Expressible semantics for expressible counterfactuals*

Emmanuel Chemla (IJN/LSCP) — September 19, 2010

Abstract

Lewis (1981) showed the equivalence between two dominant semantic frameworks for counterfactuals: *ordering semantics*, which relies on orders between possible worlds, and *premise semantics*, which relies on sets of propositions (so-called ordering sources). I define a natural, restricted version of premise semantics, *expressible premise semantics*, which is based on ordering sources containing only *expressible* propositions. First, I extend Lewis' (1981) equivalence result to expressible premise semantics and some corresponding expressible version of ordering semantics. Second, I show that expressible semantics are strictly less powerful than their non-expressible counterparts, even when attention is restricted to the truth values of expressible counterfactuals. Assuming that the expressibility constraint is natural for premise semantics, this result breaks the equivalence between ordering semantics and (expressible) premise semantics. Finally, I show that these results cast doubt on various desirable conjectures, and in particular on a particular defense of the so-called limit assumption.

Keywords: counterfactuals; ordering semantics; premise semantics; expressive power; limit assumption; relevance

1 Constraining dominant semantics for counterfactuals

1.1 Unrestricted ordering and premise semantics

(1) If Mary had set her alarm clock right, she would not have been late.

Counterfactual sentences like (1) are statements about how some fact of the world would be altered or unaltered along with some imaginary change — how Mary's time schedule would have been influenced by her setting her alarm clock right, in the case of (1). Intuitively, a counterfactual sentence is true if the consequent is true in the hypothetical worlds in which the antecedent is true (contrary to fact) but which are otherwise as similar as possible to the actual world. For instance, (1) is judged true if, assuming that Mary did set her alarm clock right, it is natural to assume as well that she would have heard it,

*Many thanks are due to Nathan Klinedinst, who first guided me into thinking about these issues, to Benjamin Spector for numerous challenging discussions, which led to the proofs based on the existence of an optimally compliant order, to Paul Egré for generous comments on earlier versions of this work, and to Denis Bonnay, Angelika Kratzer, Daniel Rothschild, Philippe Schlenker and Robert Stalnaker. Mistakes are unquestionably mine. This work was supported by a 'Euryi' grant from the European Science Foundation ("Presupposition: A Formal Pragmatic Approach").

would have woken up quickly and caught the 8am subway, that this new situation would not have led Mary's train to be overloaded and stop working properly at some subway station A or B, etc.

Most prominently, Lewis and Stalnaker offered an explicit formalization of this intuition based on an explicit order of the universe of possible worlds. In this framework of *ordering semantics*, the order is thus designed to model the relative similarity between various states of affairs and the actual world (Stalnaker 1968, Lewis 1973). The idea is that a world x is considered "smaller" than another world y if x is *closer* to the actual world, i.e. if the actual world differs less with x than it differs with y. Back to the opening example, (1) would be judged true if worlds in which Mary sets her alarm clock and catches her train alright are given as "smaller" than worlds in which she sets her alarm clock right but gets stuck in a defective train.

In the tradition of *premise semantics*, such an ordering relation between worlds is derived from an *ordering source*, i.e. a set of propositions which are true in the actual world (Kratzer 1979, 1981, see also Ginsberg 1986). Intuitively, an ordering source contains propositions which ought to be true for a state of affairs to be judged similar to the actual state of affairs: the more propositions from the ordering source hold in a world x, the more similar x is to the actual world (e.g., the ordering source may contain "Mary is able to hear her alarm clock when it rings", "the 8am train runs on time", etc.). One may also see an ordering source as the stock of premises available to justify a counterfactual statement. Concretely, (1) is judged true if one can construct an argument to the conclusion that 'Mary does not arrive late' by assuming that 'Mary set her alarm clock right' (the antecedent of (1)) and combining this assumption with other compatible premises chosen in the ordering source.

Premise semantics seems to take us one step further in the understanding of the order on worlds necessary to evaluate counterfactuals: the order postulated in ordering semantics is determined by a set of propositions which describe essential properties of the actual situation. Moreover, there is no formal advantage in using orders from ordering semantics rather than ordering sources from premise semantics: any Lewis/Stalnaker order can be obtained from an ordering source *à la* Kratzer and *vice-versa* (see Lewis 1981). Consequently, any consistent distribution of truth values over a set of counterfactuals can be obtained equivalently from an order or from an ordering source.

1.2 The human factor

Orders and ordering sources are constructed to account for *speakers'* semantic intuitions. This human factor should naturally translate into constraints on the possible forms of these devices. Most clearly, an ordering source would lose some of its *raison d'être* if it were made of non-expressible propositions, i.e. propositions which cannot be manipulated or entertained by a speaker. Accordingly, I will introduce a version of premise semantics, *expressible premise semantics*, in which ordering sources may only contain expressible propositions.

We may reduce the expressive power of premise semantics by constraining it to use only expressible ordering sources. However, this loss should be immaterial as far as only *expressible* counterfactuals are concerned. Hence, it seems that expressible premise semantics could efficiently and straightforwardly replace unconstrained versions of premise semantics (or ordering semantics). Let me summarize this as the following conjecture:

(2) **Conjecture 1:** An ordering source can be replaced with an *expressible* ordering source while preserving the truth values of the accordingly *expressible* counterfactuals.

My main goal is to show that this appealing and desirable conjecture is not correct.

1.3 Goals and results

The first result of this work is a refinement of Lewis's equivalence result between premise and ordering semantics. It shows how the constraints described for premise semantics translate in the framework of ordering semantics to yield two equivalent frameworks: expressible premise semantics and expressible ordering semantics (Result 14, §2.2).

The second set of results target conjecture 1 above more directly. They question the existence of a process which would turn any order or ordering source into an expressible order or ordering source leading to the same truth value for every expressible counterfactual. I first exhibit a procedure which *maximally* satisfies this requirement (§3). However, "maximally" is not enough, and, as was announced above, this procedure fails in the general case (§4). The situations in which there is no fully satisfying simplification process are characterized exhaustively in §4.2.

The consequences of these results are discussed in §5 from a more general perspective. In particular, the formal results remain identical if we reinterpret the constraining set of expressible propositions $\mathcal{L}$ as a set of *relevant* expressible propositions. Consequently, the following desirable and apparently reasonable conjectures below may not be sustainable, which is problematic for a certain defense of the so-called limit assumption (see §5).

(3) **Conjecture 2:** There are indefinitely many differences between possible worlds and most of these differences are irrelevant to evaluate the truth value of a given counterfactual. Hence, it is not parsimonious to assume that every possible difference is taken into account at every utterance of any counterfactual. More plausibly, irrelevant differences are straightforwardly disregarded. In other words, a simplified order is probably used to evaluate a given counterfactual, a context-dependent order which does not distinguish worlds whose differences are irrelevant. Hence, there should exist a process by which an order is turned into an "expressible relevant order" *without altering the truth conditions of the relevant counterfactuals*.

(4) **Conjecture 3:** Conjecture 2 above suggests that semantic mechanisms could be simplified on a case by case basis by focussing on a subset of relevant properties, while preserving the truth value of the relevant counterfactuals. This type of issue arises explicitly in discussions about the so-called limit assumption: given some proposition P, can we find a P-world such that there is no P-world closer to the actual world? Proponents of the limit assumption have argued that it is not a problematic assumption if only we have a way to restrict our attention to a (finite) subset of worlds — or equivalently to disregard various small differences between worlds — so that only a finite number of selected classes of worlds emerge.

2 Definitions and preliminary results

2.1 Ordering semantics and premise semantics: general definitions and comparison

Consider a universe of worlds W with a distinguished actual world and a strict order $<$ of W such that no world is smaller than the actual world.[1] The following semantics for counterfactuals is classically derived from the order $<$ (ordering semantics):

(5) **Definition (ordering semantics):**
"If P, Q" is true iff $\forall x \in P : \exists y \leqq x : y \in P \wedge Q$ and $\forall z < y : z \in \neg P \vee Q$.

Alternatively, one may start with an "ordering source" $\mathcal{O}$ — i.e. a set of propositions — to obtain another type of semantics for counterfactuals (premise semantics):

[1] This definition does not require "centering", i.e. the actual world is not necessarily smaller than every other world. Formally, centering does not play much role (see Lewis 1981, §7) and it is convenient to avoid this constraint for the present purposes: it will allow us to assimilate orders of classes of indistinguishable worlds and orders of the original universe, even though the actual world may be clustered with other indistinguishable worlds.

(6) **Definition (premise semantics):**
"If P, Q" is true iff $\forall H \subseteq \mathcal{O}$ compatible with P, $\exists J : H \subseteq J \subseteq \mathcal{O}$ such that J is compatible with P and together with P, it entails Q.

Interestingly, the premise semantics obtained from an ordering source $\mathcal{O}$ is equivalent to the ordering semantics obtained from the order defined as in (7) below. This correspondence captures the original motivation for ordering sources: the propositions in the ordering source distinguish fundamental properties of the actual world and the more propositions of $\mathcal{O}$ a world satisfies, the closer it is to the actual world.

(7) $x < y$ iff $\mathcal{O}(x) \supset \mathcal{O}(y)$, where $\mathcal{O}(w) = \{P \in \mathcal{O} : w \in P\}$.

The mapping (7) from ordering sources to orders with equivalent semantics constitutes the first half of Lewis' (1981) result. The other half completes the equivalence between premise semantics and ordering semantics. It states that from an order $<$, the following ordering source leads to the same distribution of truth values:

(8) $O = \{H_w : w \in \mathrm{W}\}$, where $H_w = \{x : x < w$ or $x = w\}$ (i.e. $=\{x : x \leqq w\}$).

2.2 Expressible propositions and counterfactuals, expressible ordering sources and orders

As already illustrated in (8), a proposition can be seen as a subset of worlds of the universe, those worlds in which the proposition is meant to be true. Let us distinguish a subset $\mathcal{L}$ of the propositions as the *expressible* propositions. Concretely, $\mathcal{L}$ models the target language: it is the set of propositions which can be manipulated by a speaker (for some later applications, one may also think of $\mathcal{L}$ as the set of *relevant* propositions). I will assume throughout that $\mathcal{L}$ is closed under boolean operations, and in most cases I will assume closure under infinite boolean operations, see discussion in §2.3.

Accordingly, expressible counterfactuals can be defined as the counterfactuals "If P, Q" such that both P and Q are expressible. Notice that I may freely use the more explicit set notation for propositions even when these propositions are presented within counterfactuals. This notational convention legitimates apparently hybrid formulae of the form: "If $\{x, y, \ldots\}$, $\{x, \ldots\}$".

As discussed in the introduction, it is natural to define a restricted version of premise semantics, expressible premise semantics, in which ordering sources are required to be expressible in the following sense :

(9) **Definition (expressible ordering source):**
An ordering source is *expressible* iff it contains only expressible propositions.

A natural notion $\sim$ of indistinguishability between worlds can also be derived from the set of expressible propositions $\mathcal{L}$. Two worlds are indistinguishable if no expressible proposition is true in one and false in the other:

(10) **Definition (indistinguishability relation $\sim$):**
a. $x \sim y$ iff $\forall P \in \mathcal{L} : x \in P \leftrightarrow y \in P$
b. Equivalently: $x \not\sim y$ iff $\exists P \in \mathcal{L} : x \in P$ and $y \notin P$.[2]

(11) **Lemma:** The indistinguishability relation $\sim$ is an equivalence relation.

Proof: Reflexivity: No proposition can be true and false in the same world. Symmetry and transitivity follow from symmetry and transitivity of logical equivalence. □

I may refer to the indistinguishability equivalent class of a world x as $[x]$, i.e. $[x] = \{w : w \sim x\}$. I may also extend this notation to propositions (i.e. sets of worlds): $[P] = \bigcup\{[x] : x \in P\}$. The following lemma which relates indistinguishability back to expressibility will prove useful:

(12) **Lemma:** Assuming closure under infinite boolean operations, a proposition P is expressible iff it is closed under the indistinguishability relation, i.e. $P = [P]$.

Proof: $\boxed{\rightarrow:}$ No expressible proposition P may attribute different truth values to two worlds which are indistinguishable (by definition of the indistinguishability relation). $\boxed{\leftarrow:}$ Assume that $P = [P]$. For every pair of distinguishable worlds (x, y), there is an expressible proposition $P_{x,y}$ which is true in x and false in y. One can check that $P = \bigcap\{P_{x,y} : x \in P \text{ and } y \notin P\}$. Hence, P is an infinite conjunction of expressible propositions. □

With these tools in hand, we can define a restricted version of ordering semantics, expressible ordering semantics, which will be equivalent to expressible premise semantics as defined above (as it is proved in §2.3). An expressible ordering semantics is an ordering semantics based on an *expressible* order, i.e. an order which does not treat differently worlds which are otherwise indistinguishable:[3]

[2] The equivalence between (10a) and (10b) requires closure under negation of the language $\mathcal{L}$.

[3] This specific definition is rather liberal. In particular, an order $<$ might be labeled expressible, even though predicates of the form "smaller than x" are not expressible. This subtlety can be disregarded because it vanishes if the language is assumed to be finite or, more generally, closed under infinite boolean operations (see §2.3).

(13) **Definition (expressible order):**
An order $<$ is *expressible* iff $\forall x \sim y : \forall a : [(x < a \leftrightarrow y < a)$ and $(a < x \leftrightarrow a < y)]$.

2.3 Expressible premise semantics vs. expressible ordering semantics

The first main result of this work is the following refinement of Lewis's equivalence result between ordering semantics and premise semantics:

(14) **Result:**
Expressible ordering semantics and expressible premise semantics are equivalent. More precisely, assuming that the set of expressible propositions is closed under infinite boolean operations, an order $<$ is expressible iff it can be derived (in the sense of (7)) from an expressible ordering source $\mathcal{O}$.

Proof: $\boxed{\leftarrow:}$ Let $\mathcal{O}$ be an expressible ordering source leading to $<$. Two indistinguishable worlds x and y are ordered similarly with respect to any world because they necessarily satisfy the same expressible propositions of $\mathcal{O}$ (i.e. $\mathcal{O}(x) = \mathcal{O}(y)$, see (7)).
$\boxed{\rightarrow:}$ Starting from an expressible order $<$, I will imitate Lewis's construction of an ordering source (see (8)). First notice that for every world x, $H^*_x = \{w : w \lesssim x\}$ is expressible (where by definition, $x \lesssim y$ iff $x < y$ or $x \sim y$). This is because the order is expressible and thus $H^*_x = [H^*_x]$ (see lemma (12)).
Now, let $\mathcal{O} = \{H^*_x : x \in \mathrm{W}\}$.[4] $\mathcal{O}$ is expressible and the order derived from it is $<$: $\boxed{\text{i}}$ Assume $x < y$. Let $H^*_z \in \mathcal{O}(y)$, then either $\underline{\text{(a) } y < z}$ and then $x < z$ by transitivity, or $\underline{\text{(b) } z \sim y}$ and then $x < y \sim z$ because the order is expressible. In both cases we obtain $H_z \in \mathcal{O}(x)$. Hence $\mathcal{O}(x) \supseteq \mathcal{O}(y)$. Furthermore, the inclusion is strict because H^*_x is in the first one and not in the second. $\boxed{\text{ii}}$ Conversely, assume $\mathcal{O}(x) \supset \mathcal{O}(y)$. The difference between the two sets immediately entails that $x \not\sim y$. The fact that $x < y$ follows from $x \in H^*_y \in \mathcal{O}(y)$. □

The assumption that the language is closed under boolean operations is not particularly problematic. Notice however that closure under *infinite* boolean operations is necessary:

(15) **Side result:** If the language $\mathcal{L}$ is not closed under infinite boolean operations, some expressible order may not derive from any expressible ordering source.

Proof: Assume that $\mathcal{L}$ is not closed under infinite boolean operations. There is an infinite set of expressible propositions $\{P_n : n \in \mathrm{N}\}$ such that $P_\infty = \bigcap P_n \notin \mathcal{L}$. (If

[4] The only difference with Lewis's ordering source as in (8) is the difference between H^*_x and H_x where the symbol $\lesssim$ replaces $\leqq$.

it is an infinite *disjunction* which is not expressible, there is also a non-expressible infinite conjunction: the infinite conjunction of the *negations* of the P_is.) Assume that P_0 is not tautologous, so that we can find $x \notin P_0$ and $y \in P_0 - P_\infty$. Consider the order defined as follows: (i) the worlds in P_∞ are the smallest, (ii) worlds in $[x]$ are immediately above, (iii) worlds in $[y]$ are immediately above, (iv) other worlds are above. Schematically: $P_\infty < [x] < [y] <$ *all other worlds*.

$\boxed{< \text{ is expressible:}}$ For any $a < b$, by considering the various cases ($a \in P_\infty$ and $b \in [x]$, or $a \in P_\infty$ and $b \in [y]$, ...) one can see that $a \not\sim b$. The easiest way to prove so is to check that for $S = P_\infty$, $S = [x]$, $S = [y]$, or $S =$ *all other worlds*, then $[S] = S$.

$\boxed{< \text{ may not derive from any expressible ordering source:}}$ Assume for *reductio* that $\mathcal{O}$ is an expressible ordering source which leads to $<$. Then, because $x < y$, there is a proposition $Q \in \mathcal{O}$ such that (a) $x \in Q$ and (b) $y \notin Q$ (and similarly for worlds in their respective equivalence class since Q is expressible). Furthermore, just like any other proposition in $\mathcal{O}$: (c) $a < b$ and $b \in Q$ entail that $a \in Q$. From (a) and (c) we obtain that $P_\infty \subseteq Q$, and from (b) and the contraposition of (c), we obtain that $\{$*all other worlds*$\} \cap Q = \emptyset$. Hence, $Q = P_\infty \cup [x]$. But then it follows that P_∞ is expressible as it is the *finite* conjunction of expressible propositions: $P_\infty = Q \cap P_0$. $\Box$

Yet, this assumption of infinite closure does not affect the conclusions of this work. First, notice that this assumption is necessary only because we relied on a rather liberal definition of an expressible order (see footnote 3). Second and more importantly, this assumption is vacuous when the language is finite. The following results mostly rely on counterexamples, and counterexamples will be exhibited both for infinite and finite languages.

3 Constructing an optimally compliant order

The main topic of this paper is expressible premise semantics. Given the equivalence result from the previous section, I will frame most of the discussions with the more usual and convenient terminology of ordering semantics.

The main goal of this paper is to try to replace non-expressible orders with expressible orders. This constraint will be counterbalanced by our focussing attention on *expressible* counterfactuals, i.e. counterfactuals whose antecedents and consequents are expressible. In that respect, two orders are interchangeable for our purposes independently of the truth value they attribute to non-expressible counterfactuals. This notion can be formalized as follows:

(16) **Definitions:**

Two orders $<$ and $\overset{*}{<}$ are *compliant* if they yield the same truth values for every expressible counterfactual.

By extension, an order $\overset{*}{<}$ is *compliant* with a list δ of truth values over counterfactuals if $\overset{*}{<}$ yields the truth values given in δ for all expressible counterfactuals.

As a shortcut, I will simply say that an order is *compliant* whenever context makes clear which is the order $<$, or the list of truth values δ of reference.

The main target of this paper is to exhibit the conditions under which we can associate an expressible compliant order $\overset{*}{<}$ with a given order $<$. The main result of this section is in §3.2:(18/19) where I exhibit a systematic procedure to construct an *optimally* compliant order $\overset{*}{<}$, i.e. an expressible order which is compliant whenever it is possible.

3.1 Distribution of truth values

A semantics for counterfactuals is primarily designed to yield a distribution of truth values over all counterfactuals. Valid inference rules can be seen as constraints on the structure of this list of true and false counterfactuals enforced by the underlying machinery, e.g., orders. For instance, it cannot be that the list of true counterfactuals contains both "If P, Q_1" and "If P, Q_2" without containing "If P, $Q_1 \wedge Q_2$". Practically, if the former are known to be true, there is no need to specify that the latter is true as well. Conversely, if the list of truth values respects all necessary constraints, we should be able to uncover the underlying machinery that led to this particular distribution of truth values.

More specifically, the whole distribution of truth values is constrained by the truth values of *primitive* counterfactuals of a particular format, namely counterfactuals whose consequent is a single world, and whose antecedent is a pair of worlds. In essence, this result reveals that there are valid inference rules which constrain the truth value of any counterfactual from the truth values of these primitive counterfactuals alone.

(17) **Lemma:** In ordering or premise semantics, the truth value of any counterfactual is fully determined by the truth values of the *primitive* counterfactuals of the form: "If $\{a, b\}$, $\{a\}$" (a and b being any worlds).

Proof: Assume that $<$ is an order on a universe W. Notice that $a < b$ iff $a \neq b$ and "If $\{a, b\}$, $\{a\}$" is true (see (5)). Hence, the underlying order $<$ can be recovered from the true counterfactuals of the form "If $\{a, b\}$, $\{a\}$". The truth values of these counterfactuals thus constrain the truth value of every other counterfactual. □

3.2 An optimally compliant order

The goal of this paper is to investigate procedures which transform ordering sources into compliant expressible ordering sources. The results obtained so far show that 1) this question can be formulated equivalently in terms of ordering sources or in terms of orders [Result (14)] and 2) we should be able to focus on a sublist of primitive counterfactuals [Lemma (17)].

Concretely, we may start from an order $<$, or directly from the list of truth values it attributes to counterfactuals. We can first define a new expressible order $\overset{*}{<}$:

(18) **Lemma:**
Starting from the distribution of truth values for counterfactuals, the relation $\overset{*}{<}$ defined below is a strict expressible order:
$x \overset{*}{<} y$ iff $x \not\sim y$ and "If $[x] \cup [y]$, $[x]$" is true

Proof: We need to check that $\overset{*}{<}$ is transitive (expressibility and irreflexivity: by construction). Even though there is no general valid inference rule which guarantees such a result, transitivity will come out in this particular case because we are concerned with disjoint propositions (equivalent classes). Specifically, we need to prove that when $A \cap B$ and $B \cap C$ are empty, (i) "If $A \cup B$, A" and (ii) "If $B \cup C$, B" entail (iii) "If $A \cup C$, A". Assuming (i) and (ii), let us thus show that for any C- or A-world, there is a smaller "stable" B-world (in the sense of (5)).

<u>Let $w_C \in C$.</u> According to (ii), let $w_B \in B$ be smaller than w_C such that there is no smaller C-world (all smaller $B \cup C$-worlds are B-worlds). According to (i), let $w_A \in A$ be smaller than w_B such that there is no smaller B-world (all smaller $A \cup B$-worlds are A-worlds). This w_A is an A-world smaller than w_C (by transitivity) such that all smaller $A \cup C$-worlds are A-worlds.

<u>Let $w_A \in A$.</u> If there is a smaller C-world, apply the procedure above to such a world. If there is no smaller C-world, then all smaller $(A \cup C)$-worlds are A-worlds. □

Notational shortcut: I will qualify as $*$*true* or $*$*false* a counterfactual which is accordingly true or false under the semantics associated with the order $\overset{*}{<}$. Unless otherwise stated, the non-starified versions of *true* and *false* thus refer to the order $<$.

This order $\overset{*}{<}$ will be particularly useful in our quest for an expressible compliant order because it is *optimally* compliant in the following sense:

(19) **Result:**
If there is any compliant expressible order, the expressible order $\overset{*}{<}$ is compliant.

Proof: Assuming that there is a compliant expressible order, the only thing left to prove is that $\overset{*}{<}$ preserves the truth values of the crucial counterfactuals of the form "If $[a] \cup [b]$, $[a]$", with $a \not\sim b$. Indeed, in the target language $\mathcal{L}$ (which contains the expressible propositions), these counterfactuals are the primitive counterfactuals used in lemma (17) (classes of worlds in the target language $\mathcal{L}$ play the same role as worlds in the source language, $\mathcal{P}(\mathrm{W})$, the set of all propositions). Hence, we must prove that: "If $[a] \cup [b]$, $[a]$" is true with respect to $<$ if and only if it is true with respect to $\overset{*}{<}$. Given that classes of worlds play the same role in the target language as single worlds in the usual case, "If $[a] \cup [b]$, $[a]$" is *true iff $[a]\overset{*}{<}[b]$ and, by construction, this is true iff the counterfactual is true according to the original order $<$. $\square$

3.3 Other natural or compliant procedures?

I highlighted a particular procedure to construct an expressible order. This order $\overset{*}{<}$ is *optimally compliant*: whenever there is one, $\overset{*}{<}$ is a compliant expressible order. However, there are various other natural candidates that come to mind to transform directly any order or ordering source into an expressible order or ordering source. Here is a couple of them:

(20) Transform every proposition in the ordering source into a minimally different expressible one (e.g., we could replace any proposition P in $\mathcal{O}$ with its expressible counterpart $[P] = \{[x] : x \in P\}$, see lemma (12)).

(21) Drop all non-expressible propositions from the ordering source.

(22) Transform the ordering source to make it fit some property (e.g., closure under conjunction, see (23)) and then apply (20) or (21).

(23) **Remark:** Adding conjunctions to an ordering source does not affect the associated order (or the distribution of truth value for counterfactuals).

Proof: Let $\mathcal{O}$ be an ordering source, and $\hat{\mathcal{O}}$ the same ordering source augmented with some conjunctions (e.g., $\hat{\mathcal{O}}$ could be the closure under conjunction of $\mathcal{O}$). The associated orders are the same: $\mathcal{O}(x) \supset \mathcal{O}(y) \leftrightarrow \hat{\mathcal{O}}(x) \supset \hat{\mathcal{O}}(y)$. $\boxed{\rightarrow:}$ First, if P is in $\mathcal{O}(x) - \mathcal{O}(y)$, P is in $\hat{\mathcal{O}}(x) - \hat{\mathcal{O}}(y)$. Second, let $P \in \hat{\mathcal{O}}(y)$. P is the conjunction of the elements of a subset $H \subseteq \mathcal{O}$. $\forall Q \in H, Q \in \mathcal{O}(y) \subset \mathcal{O}(x)$. Hence, x is also in $\bigcap H = P$. $\boxed{\leftarrow:}$ First, let $P \in \mathcal{O}(y)$. $P \in \mathcal{O}(y) \subseteq \hat{\mathcal{O}}(y) \subset \hat{\mathcal{O}}(x)$, hence $P \in \mathcal{O}(x)$. Second, we also need to find a proposition in $\mathcal{O}(x) - \mathcal{O}(y)$. Let $P \in \hat{\mathcal{O}}(x) - \hat{\mathcal{O}}(y)$, i.e. $x \in P$ and $y \notin P$. P is a conjunction of elements of some subset $H \subseteq \mathcal{O}$, hence $x \in \bigcap H$ and $y \notin \bigcap H$. Hence there is $Q \in H$ such that $x \in H$ and $y \notin H$, i.e. $Q \in \mathcal{O}(x) - \mathcal{O}(y)$. $\square$

A priori, procedures such as (20-22) may sound more natural than (18) since they rely on direct manipulations of ordering sources, rather than referring to specific counterfactuals for instance. However, there are various concerns about this subjective preference. First, notice that given the problem, the non trivial cases for our discussion are those where the original ordering source is not expressible. Hence, there is no reason to assume that it is easier to manipulate this ordering source rather than the order it gives rise to, or than the truth values of a subset of expressible counterfactuals as is done above. Second, these procedures are unsteady in that they may yield different outcomes for different otherwise compliant ordering sources. Third, these procedures are not optimal: they may not preserve the truth values of the expressible counterfactuals (even in cases where there clearly is a way to do so).

Example (24) below illustrates the last two concerns. Specifically, this example shows that 1) the procedure (20) may yield a non-compliant order (even when there clearly is one) and 2) the procedure may yield a compliant order with different but equivalent ordering sources. The equivalent ordering sources considered in this example are: the original ordering source first closed under conjunction (as suggested in procedure (22)) and an ordering source *à la* Lewis (see (8)).

(24) **Example:** Consider the structure represented below: $x \sim x'$, $y \sim y'$, and both x and x' are smaller than y and y' [arrows like "$a \leftarrow b$" mean "$a < b$"]. The truth conditions derived from this order also derive from the following ordering source: $\mathcal{O} = \{\{x, x', y\}, \{x, x', y'\}\}$. Replacing every proposition P with its expressible counterpart $[P]$ leads to the non compliant ordering source $[\mathcal{O}] = \{\{x, x', y, y'\}\} = \{[x] \cup [y]\}$.
If $\mathcal{O}$ is first closed under conjunction ($\mathcal{O}^+ = \{\{x, x'\}, \{x, x', y\}, \{x, x', y'\}\}$) or if we start with a compliant ordering source *à la* Lewis ($\mathcal{O}^\mathcal{L} = \{\{x\}, \{x'\}, \{y, x, x'\}, \{y', x, x'\}\}$), we do obtain a compliant ordering source: $[\mathcal{O}^+] = [\mathcal{O}^\mathcal{L}] = \{\{x, x'\}, \{x, x', y, y'\}\} = \{[x], [x] \cup [y]\}$.

$$\begin{tikzcd} x & y \arrow[l] \arrow[dl] \\ x' & y' \arrow[l] \arrow[ul] \end{tikzcd}$$

Example (24) may suggest that the procedure (22) is optimally compliant: closing under conjunction an ordering source before transforming it into an expressible one as in (20) leads to a compliant order (at least when there is one). This is not true, as the following example shows.

(25) **Example:** Consider that there are three worlds x, x' and y with no world smaller than another and $x \sim x'$. This trivial order can be derived from the ordering source

$\mathcal{O} = \{\{x, y\}, \{x'\}\}$. $\mathcal{O}$ is closed under conjunction.[5] However, the expressible portion of it leads to $[\mathcal{O}] = \{\{x, x', y\}, \{x, x'\}\}$, which makes the counterfactual "If $\{x, x', y\}$, $\{x, x'\}$" true, while it was originally false.
Notice that the empty ordering source is compliant and expressible.

Let's take stock. There are various ways to obtain an expressible order or ordering source from a non-expressible order or ordering source. Which procedure one may choose depends on the ranking of two constraints. (i) Do we want to keep the semantics (truth values) constant? (ii) Do we want to work directly with the ordering source, on the (rather artificial) assumption that it is somewhat cognitively more accessible? Given our initial purposes, the first constraint clearly has priority. Hence, the procedure described in §3.2:(18/19) to recover a compliant expressible order $\overset{*}{<}$ is optimal because it guarantees to offer a solution, whenever there exists a solution. This optimal procedure will be used to describe counterexamples (§4.1) and characterize the situations where the answer to the following question is positive: is there a compliant expressible order? (§4.2)

4 Main result: no compliant order in general

This section contains the most technical proofs, but the end result is simple: contrary to expectations, there are situations in which there is no compliant expressible order. Section §4.1 presents such situations in various formats. Section §4.2 shows that all counterexamples are of the form described in §4.1. Interesting distinctions will appear depending on whether the target language $\mathcal{L}$ is finite or infinite.

4.1 Counterexamples: no compliant order

Let us start from a list of true counterfactuals and a set of expressible propositions. Technically, the following configurations will be proved to allow no expressible compliant order as follows: the optimal procedure described in §3.2:(18/19) does not deliver a compliant expressible order, hence there is none.

[5]The contradiction is omitted for simplicity, it does not play any role.

4.1.1 TRUE BECOMES FALSE (WITH A POSSIBLY FINITE LANGUAGE $\mathcal{L}$)

Let the universe be as pictured in example (26).[6] As before, arrows like "$a \leftarrow b$" mean "$a < b$". Hence, there are four worlds x, x', y_1 and y_2, with only two ordering relations between them: $y_1 < x$ and $y_2 < x'$.

(26)
$$\boxed{\begin{array}{l} y_1 \longleftarrow x \\ y_2 \longleftarrow x' \end{array}}$$

Assume now that x and x' are indistinguishable, and all other pairs of worlds are distinguishable. As a result, the proposition $\{x, x'\}$ is expressible, and no proposition containing exactly one of the two xs is. In this situation, the expressible counterfactual "If $\{x, x', y_1, y_2\}$, $\{y_1, y_2\}$" is true, but it comes out false with the order $\overset{*}{<}$ provided by the optimal procedure described above.

> **Proof:** All primitive expressible counterfactuals with an antecedent containing two classes of worlds, and a single class as a consequent are originally false. For instance, "If ($\{x, x'\} \cup \{y_1\}$), $\{y_1\}$" is false because from x' we cannot walk down to the world y_1. Hence, this situation gives rise to a trivial optimal order $\overset{*}{<}$: no two worlds come out ordered. As a result, the counterfactual "If ($\{x, x'\} \cup \{y_1\} \cup \{y_2\}$), $\{y_1\} \cup \{y_2\}$" comes out false as well: starting from x or from x' in the target universe, neither of y_1 or y_2 is smaller and thus (5) does not hold. □

This counterexample can be generalized. In short, the problem could be the same with more than two indistinguishable worlds (previously: x and x'), and potentially more than two distinguishable worlds below them (previously: y_1 and y_2). Suppose we can find an equivalent class of indistinguishable worlds X and a set of worlds Y such that:

(27) General counterexample 1:

a. $\forall x \in X : \exists y \in Y : y < x$.

b. No subset of Y included in a single class would be sufficient for (27a).
$\neg\exists\overline{y} \in Y : \forall x \in X : \exists y \in [\overline{y}] \cap Y : y < x$

c. $\neg\exists(x, y) \in X \times Y : x < y$.

Example (26) satisfies these constraints with $X = [x] = \{x, x'\}$ and $Y = \{y_1, y_2\}$. Under the more general conditions in (27), the following expressible counterfactual is originally true, but it comes out false with the optimally compliant order $\overset{*}{<}$: "If ($X \cup [Y]$), $[Y]$".

[6] It is easier to consider that the actual world is not represented but it could in principle be y_1 or y_2. Otherwise, it could also be an additional distinguishable world which should be added somewhere to the left of the diagram, where the smallest worlds belong to.

Proof: $\boxed{\text{True.}}$ (i) For $x \in X$, there is $y \in Y$, $y < x$ (27a). According to (27c), there is no X-world smaller than y. (ii) For $y \in [Y]$, either there is no x smaller, or we can apply (i) above. $\boxed{*\text{False.}}$ Assume that the counterfactual is $*$true. Then there is $y \in Y$, $y \overset{*}{<} x$. But then $[y] \cap Y$ would be sufficient for (27a), and this would contradict (27b). □

Let us make sense of this counterexample by considering the situation in (26) again. There are two parallel ways to get closer to the actual world, the row on top and the row at the bottom of (26). Hence, there is some property which gives rise to parallel branches in terms of which state of affairs is more similar to the actual state of affairs (e.g., the property true of the worlds in the top row and not of the worlds in the bottom row). In this example, we are trying to partially ignore this property assuming that some worlds in two different branches are sufficiently similar to be collapsed. The present result shows that a loss of information of this sort necessarily alters the evaluation of some counterfactual, at least if we do not simultaneously ignore similar differences all the way down the parallel branches.

Overall, there are two properties of this counterexample which should be highlighted. First, the target language is finite. Consequently, the assumption about infinite boolean closure of the language does not play any role in this counterexample. This failure in the finite case also yields consequences for the project which would consist in justifying the limit assumption from coarse-graining (as presented in (4), see discussion in §5). Second, some counterfactual which was originally true became false when moving to our optimally compliant expressible order. The (potential) finiteness of the target language and the direction of the alteration of truth value are two differences with the following class of counterexamples.

4.1.2 FALSE BECOMES TRUE (WITH A NECESSARILY INFINITE LANGUAGE $\mathcal{L}$)

Consider that the universe is as represented in (28): an infinitely decreasing sequence of indistinguishable worlds $\{x, x', x'', ...\}$ interspersed with a series of distinguishable worlds $\{y_0, y_1, y_2, ...\}$.

(28) $$\boxed{\ \dots \longleftarrow y_2 \longleftarrow x'' \longleftarrow y_1 \longleftarrow x' \longleftarrow y_0 \longleftarrow x\ }$$

Alternatively, consider that the universe is an infinitely decreasing sequence of worlds which are all distinguishable $\{x_0, y_0, x_1, y_1, x_2, y_2, ...\}$, but where there is an indistinguishable alter-ego for every odd world in the sequence $\{x'_0, x'_1, ...\}$, with $\forall n : x_n \sim x'_n$.

(29)
$$\begin{array}{ccccccccccccc} \dots & \longleftarrow & y_2 & \longleftarrow & x_2 & \longleftarrow & y_1 & \longleftarrow & x_1 & \longleftarrow & y_0 & \longleftarrow & x_0 \\ & & & & \wr & & & & \wr & & & & \wr \\ \dots & & & & x'_2 & & & & x'_1 & & & & x'_0 \end{array}$$

Intuitively, in both of these examples (28) and (29), some irrelevant or imperceptible differences (between xs) cut across infinitely many other differences (between ys) which we are *not* in a position to disregard as well. As a result, the counterfactual "If $([x] \cup \bigcup [y_n])$, $[x]$" —or "If $(\bigcup[x_n] \cup \bigcup [y_n])$, $\bigcup[x_n]$" in the case of (29)— is false, but it comes out true if we use the optimally compliant expressible order $\overset{*}{<}$. Instead of proving this result in these two particular cases, we can right away describe a more general version.

The following more general description does not affect the structure of these examples, which will always rely on a backbone of infinitely many worlds with the odd worlds being indistinguishable from some other world. However, it is important to specify explicitly how such a sub-counterexample can and cannot be embedded within potentially many other worlds without loosing its crucial features. For instance, if there were plenty of worlds smaller but indistinguishable from the xs and the ys which actively participate to the counterexample, the sub-example itself would not participate in any way to the evaluation of any counterfactual, for which priority is given to worlds closer to the actual world. Here comes the general description of the counterexample. Suppose we can find two non-empty sets of worlds X and Y such that $[X] \cap [Y] = \emptyset$ and:

(30) General counterexample 2:

a. There is a y below every x.
$\forall x \in X : \exists y \in Y : y < x$

b. There is an x below and *below every y.
$\forall y \in Y : \exists x \in X : x < y$ and $x \overset{*}{<} y$ [i.e. "If $([y] \cup [x])$, $[x]$" is true].

c. No y is *below an x (i.e. there is always an element indistinguishable from an x below an element indistinguishable from an y).
$\forall (x, y) \in X \times Y : \neg(y \overset{*}{<} x)$ [i.e. "If $([y] \cup [x])$, $[y]$" is false].

d. Every element indistinguishable from an element of X and below an(other) element of X is above some y [this roughly guarantees that there is no x again below the whole decreasing sequence with ys].
$\forall x \in [X] : (\exists x' \in X : x < x') \to (\exists y \in Y : x < y)$[7]

[7]Notice that we could merge (30a) and (30d) by rewriting the latter as (30d'):

The first two conditions guarantee the presence of an infinitely decreasing sequence with Xs and Ys alternatively (in fact, there is no constraint for it to be a sequence rather than a graph with multiple infinite branches). Notice that the X-worlds may or may not be distinguishable, as in (28) or as in (29). If they are not distinguishable, (30c) is (mostly) redundant because of the repetitive presence of elements of the same class $[x]$ below each y. If they are distinguishable, then the counterexample relies on there being other reasons for the elements of X to resist being ordered above the Ys (in (29), this is achieved by the bottom row). In all cases, the conditions (30c) and (30d) are less essential, they mostly guarantee that situations like (28) and (29) are not altered by other surrounding worlds.

Most importantly for our purposes, the conditions in (30) ensure that there is an expressible counterfactual which is originally false and which comes out true with the new order $\overset{*}{<}$: "If $([X] \cup [Y])$, $[X]$".

> **Proof:** $\boxed{\text{False.}}$ Let $x \in X \subseteq [X]$. Suppose you can choose $x' \in [X] : \forall w < x' : w \notin [Y]$. Then (30d) entails that there is a y below this x'. This is absurd and hence (5) is falsified and the counterfactual comes out false. $\boxed{*\text{True.}}$ (i) Let $x \in [X]$. Then according to (30c) there is no $[Y]$-world *smaller (notice that $\overset{*}{<}$ is a relation between equivalent classes and whether x belongs to X or is indistinguishable from an element of X does not make a difference). (ii) Let $y \in [Y]$. We can choose $x \in X$ such that $x \overset{*}{<} y$ (see (30b)). As proved by (i), there is no Y-world below that x. Together (i) and (ii) justify (5) for all elements of the antecedent $[Y] \cup [X]$. $\square$

4.2 FULL CHARACTERIZATION: PROOF

The counterexamples from §4.1 cover all situations under which there is no expressible compliant order. The result is actually twofold.

(31) **Result A:** The optimally compliant order $\overset{*}{<}$ makes some counterfactual true instead of false iff the original order satisfies (27)
[i.e. roughly, if there is a sub-portion of the universe as in (26)].

Result B: The optimally compliant order $\overset{*}{<}$ makes some counterfactual false instead of true iff the original order satisfies (30)
[i.e. roughly, if there is a sub-portion of the universe as in (28) or (29); notice that this incidentally requires that the target language be infinite].

(30d') $\forall x \in [X] : (\exists x' \in X : x \leqq x') \rightarrow (\exists y \in Y : x < y)$.

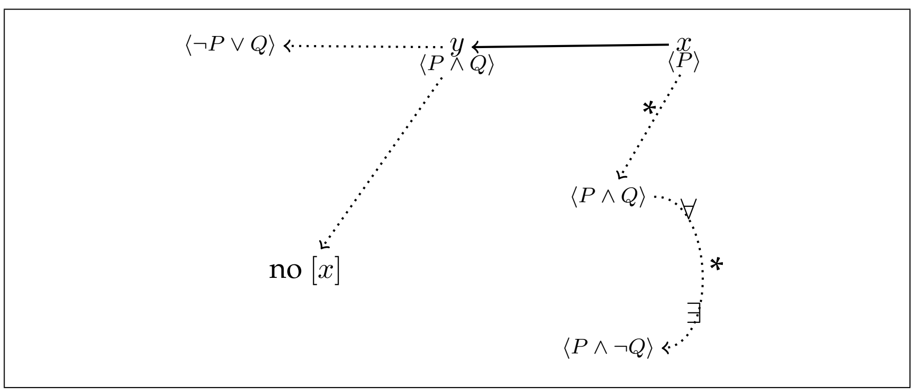
The rest of this section contains the proof of these two results. These proofs mainly consist in constructing incrementally the relevant sub-examples (26) or (28-29) when the failure described in results A or B arises.

Proof of result (31A): The right to left direction of the equivalence follows from the fact that (27) yields counterexamples (§4.1.1). Let us prove the other direction. Assume that P and Q are expressible propositions such that: (i) "If P, Q" is originally true and (ii) "If P, Q" is *false.

Step 1: construct the diagram below incrementally, starting at the top right, and adding smaller and smaller worlds.

Let $x \in P$ be a "falsifier" according to (ii): for all $P \wedge Q$-worlds *smaller than x, there is a further *smaller $P \wedge \neg Q$-world. This is represented by the consecutive dotted arrows at the right hand side of the diagram below. According to (i), we can find a $P \wedge Q$-world y smaller than x such that no $P \wedge \neg Q$-world is smaller (see the top row in the diagram).

Let us prove as well that there can be no element of $[x]$ below y. Otherwise, x is a $P \wedge Q$-world (top row). x is clearly *smaller or identical to x and so there is a $P \wedge \neg Q$-world *smaller than it (right hand side of the diagram). Hence, there is a $P \wedge \neg Q$-world z *smaller than x. Thus we can find $z' \sim z$ smaller than x, and thus z' would be a $P \wedge \neg Q$-world smaller than y. This is impossible because all worlds below y should be $\neg P \vee Q$ (top row).



Step 2: Generalize to $[x]$. A similar diagram can be built starting from any world $x \in [x]$. Indeed, similar consequences arise for indistinguishable worlds from the truth of a counterfactual —indistinguishable worlds satisfy the same combination of P and Q— or from the *falsity of a counterfactual —indistinguishable worlds are in *relations with the same worlds.

Step 3: X and Y satisfy (27). Let $X = [x]$, and Y a set of ys we obtain constructing such diagrams for each element of X. We can prove that (27) holds.

$\boxed{\text{(27a)}}$ was at the heart of the construction: there is always a y below an x.

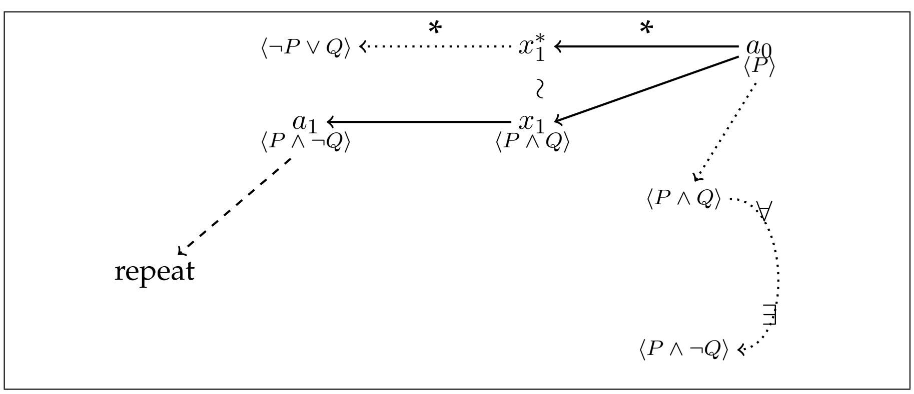
$\boxed{\text{(27b)}}$ Suppose there is $\overline{y} \in Y$ such that for all $x \in X$, $\exists y_x \in [\overline{y}] \cap Y : y_x < x$. Then $\overline{y} \overset{*}{<} x$. This is because there is no world indistinguishable from x below any element of Y, hence no X-world below the y_xs. But then, according to the right hand side of the diagram again, we can find a $P \wedge \neg Q$-world *smaller than $\overline{y}$, and hence some world indistinguishable from it smaller than y. This is excluded because all worlds smaller than y are $\neg P \vee Q$-worlds (top row).

$\boxed{\text{(27c)}}$ follows from the construction of the diagonal arrow from y to no $[x]$. □

Proof of result (31B): The right to left direction of the equivalence follows from the fact that (30) yields counterexamples (§4.1.2). Now, assume that P and Q are expressible propositions such that: (i) "If P, Q" is false and (ii) "If P, Q" is *true.

<u>Step 1: construct the diagram below.</u> Let $a_0 \in P$ falsify the original counterfactual (i): for all $w \leqq a_0$, if $w \in P \wedge Q$, there is a $P \wedge \neg Q$-world smaller (right hand side of the diagram). On the other hand, because the counterfactual is *true, there is a $P \wedge Q$-world x_1^* *smaller than a_0 and such that all *smaller worlds are $\neg P \vee Q$-worlds (top row in the diagram). x_1^* can be changed for an indistinguishable world $x_1 \leqq a_0$ (because $x_1^* \overset{*}{\leqq} a_0$). Define $X(a_0) = \{x \in P \wedge Q | x < a_0; x \overset{*}{<} a_0; \forall y \overset{*}{<} x : y \in P \wedge Q\}$, we proved that this set is not empty. More generally, $X(w)$ is not empty if $w \in P$ and $w \leqq a_0$.

Now for any $P \wedge Q$-world w smaller than a_0, $A(w) = \{a \in P \wedge \neg Q : a < w\}$ is also not empty. This applies for instance to any world in $X(a_0)$, and we can place a_1 below x_0 in the diagram.



<u>Step 2: Construct X and Y by applying the diagram recursively.</u> Let $X = \bigcup X^{(n)}$ and $A = \bigcup A^{(n)}$ where $X^{(n)}$ and $A^{(n)}$ are constructed as follows.

$A^{(0)} = \emptyset$, $X^{(0)} = X(a_0)$ and for all $n \geq 0$:

– $A^{(n+1)} = A^{(n)} \cup \bigcup\{A(x) : x \in X^{(n)}\}$

– $X^{(n+1)} = X^{(n)} \cup \bigcup\{X(a) : a \in A^{(n)}\}$

Intuitively, the sequence of constructions represented in the schema above is applied recursively.

Step 3: X and Y satisfy (30). X and A are not empty and they contain strictly different indistinguishability classes of worlds ($[X] \cap [Y] = \emptyset$), because $X \subseteq Q$ and $A \subseteq \neg Q$.
$\boxed{\text{(30a) and (30b)}}$ follow from the non-emptyness of all sets $A(x)$ and $X(a)$ for $x \in X^{(n)}$ and $a \in A^{(n)}$.
$\boxed{\text{(30c)}}$ Every world *below an x is a $\neg P \vee Q$-world (top row), hence it cannot belong to $A \subseteq P \wedge \neg Q$.
$\boxed{\text{(30d)}}$ Let $x \in [X]$ such that we can choose $\exists x' \in X : x < x'$. First, $A(x)$ is not empty because x is a $P \wedge Q$-world smaller than a_0. Second, $A(x) \subseteq A(x')$, because any $P \wedge \neg Q$-world smaller than x' is smaller than x. Hence, all potential A-worlds below x have been added as A-worlds below x'. □

4.3 Summary

The previous results describe the conditions under which expressible orders or ordering sources cannot recover the truth values of expressible counterfactuals as given by some non-expressible order or ordering source. In fact, the counterexamples which were described prove that as soon as the language $\mathcal{L}$ is such that two worlds cannot be distinguished (i.e. $\mathcal{L}$ is not maximally expressive), and two other worlds can be distinguished ($\mathcal{L}$ is not trivial, roughly), then $\mathcal{L}$ is not self-sufficient in the sense that there exists an order/ordering source on the universe which is a) not expressible and b) not compliant with an expressible order/ordering source. This proves how incorrect the conjecture (2) is.

5 Coarse-graining is not easy to obtain

Certain differences between possible worlds are not worth considering, they might not be relevant or expressible or "graspable" (e.g., Kratzer 1989). The process I presented can be thought of as a mechanism to focus on a set of select propositions $\mathcal{L}$ and to disregard irrelevant differences. There is a risk that such a process also erases key information about the structure of the universe, and the process I presented was thus optimized to preserve the truth value of the counterfactuals built from $\mathcal{L}$. Yet, there are situations in which there is simply no way to attain this goal. Specifically, the optimal simplification process I discussed may render false some relevant counterfactuals which were originally true (and the other way around but only if the target language is infinite).

This perspective on the previous results makes the conjectures in (3-4) as questionable as the conjecture (2) has become. Let me discuss conjecture (4) again. One of the debates about counterfactuals *à la* Lewis or *à la* Stalnaker concerns the so-called limit assumption: given some property P, can we always go down to a P-world with no closer P-world? The limit assumption is false in general because there is no reason to ban a universe where there would be a sequence of worlds indefinitely closer to the actual world without ever failing to satisfy P. Yet, proponents of the limit assumption have argued that if this type of sequence is in principle possible, there is reason to believe that human beings would not be willing or able to deal with all subtle and irrelevant differences between these infinitely many worlds. Hence, the sequence would standardly be reduced to a simpler one where the limit assumption may hold. This is the case for instance if we always consider a finite number of differences between worlds at a time. Lewis phrased this explicitly as follows:

> If we want the Limit Assumption, I take it that what we need is some sort of coarse-graining. We must imitate the finite case by ignoring most of the countless respects of difference that make the possible worlds infinite in number. *(Lewis 1981, p.91)*

Unfortunately, this suggestion now faces the following dilemma. There is a need for a process to simplify universes which are too complex (e.g., by focussing on a finite language $\mathcal{L}$, in the terms introduced before). However, no such process could succeed without altering the truth value of certain *relevant* counterfactuals.

Various questions remain open. First, there are reasons to examine further the contribution of various assumptions I relied on. For instance, there might be a distinction between the language which constrains orders or ordering sources — the language of thought? — and natural language, which delineates the set of expressible or relevant counterfactuals. As a result, the language of thought may allow fine-grained semantics for counterfactuals (possibly finer than what is relevant for natural language). Yet, the present work shows that disappointing surprises may easily arise and a proper formalization cannot be avoided to decide whether this refinement runs into obstacles similar to the counterexamples I presented. Second, how should we optimize the universe-simplification process if the constraint of preserving truth values is not sufficient? To investigate this question, one would need to consider and compare more thoroughly various simplification processes (see §3.3). Concretely, the alterations predicted by different processes should be evaluated both on conceptual grounds and with respect to speakers' intuitions (relative to cases where, e.g., we consciously decide to ignore certain differences between worlds).

References

Ginsberg, M. L. (1986). Counterfactuals. *Artificial Intelligence 30*(1), 35–79.

Kratzer, A. (1979). Conditional necessity and possibility. *Semantics from Different Points of View: With 15 Figures*, 117.

Kratzer, A. (1981). Partition and revision: The semantics of counterfactuals. *Journal of Philosophical Logic 10*(2), 201–216.

Kratzer, A. (1989). An investigation of the lumps of thought. *Linguistics and Philosophy 12*(5), 607–653.

Lewis, D. (1973). *Counterfactuals*. Harvard University Press.

Lewis, D. (1981). Ordering semantics and premise semantics for counterfactuals. *Journal of Philosophical Logic 10*(2), 217–234.

Stalnaker, R. C. (1968). A theory of conditionals. In N. Rescher (Ed.), *Studies in Logical Theory*, Volume 2 of *American Philosophical Quarterly Monograph Series*, pp. 98–112. Oxford: Blackwell.