating the deontic information parameter.

15 An anonymous reviewer asks how the proposed semantics can deal with the miners puzzle (Kolodny & MacFarlane 2010), which has been taken as evidence for information-sensitive ordering sources. I admit that the proposed semantics as it stands cannot handle the miners puzzle, but I speculate that it can be extended along the lines of Cariani, Kaufmann & Kaufmann (2013). Cariani et al. introduce a decision problem parameter to partition the modal background of *ought*. They utilize a deliberative preference ranking to select the best partition within the modal background, and this allows them to account for the miners puzzle without utilizing an information-sensitive ordering source. Given that the proposed semantics (i) defines the modal background of '*ought* p' as the union of the counterfactual p-worlds and the counterfactual alternative-to-p-worlds and (ii) assumes that that the alternatives are mutually exclusive, each counterfactual worlds can naturally be understood as a cell of the decision problem. The proposed semantics can then reproduce the same results as Cariani et al. by replacing the deontic ordering source with a deliberative preference ranking.

10 Conclusion

This paper proposes that deontic modals involve counterfactual reasoning, picking out the best worlds among the counterfactual prejacent-worlds and the counterfactual alternative-to-the-prejacent-worlds. As a side effect, law-like dependencies between the prejacent and facts that causally affect the prejacent are ignored in interpreting *must/should/ought*. The upshot for the theory of modality is that for intuitively true cases of 'if p, must/should/ought p' discussed in Zvolenszky 2002 and Carr 2014, we can set up a coherent modal background of the deontic modal. The task has been known to be difficult without making certain unnatural assumptions about the worlds under consideration. The issue was that given the widely accepted assumption that certain facts stand and fall together, it is unclear how to construct a modal background where p is an open possibility but the facts that are inferred from p are necessarily true. By hypothesizing that deontic modals are inherently counterfactual, the proposed semantics motivates the construction of such a modal background and offers a systematic solution to the 'if p, ought p' problem. In addition, it offers a principled explanation of why the 'if p, ought p' examples do not have a single modal construal, circumventing the overgeneration issue.

References

Arregui, Ana. 2005. Layering modalities: The case of backtracking counterfactuals. Unpublished manuscript. https://anaarreguidotcom1.files.wordpress.com/2014/08/layering-modalities.pdf.

Arregui, Ana. 2011. Counterfactual-style revisions in the semantics of deontic modals. *Journal of Semantics* 28(2). 171–210. https://doi.org/10.1093/jos/ffq017.

Bennett, Jonathan. 2003. *A philosophical guide to conditionals*. Oxford: Oxford University Press. https://doi.org/10.1093/0199258872.001.0001.

Briggs, R. A. 2012. Interventionist counterfactuals. *Philosophical studies* 160(1). 139–166. https://doi.org/10.1007/s11098-012-9908-5.

Cariani, Fabrizio, Magdalena Kaufmann & Stefan Kaufmann. 2013. Deliberative modality under epistemic uncertainty. *Linguistics and Philosophy* 36(3). 225–259. https://doi.org/10.1007/s10988-013-9134-4.

Carr, Jennifer. 2014. The *If P, Ought P* problem. *Pacific Philosophical Quarterly* 95(4). 555–583. https://doi.org/10.1111/papq.12048.

Chung, WooJin. 2019. Decomposing deontic modality: Evidence from Korean. *Journal of Semantics* 36(4). 665–700. https://doi.org/10.1093/jos/ffz016.

Ciardelli, Ivano, Linmin Zhang & Lucas Champollion. 2018. Two switches in the theory of counterfactuals. *Linguistics and Philosophy* 41(6). 577–621. https://doi.org/10.1007/s10988-018-9232-4.

Condoravdi, Cleo. 2002. Temporal interpretation of modals: Modals for the present and for the past. In David I. Beaver, Luis D. Casillas, Brady Z. Clark & Stefan Kaufmann (eds.), *The construction of meaning*, 59–88. Stanford: CSLI Publications.

von Fintel, Kai. 2012. *The best we can (expect to) get? Challenges to the classic semantics for deontic modals*. Paper presented at the APA Central, February 17. http://web.mit.edu/fintel/fintel-2012-apa-ought.pdf.

von Fintel, Kai & Sabine Iatridou. 2017. *X-marked desires: What wanting and wishing cross-linguistically can tell us about the ingredients of counterfactuality*. Slides from a talk at PhLiP Workshop, Tarrytown, NY. https://web.mit.edu/fintel/ks-x-phlip-slides.pdf.

Frank, Anette. 1997. *Context dependence in modal constructions*. Stuttgart University dissertation. https://www.cl.uni-heidelberg.de/~frank/papers/header.pdf.

Geurts, Bart. 2004. On an ambiguity in quantified conditionals. Unpublished manuscript. https://pdfs.semanticscholar.org/3229/e325293d8fb47ab7b547ed473941de387618.pdf.

Gibbard, Allan & William L. Harper. 1978. Counterfactuals and two kinds of expected utility. In Clifford Alan Hooker, James J. Leach & Edward Francis McClennen (eds.), *Foundations and applications of decision theory*, 125–162. Dordrecht: Springer. https://doi.org/10.1007/978-94-009-9789-9_5.

Halpern, Joseph Y. 2013. From causal models to counterfactual structures. *The Review of Symbolic Logic* 6(2). 305–322. https://doi.org/10.1017/S1755020312000305.

Heim, Irene. 1992. Presupposition projection and the semantics of attitude verbs. *Journal of semantics* 9(3). 183–221. https://doi.org/10.1093/jos/9.3.183.

Hiddleston, Eric. 2005. A causal theory of counterfactuals. *Noûs* 39(4). 632–657. https://doi.org/10.1111/j.0029-4624.2005.00542.x.

Jackson, Frank. 1985. On the semantics and logic of obligation. *Mind* 94(374). 177–195. https://doi.org/10.1093/mind/XCIV.374.177.

Jackson, Frank & Robert Pargetter. 1986. Oughts, options, and actualism. *The Philosophical Review* 95(2). 233–255. https://doi.org/10.2307/2185591.

Kaufmann, Stefan. 2013. Causal premise semantics. *Cognitive Science* 37(6). 1136–1170. https://doi.org/10.1111/cogs.12063.

Kaufmann, Stefan. 2017. The limit assumption. *Semantics and Pragmatics* 10(18). https://doi.org/10.3765/sp.10.18.

Kolodny, Niko & John MacFarlane. 2010. Ifs and oughts. *The Journal of Philosophy* 107(3). 115–143. https://doi.org/10.5840/jphil2010107310.

Kratzer, Angelika. 1981a. Partition and revision: The semantics of counterfactuals. *Journal of Philosophical Logic* 10(2). 201–216. https://doi.org/10.1007/BF00248849.

Kratzer, Angelika. 1981b. The notional category of modality. In Hans J. Eikmeyer & Hannes Rieser (eds.), *Words, worlds, and contexts: New approaches to word semantics*, 38–74. Berlin: Walter de Gruyter. https://doi.org/10.1515/9783110842524.

Kratzer, Angelika. 1991a. Conditionals. In Arnim von Stechow & Dieter Wunderlich (eds.), *Semantics: An international handbook of contemporary research*, 651–656. Berlin: de Gruyter. https://doi.org/10.1515/9783110126969.

Kratzer, Angelika. 1991b. Modality. In Arnim von Stechow & Dieter Wunderlich (eds.), *Semantics: An international handbook of contemporary research*, 639–650. Berlin: de Gruyter. https://doi.org/10.1515/9783110126969.

Kratzer, Angelika. 2012. *Modals and conditionals: New and revised perspectives*. Oxford, UK: Oxford University Press.

Lassiter, Daniel. 2017. Probabilistic language in indicative and counterfactual conditionals. *Semantics and Linguistic Theory (SALT) 27*. 525–546. https://doi.org/10.3765/salt.v27i0.4188.

Lewis, David. 1973. *Counterfactuals*. Oxford: Oxford University Press. https://doi.org/10.2307/2215339.

Lewis, David. 1979. Counterfactual dependence and time's arrow. *Noûs* 13(4). 455–476. https://doi.org/10.2307/2215339.

Ogihara, Toshiyuki. 2014. The semantics of *-ta* in japanese future conditionals. In Luka Crnič & Uli Sauerland (eds.), *The art and craft of semantics: A festschrift for irene heim*, vol. 2, 1–21. Cambridge, MA: MITWIPL.

Pearl, Judea. 2000. *Causality: Models, reasoning, and inference*. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511803161.

Portner, Paul. 2009. *Modality*. Oxford, UK: Oxford University Press.

Rubinstein, Aynat. 2017. Straddling the line between attitude verbs and necessity modals. In Ana Arregui, María Luisa Rivero & Andrés Salanova (eds.), *Modality across syntactic categories*, 610–633. https://doi.org/10.1093/acprof:oso/9780198718208.003.0007.

Santorio, Paolo. 2019. Interventions in premise semantics. *Philosophers' Imprint* 19(1). 1–27. http://hdl.handle.net/2027/spo.3521354.0019.001.

Schulz, Katrin. 2011. If you'd wiggled A, then B would've changed. *Synthese* 179(2). 239–251. https://doi.org/10.1007/s11229-010-9780-9.

Spohn, Wolfgang. 1975. An analysis of Hansson's dyadic deontic logic. *Journal of Philosophical Logic* 4(2). 237–252. https://doi.org/10.1007/BF00693275.

Stalnaker, Robert C. 1968. A theory of conditionals. *Studies in Logical Theory*. 98–112. https://doi.org/10.1007/978-94-009-9117-0_2.

van Fraassen, Bas C. 1972. The logic of conditional obligation. *Journal of Philosophical Logic* 1(3–4). 417–438. https://doi.org/10.1007/BF00255570.

Villalta, Elisabeth. 2008. Mood and gradability: An investigation of the subjunctive mood in Spanish. *Linguistics and Philosophy* 31(4). 467. https://doi.org/10.1007/s10988-008-9046-x.

Werner, Tom. 2006. Future and non-future modal sentences. *Natural Language Semantics* 14(3). 235–255. https://doi.org/10.1007/s11050-006-9001-8.

Williams, Bernard. 1981. Ought and moral obligation. In *Moral luck*, 114–123. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9781139165860.010.

Zvolenszky, Zsófia. 2002. Is a possible-worlds semantics of modality possible? A problem for Kratzer's semantics. *Semantics and Linguistic Theory (SALT) 12*. 339–358. https://doi.org/10.3765/salt.v12i0.2866.

WooJin Chung
Institut Jean Nicod
École Normale Supérieure
29 Rue d'Ulm
75005 Paris, France
woojin@nyu.edu