A causal explanation for the contrafactive gap
v3 under revision—working document

Tom Roberts and Deniz Özyıldız

Institute for Language Sciences, Utrecht University.

University of Stuttgart.

Abstract

Contrafactive predicates, which assert belief in their complement while presupposing that that complement is false, appear to not exist (Holton, 2017). This lexical gap is surprising, given the widespread cross-linguistic existence of factive verbs like *know* which presuppose the *truth* of their complement, and a variety of expressions which entail a proposition's falsity, like *be wrong* and *falsely believe*. However, the contrafactive gap is not universally believed to exist (Sander, 2025), and there is no consensus on its source even if it does (Holton, 2017; Strohmaier and Wimmer, 2022; Glass, 2025: a.o.).

We examine a variety of potential contrafactive predicates and conclude that the contrafactive gap is in fact real, or at the very least a strong typological tendency, as none of these candidates quite meet the bar of encoding a falsity inference that is strong enough to truly be considered a bonafide counterfactive. We then derive the contrafactive gap from a general constraint on lexical meaning that requires presupposed content to be causally upstream of at-issue content, drawing inspiration from the notion of presupposition as ontological precondition (Roberts and Simons, 2024). We propose that the inability of the presupposition 'not $\boldsymbol{p}$' to be upstream of the assertion '$\boldsymbol{x}$ believes $\boldsymbol{p}$' makes contrafactives 'impossible predicates.' We then show that a variety of other attested attitude predicates can naturally receive an analysis in these causal terms and suggest that our account can help explain why some other logically possible unattested predicates do not exist. Finally, we speculate on possible underlying bases for the constraint, such as pressures from communicative efficiency (Enguehard and Spector, 2021).

1 Introduction

Natural languages are replete with clause-embedding predicates which generate inferences about the truth or falsity of their complements. For example, factive predicates like *know* presuppose that their complements are true, whereas veridical and anti-veridical predicates like *be right/be wrong* entail (though do not presuppose) truth and falsity, respectively. Here, we take entailment persistence under the scope of negation as a diagnosis for projection (following, e.g., Chierchia and McConnell-Ginet 1990).

(1) **Factives**

a. Alice knows that it's raining in Leiden. → It's raining in Leiden.
b. Alice doesn't know that it's raining in Leiden. → It's raining in Leiden.

(2) **Veridicals**

a. Alice is right that it's raining in Leiden. → It's raining in Leiden.
b. Alice isn't right that it's raining in Leiden. → It's not raining in Leiden.

(3) **Anti-veridicals**

a. Alice is wrong that it's raining in Leiden. → It's not raining in Leiden.
b. Alice isn't wrong that it's raining in Leiden. → It's raining in Leiden.

However, Holton (2017) suggests the existence of a curious lacuna with respect to truth/falsity inferences in natural language lexica: the absence of predicates which are *contrafactive*—belief predicates which are like factives but presuppose the falsity of their complement rather than its truth, yielding the pattern of inference in (4). We illustrate this pattern here and throughout with a putative contrafactive verb *contra*[1], which is like *know* except it presupposes that its complement is false instead of true:

(4) **Contrafactives** (unattested)

a. Alice contras that it's raining in Leiden. → It's not raining in Leiden.
b. Alice doesn't contra that it's raining in Leiden. → It's not raining in Leiden.

The lack of contrafactives is surprising because false beliefs are not only conceivable but arguably commonplace, and further there is nothing inherently wrong with predicates that entail (but do not presuppose) false belief, or something close to it, such as *be wrong* (Rosenberg, 1975). And as we have seen by the existence of factive predicates, there is also nothing

For helpful discussion, we would like to thank audiences at SuB 28, TbiLLC 19, the ILLC MLC Seminar, the MECORE workshop in Edinburgh, and the False Beliefs workshop in Düsseldorf, as well as Pranav Anand, Ryan Bochnak, Milica Denić, Kajsa Djärv, Emily Hanink, Dan Hoek, Morwenna Hoeks, Mora Maldonado, Alda Mari, Dean McHugh, Raquel Montero Estebaranz, Prerna Nadathur, Doris Penka, Jed Pizarro-Guevara, Thorsten Sander, Sheng-Fu Wang, and Simon Wimmer. This research was funded in part by the project "MECORE: Cross-linguistic investigation of meaning-driven combinatorial restrictions in clausal embedding" jointly awarded by the AHRC (AH/V002716) and DFG (RO 4247/5-1) to Romero, Uegaki and Roelofsen, and the project "A Sentence Uttered Makes a World Appear—Natural Language Interpretation as Abductive Model Generation" awarded by the NWO (406.18.TW.009) to Muskens.

[1] Credit to Simon Wimmer for coining this term.

wrong with presupposing the truth of a clausal complement. Rather, there is something about the combination of presupposition and falsity inference that is deficient. We refer to this pattern as the **contrafactive gap**.

The contrafactive gap

	inference of truth	inference of falsity
via entailment	*be right*	*be wrong*
via presupposition	*know*	

Holton (2017) proposes that the contrafactive gap should be explained on ontological grounds. The idea is that *that*-clause complements of factive verbs refer to facts (as opposed to propositions), so by analogy, *that*-clause complements of contrafactive verbs must refer to *contrafacts* ('facts, only false'). And since there are no contrafacts, there are therefore no contrafactive predicates.

While we believe Holton's account is in some sense getting at the core intuitive oddity of contrafactive predicates, it has several shortcomings. Empirically, Holton considers only a handful of mostly genetically related languages, so it's not clear whether the lack of contrafactives is a deep property of language or an accidental one of the languages he considers. This explanation is also difficult to falsify–since contrafacts are hypothetical formal objects, so it is not clear how (or if) we could identify one. It is also worth noting that in other domains, 'negative' events and individuals (e.g. a non-leaving event in *I saw Mary not leave*) have been posited to model semantic phenomena (Bernard and Champollion, 2018; Bledin, 2022: a.o.), so it is not at all obvious that contrafacts shouldn't exist or be part of linguistic theorizing. But even if we could prove that contrafacts, whatever they may be, are not part of our natural language ontology—and facts are—we then face the question of why this ontological asymmetry should be the case, arguably an analogous problem to the original linguistic question about the contrast between factives and contrafactives.

Assuming the contrafactive gap is robust, we would like to find a reason for the contrafactive gap rooted in some more general feature of language or cognition, rather than kicking the explanatory can down the road. In this paper, we concur with Holton's generalization that the contrafactive gap exists, but argue that it follows from a more general constraint on **lexical coherence**: languages do not like to lexicalize predicates whose presuppositions do not 'support' their at-issue content, in a sense we will make precise using causal models (Pearl, 2000).

This paper proceeds as follows. We first substantiate the contrafactive gap as a real phenomenon—either a universal property of human languages or a strong typological tendency—by examining a variety of plausible cross-linguistic counterexamples but concluding that none of them quite fit the bill. We then propose that the contrafactive gap can be explained theoretically, by positing a general constraint on the semantics of predicates requiring that their presupposed content be causally upstream of their at-issue content, and show that this explanation correctly rules out other types of unattested predicates, while allowing for a diversity of attested predicates, including close neighbors of *contra*. We then argue that

our proposal is more explanatory than alternative possible accounts which derive the contrafactive gap from pragmatic competition or informativity, while leaving a path forward for uniting our proposal with recent work on the learnability of unattested predicates.

2 What makes a contrafactive?

Holton (2017) proposes four criteria that a predicate must satisfy in order to count as a contrafactive: It must (a) presuppose that its complement is false, and be (b) doxastic, (c) stative, and (d) unanalyzable. The latter three criteria are based on Williamson's (2000: p. 33 et seq.) characterization of the (denotation of the) verb *know* as the most general factive mental state operator, which is to say that *S knows that p* is argued to be entailed by any other sentence of the form *S Vs that p* where *V* is stative, factive, and doxastic (e.g., *remembers*, *sees*, etc.). In effect, the idea is that *contra* is the hypothetical most general *contrafactive* mental state operator.[2]

2.1 Motivating the criteria

This choice of criteria might seem arbitrary or only historically motivated: why is this particular combination of features interesting? While perhaps imperfect, they are useful in that they set *some* standard that allows for a direct comparison between existing work on the contrafactive gap (Holton, 2017; Strohmaier and Wimmer, 2022; Sander, 2023) and false belief reports (e.g. Glass, 2023; Anvari et al., 2019; Maldonado and Percus, 2024), and that they might serve as a guide for identifying the properties candidates for contrafactivity might indeed have. We briefly outline below some reasons to hold on to these criteria for contrafactivity, while acknowledging that searching for predicates with some other possible combination of properties might also bear interesting fruit. For views that question the validity of this choice of properties and suggest alternative formulations of contrafactivity, see Sander (2025); Glass (2025); Strohmaier and Wimmer (2022).

It is uncontroversial that languages can express false belief reports, but what we seek to explain is why contrafactivity seems to resist being rolled into a single lexical package. The unanalyzability condition is motivated for contrafactives since we are ultimately interested in properties of the lexicon: we want to rule out predicates which transparently contain meaning components contributed by adverbs like *wrongly* or prefixes like *mis-*, in the case of English, among other effects. Thus, we take 'unanalyzable' here to mean 'exists as a singular lexical item which cannot be semantically decomposed', rather than a more philosophical definition of having a meaning which can be expressed as multiple concepts that can be individuated. [3]

Turning to the other criteria besides falsity of the complement, let us first examine doxasticity. Since we are interested in false *belief* in particular, we need to rule out attitude verbs which are not belief-oriented, which is to say they don't express a belief relation between their subject and an embedded clause. Interestingly, it is readily apparent that other kinds of attitude predicates, such as speech act predicates, can carry strong falsity inferences (e.g. *lie*, for English speakers who accept it as a clause-embedder, or *pretend*), although speech act

[2] Many thanks to Thorsten Sander and Simon Wimmer for discussion surrounding these criteria.
[3] Note that this definition of unanalyzability is also slightly different from a requirement that contrafactives be monomorphemic, since words can be morphologically complex but have idiomatic interpretations.

predicates in particular might be ruled out on the basis of being non-stative, in addition to being non-doxastic.

Finally, why should we require that a putative contrafactive be stative? While we will return to this point in section 4 after presenting our proposal, the basic idea here is that we are looking for the simplest kind of contrafactive attitude. If this is measured in terms of the internal complexity of the eventualities that an attitude predicate describes, states are arguably simpler than events. Indeed, states are homogeneous, i.e. contain only subparts of the same kind as the whole, whereas events are not (Dowty, 1979). Change-of-state predicates—accomplishments and achievements in the Vendler (1967) sense—have result states as strict subparts, with some dynamic event resulting in those states, so such predicates will always have more complex event structures than dyed in the wool states.

This raises the question of whether all predicates that are both factive and describe a change of state imply coming into or out of a state of knowledge. And while this is the hallmark of factive CoS verbs like *discover* (5), the answer seems to be no for other cases. For instance, in (6), we have a factive main predicate, but the change of state described is in the subject's emotive attitude towards p, not their knowledge.[4]

(5) When Constance left her phone unlocked, I discovered that she has a terrible secret.
Presupposed: Constance has a terrible secret
Presupposed: I used not to know that Constance has a terrible secret
Entailed: I now know that Constance has a terrible secret

(6) This morning when I woke up, I randomly got angry that Mary smokes (even though I already knew that).

In section 5, we will examine some predicates that come close to being contrafactive according to these criteria, but suggest either that they systematically violate at least one of the four, or that we lack the evidence necessary to conclude that they satisfy all of them. In the end, we will conclude that we were not able to find contrafactives *in the sense that* they presuppose falsity, are doxastic, stative and unanalyzable. For a weaker, less controversial claim, the reader may assume that verbs that presuppose the falsity of their complement are rare, rather than not attested at all—see White and Rawlins (2018: ex. (14)) for their rarity across the lexicon of English.[5]

What we mean by 'predicates'

Languages may use yet other means for ascribing false belief which do not involve the lexical semantics of embedding predicates *per se*, but rather some type of clausal morphology or discourse particles. For example, Daakaka, an Oceanic language spoken in Vanuatu, employs mood marking in complement clauses of belief predicates to differentiate between neutral vs. false belief ascription, respectively the realis (REAL) and distal (DIST) moods, expressed by the particle *mo* and clitic *=t* respectively (von Prince, 2015):

[4] See Klein (1975), or Egré (2008) for a more detailed discussion of the factivity of emotive factives.
[5] *Qua* claim of non-existence, an airtight argument for the complete absence of contrafactives requires either an exhaustive survey of existing predicates or a proof based on logical reasoning. We believe that both of these possibilities will be hard to come by, but we believe the evidence at least supports the weaker claim that contrafactives are rare.

(7) Na=m dimye ka ko=t penin seaa ∅-ada via a ∅-ada
1S=REAL think MOD.COMP 2S=DIST roast.PL every 1D.IN banana and 1D.IN
vis mwe pwer kyun
banana REAL stay just
'I thought you had roasted all our bananas, but our bananas are just here.'
von Prince (2015: ex. 831a)

(8) Na=m dimye (na) pun=an=ne ló-ó mon mo nok ma
1S=REAL think (COMP) tell=NM=TRANS palm-coconut also REAL finish REAL
ge=te kyun.
be.like.MED just
'I think the story of the coconut ends like this.' von Prince (2015: ex. 830)

While such linguistic tools are certainly worthy of further study, mood and particles are not contrafactives in our sense because they do not describe eventualities, and cannot in and of themselves be stative. In this paper, we will restrict our attention to attitude verbs (and a few adjectives) which uncontroversially describe properties of eventualities, but see McGregor (2023) for a recent survey of different ways in which false belief is encoded in Australian languages.

2.2 Presupposition

Before considering potential contrafactives, a word about presupposition. We make the simplifying assumption that the factive presupposition is encoded in the lexical semantics of verbs like *know*, *remember*, etc., i.e., the presupposition of the truth of complement p is specified as such in the lexical entries for these verbs. Changing what needs to be changed, we assume that the contrafactive presupposition would be encoded wholesale in the lexical semantics of contrafactive predicates, if any are found to exist.

One reason that this assumption is important is for diagnostic purposes. It leads us to expect that we should be able to diagnose a candidate contrafactive *contra* as contrafactive in the same way that we are able to diagnose *know* as factive. We want, for example, both (9a) and (9b) to imply that it is raining in Leiden (projection).

(9) a. Alice contras that it's raining in Leiden. → It's not raining in Leiden.
b. Alice doesn't contra that it's raining in Leiden. → It's not raining in Leiden.

Another set of diagnostics that we rely on is that presuppositions cannot usually be reinforced or canceled. That is, the two continuations in (10) should respectively sound redundant and contradictory, if *contra* is contrafactive.

(10) Alice contras that it's raining in Leiden...
a. #and it's not. reinforcement
b. #but it is. cancellation

This assumption about wholesale lexical encoding of contrafactivity is a simplification for two reasons: First, whether or not a predicate is factive or not is often a combined property of that predicate and particular complement clause (Schulz, 2002; White, 2014). For example, *Alice knows that she let the cat in* will presuppose that the cat was let in but *Alice knows to let the cat in* does not. Second, even when a factive inference is expected to be available (e.g., with *know that*) whether it is actually drawn, whether it projects, and how strongly it does both of these things is highly dependent on discourse context (Simons et al., 2010; Tonhauser et al., 2018; Degen and Tonhauser, 2022).

Very few would deny, however, that the availability of the factive inference is sensitive to the lexical semantics of attitude predicates. That is, while the factive presupposition associated with *know* may come and go, its availability has something to do with this verb's lexical semantics. Thus, a compelling reason to encode the presupposition wholesale in lexical meaning is to clearly differentiate predicates which do have factive-like inferences and those which do not, even if one acknowledges that these inferences may not be as sticky as traditional views of presupposition. This does, however, mean that we need to be very cautious about our interpretation of presupposition diagnostics: one particular example of a predicate failing to exhibit hallmarks of presupposition in one particular context should not be enough to invalidate the notion that that predicate is presuppositional at all.

3 Explaining the contrafactive gap: constraints on lexical coherence

Despite a proliferation of tools natural language gives us to talk about false beliefs, none of those tools of which we are aware seem to package *contrafactivity* into a single predicate, rather relying on pragmatic inferences or non-predicative modifiers. very few

In trying to explain this contrafactive gap, we are aiming to put some teeth on constraints on lexicalizability in general—the idea being that certain configurations of information simply cannot be put together in a single package. The core intuition we aim to capture is that contrafactivity is not lexicalizable because a contrafactive verb would describe an eventuality that is, in an important sense, deviant. Informally, we will propose that *know* can be lexicalized because there is an inherent connection between the fact of p and the belief in p, but *contra* cannot because the fact that $\neg p$ and the belief that p lack such an inherent connection.

This paper broadly meshes with recent work which characterizes presuppositions of predicates (as opposed to purely anaphoric presuppositions) as *preconditions* of at-issue content (for Schlenker (2021), 'epistemic' preconditions; for Roberts and Simons (2024), 'ontological'). The idea is that presuppositions are presuppositions in virtue of being (temporally, conceptually, ontologically, etc.) anterior to at-issue content. We take a slightly weaker stance: in our view, presuppositions are not preconditions *simpliciter*, but rather connected to at-issue content by a *causal chain* given typical circumstances.

We make this change because in at least some cases, the truth of a lexeme's presupposed content seems not to be a necessary precondition for the truth of its at-issue content. For example, we could treat *know* as having at-issue content (in part) *believe p* and presupposition (in part) p. We do not want to generally require the belief that p to presuppose p, since of

course false beliefs are possible, so the notion of presuppositions as necessary preconditions seems too strong.

As we mentioned in §2.2, in order to cash out this intuition more precisely, we will make the assumption that presuppositions are lexically encoded. That is to say that lexical items specify in their lexical entry which of their entailments are presuppositions, if any. Putting our cards on the table, we acknowledge that this is almost certainly much too simplistic a view, and many aspects of the analysis find inspiration in theories that assume that presuppositions are instead entailments whose not-at-issue status is derived through a pragmatic calculus (e.g., Stalnaker 1970, 1973, 1974; Rosenberg 1975; Karttunen and Peters 1979; Kadmon 2001; Abusch 2002, 2010; Simons et al. 2010, 2016; Abrusán 2011, 2016; Schlenker 2021), though see Djärv and Bacovcin (2020) for arguments for a lexicalization of presuppositions for attitude verbs in particular. We believe that the core of our account does not adjudicate between lexicalist and non-lexicalist views of presupposition itself, and our proposal can be adjusted to fit the latter view by establishing requirements on causal relations among lexical entailments, rather than between presuppositions vs. at-issue content. Because this is orthogonal to our main point, we leave a fully fleshed-out version of such an account for future work.

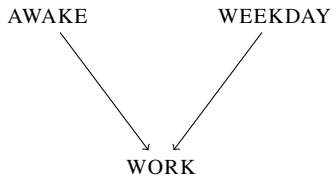
3.1 Background: Causal models

We will represent causal dependencies using the formal language of **causal models**, which have been profitably employed in much recent work describing causal phenomena in natural language (e.g. Nadathur and Lauer, 2020; Baglini and Bar-Asher Siegal, 2020; Copley and Mari, 2022; Glass, 2023; McHugh, 2023; Nadathur, 2023). Causal models are formal structures which symbolically encode causal relationships between variables in a system, represented as directed graphs in which edges denote causal links. We adopt a simplified version of causal models from Pearl (2000), defining a CM as a triple $\langle U, V, E\rangle$ as follows:

1. U is the set of exogenous variables (valued model-externally)
2. V is the set of endogenous variables (valued model-internally)
3. E is the set of structural equations f_i such that $\forall v_i \in V$, $v_i = f_i(pa_i)$, where $pa_i \subseteq U \cup V$

The role of E is to provide instructions on how to compute the values of each variable in V as a function of its parents (pa). Each node in a causal model represents a proposition in $U \cup V$, and an arrow from node A to node B represents a dependency of B on A. We distinguish only two types of causes: necessary and sufficient; the particular type of cause is determined by the structural equation which relates the parent and child nodes.

To illustrate, suppose that it is the case that you're at WORK in the morning if and only if two conditions hold: (a) that you are AWAKE on time and (b) that today is a WEEKDAY. All of AWAKE, WORK and WEEKDAY correspond to variables, in the definition above, and AWAKE and WEEKDAY are the parents of WORK. This state of affairs corresponds to a causal model in which the truth of AWAKE and WEEKDAY are jointly necessary for WORK to be true. The structural equation here is represented as a statement of first-order logic:



Exogenous variables: { AWAKE, WEEKDAY }
Endogenous variables: { WORK }
Structural equation: WORK := AWAKE ∧ WEEKDAY

In this model, the value of WORK is dependent on AWAKE and WEEKDAY—both causes are necessarily true for WORK to also be true. If both AWAKE and WEEKDAY were sufficient causes, instead of necessary ones, we would construct the same causal model with a disjunctive structural equation rather than a conjunctive one.

3.2 Presuppositions and causality

With the basic formalism in hand, we now develop our core proposal. Essentially, our idea is that there is a restriction on what kinds of meanings can be lexically encoded: discrete components of semantic information within a lexical item must be related by what we will call *causal chains*, or paths from causes to effects with potential intervening causal links.

We root this idea in research in human cognition, e.g. Radvansky and Zacks (2014) and Gärdenfors (2014), which suggests people conceive of events in terms of causal relations between entities. Assuming that individual predicates denote individual eventualities, a corollary of this view is that eventualities which are described by a single lexical item must be internally causally coherent; that is to say, its subparts are all causally related in some way so as to constitute a singular event.

Our proposal is that this causal connection underpins the structure of a lexical item's meaning. In particular, an eventuality-denoting predicate is associated with a single connected causal model in which (the situations described by) the predicate's presuppositions are causally upstream of (the situations described by) its at-issue content. We can state this more concretely as a relation between propositions:

(11) **Predicate Lexicalization Constraint (PLC; to be revised)**
An eventuality-denoting predicate with at-issue content α can have a presupposition π iff we can construct a causal model with a **chain** from π to α.

(12) **Causal chain**
For two variables v_1, v_2 in causal model M, there is a causal chain from v_1 to v_2 iff altering the value of v_1 can alter the value of v_2.

The PLC can be conceived of as a constraint on lexical **coherence**: all of the meaning packaged in a predicated has to be structured in a particular way. This first stab at a constraint will help us get a handle of the way to understanding why factives like *know* are lexicalizable but *contra* is not, but will require some amendments, to be elaborated in §3.2.2. Importantly, this initial characterization of the PLC is weak: it is satisfied if a causal model of a particular kind *can* be generated. As it turns out, this characterization is *too* weak, for reasons that

will become clear shortly; we will unpack these reasons more carefully in §3.3 and how they motivate a more precise characterization of the PLC.

3.2.1 *Know* is well-behaved

First, we will show how causal models can be used to capture the relationship between presuppositions and at-issue content of the attested predicate *know*. We must begin with a remark here about knowledge: a precise characterization of knowledge *per se* is a notoriously thorny and multifaceted issue, almost certainly requiring *justification* and other conditions on top of to true belief, or a different set of conditions altogether (see Ichikawa and Steup 2017 for an overview). Because we are concerned primarily with the relationship between belief and falsity in particular, we will not account for all of these aspects of knowledge here, but focus our attention specifically on the relationship between the factive presupposition of *know* and its belief entailment, while acknowledging that there are surely other causal factors at play in the denotation of *know*.

To see why the factive presupposition of *know* is well-behaved, consider the ordinary *know*-report in (13):

(13) Delilah knows that Konstanz is in Germany.

In virtue of containing an un-negated *know*, (13) presupposes that Konstanz is in Germany ($= p$) and asserts that Delilah believes that Konstanz is in Germany ($= B(d)(p)$). We want to construct a causal model in which this assertion is a downstream consequence of the presupposition and intuitively reflects some fragment of what it means that Delilah knows that Konstanz is in Germany. We treat the presupposition of *know* as a proposition, rather than a fact, but nothing crucial hinges upon this assumption.

Now, propositions (or facts) do not in and of themselves directly cause beliefs, since propositions can be true independently of whether someone in particular believes them. Rather, there are intermediate steps involved in the belief formation part of acquiring knowledge: beliefs are formed not *ex nihilo* but on the basis of some sort of evidence, whether that evidence turns out to be faulty or not. Evidence, in turn, depends on states of affairs: for instance, it having rained is (in most cases) a sufficient precondition for puddles on the ground to form, a sort of evidence for the truth of the proposition *it rained*.

We can break this chain of causal inferences into three links. The first link concerns the relationship between propositions and evidence. Intuitively, if a proposition p is true, this necessarily results in the world being in such a way that is consistent with that proposition being true.[6] For example, because Konstanz is in Germany, there are maps which show that Konstanz is in Germany, people in Konstanz pay taxes to the German government and vote in German elections, and so on. These results can themselves serve as evidence for the proposition p: I would be reasonably licensed to conclude that Konstanz is in Germany on the basis of my acquaintance of any of these pieces of evidence. We call such results–which can be any

[6] What is important here is that the idea that 'true propositions are observable' be something which we, as humans, assume. It could well be the case that some mathematical truths, for instance, are not actually practically discoverable, nor could we discover the truth value of a sentence like "Bob (the Diplodocus) had an even number of hairs". Rather, our assumption is that people generally believe that if p is true, we could **in principle** find that out.

number of things, including artifacts, collections of artifacts, other propositions, or something else–**indicators** for a proposition:

(14) **Indicators**
i_p is an *indicator* of proposition p iff being acquainted with i_p is sufficient to conclude that p.

We assume that, in agents' folk understanding of beliefs, the truth of any true proposition p generates some indicator for p. Let *indic(p)* stand for the existential statement $\exists i : i$ is an indicator for p. In causal terms, these assumptions mean that p is a sufficient condition for $indic(p)$. (Correspondingly i_p stands for any one of these indicators.) The resulting simple causal model is below:

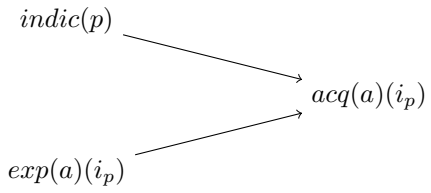
$$p \longrightarrow indic(p)$$

Structural equation: $indic(p) := 1$ if $p = 1$

Note however that p is not a necessary condition for $indic(p)$, since indicators for false propositions can exist. For instance, although Konstanz is in Germany in reality, it is still possible for someone to make forged maps showing Konstanz in Switzerland (assuming Konstanz has to be in either Germany or Switzerland, this is equivalent to $\neg p$). These forged maps could on their own be construed as indicators for the false proposition *Konstanz is in Switzerland*. Since the maker of the forged map does so with an intent to deceive, arguably p is a causally upstream of $indic(\neg p)$ in this case.

However, unlike the true indicator case, a proposition does not generally give rise to indicators for its negation on its own, but in conjunction with something else. In the case of the forger, it is a *combination* of the truth of p and the forger's explicit intention to make a false map that produces the indicator for $\neg p$. In short, for most cases, p is not sufficient on its own to cause $indic(\neg p)$. We will focus our attention only on true beliefs, though return to this issue in our discussion of *contra*.

The second link concerns the connection between indicators and belief-forming agents. By definition, in order for an indicator i_p to serve as evidence for some particular agent a's belief that p, a must be acquainted with that indicator. The existence of said indicator is, naturally, a necessary condition for being acquainted with it. Note that existence is of course not sufficient on its own for acquaintance, since an agent must have some experience with the indicator to be acquainted with it; call this requirement $exp(a)(i_p)$. The value of the proposition $acq(a)(i_p)$, namely that a is acquainted with an indicator for p, is the conjunction of these two necessary conditions:



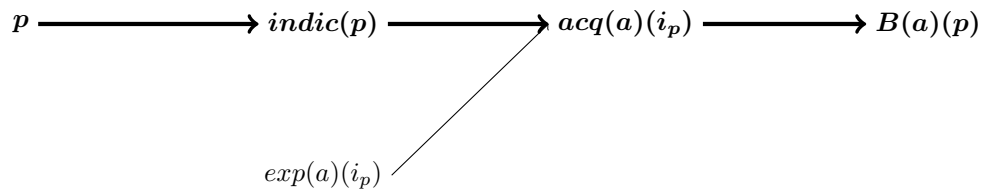
Structural equation: $acq(a)(i_p) := indic(p) \wedge exp(a)(i_p)$

The final piece of the puzzle concerns belief formation itself. We explicitly connect the acquaintance with an indicator to the formation of a belief: being acquainted with some indicator for p, by definition, is a sufficient condition to form the belief that p. Importantly, this definition does not preclude the formation of incorrect beliefs on the basis of faulty information, since an indicator itself need not align with reality, and agents may well misinterpret indicators. What is crucial, however, is that an agent who believes p nevertheless themselves believes that the indicator they observed is for p. In fact, that is the basis for their belief.

$$acq(a)(i_p) \longrightarrow B(a)(p)$$

Structural equation: $B(a)(p) := acq(a)(i_p)$

Putting these three causal links together, we have a causal chain from the truth of p to a's belief that p: that is, p is causally upstream of $B(a)(p)$. (Note that some nodes of the model are not part of the chain that links p to B(a)(p), in bold; in what follows, we will leave them out of models for simplicity.)



With this definition in hand, we can now see how *know*, as used in (13), adheres to the PLC. Consider a situation in which Delilah sees a contemporary map of Germany with Konstanz on it, and forms a belief that Konstanz is in Germany on that basis. Treating again the proposition *Konstanz is in Germany* as p, the map of Germany serves as an indicator for p, since it is reasonable to consider it sufficient that p on the basis of acquaintance with that indicator (the first link). The existence of the map facilitates Delilah's observation of it (the second link), an observation which leads to her adopting the belief that p (the third link).

In short, *know* describes a situation with a natural causal flow from the facts of the world to the formation of beliefs. For the moment, we will flag a major caveat with this conclusion: while in this example, we can trace a causal path from Konstanz being in Germany to Delilah's belief in such, it is not at all obvious that this example is illustrative of *know*-sentences in general. We will address this reasonable point of skepticism in greater detail in the following section in seeing how *know* is causally distinct from *contra*.

3.2.2 *Contra* is causally deficient

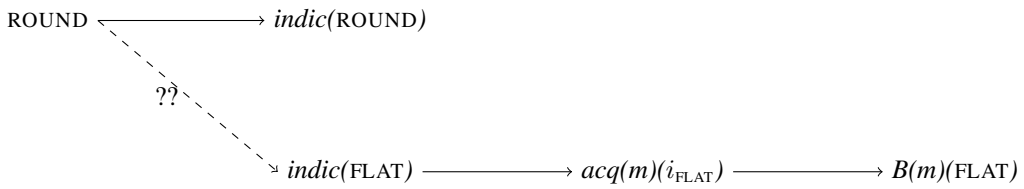
Having shown that *know* is causally well-behaved given our lexicalization constraint, we turn now to *contra*. We want to demonstrate that *contra* is unlexicalizable due to the oddness of constructing a causal chain from $\neg p$ to the belief that p. If we think about the kinds of situations under which false beliefs arise, they typically involve erroneous information or unjustified assumptions:

(15) [*Context: Mary is looking at a map from 1980, but mistakenly thinks it is from 2023*.]
Mary contras that Konstanz is in West Germany.

Rather than the fact of $\neg p$ (Konstanz not being in WG), a feature of the world consistent with evidence for p—in the case of (15), an outdated map—is what leads to the belief that p. To illustrate this more explicitly, assume, for the sake of constructing a proof by contradiction, that the *contra*-sentence in (16) is defined and true.

(16) Marianne contras that the Earth is flat.
→ The Earth is round (i.e., not flat).
→ Marianne believes that the Earth is flat.

Given our lexicalization constraint, there must then be a causal link between the fact that the Earth is ROUND, and Marianne's belief that it is FLAT. What should that link be? Recall that we proposed that beliefs are formed on the basis of indicators: that is to say, Marianne's belief is based on acquaintance with some indicator that the Earth is flat. We also proposed that propositions do not generate evidence (in our terms, indicators) for their negation, so the fact of the Earth being round does not typically generate evidence to the effect that it is flat (in fact, it generates evidence that the earth is round). In other words, ROUND inhibits or is, at best, unrelated to *indic(*FLAT*)*. This non-relation is illustrated by a dashed line in the model below:



It follows that either *B(m)*(FLAT) and ROUND are not linked, or that they impose the contradictory requirement on the model that ROUND is sufficient for indicators both for itself and its negation. Given that we don't want to admit contradictions into our model, it follows that there is no causal model which provides a chain from ROUND to *B(m)*(FLAT)—the belief in its negation.

However, even if we have shown that in this one particular situation we cannot construct a chain that links the presupposition of *contra* to its at-issue content, that is not enough to prove that *contra* is unlexicalizable. We could just as easily imagine a particular scenario in which a particular proposition *does* seem to generate evidence for its negation, which our lexicalization constraint would seem to admit. For instance, consider situations involving intentional deception:

(17) [*Context: Marianne thinks that Sasha is not a spy, because she is not behaving suspiciously. In reality, Sasha is a spy who is very good at concealing her spy-ness by acting normal.*]

Marianne contras that Sasha is not a spy.
≈ Marianne wrongly believes that Sasha is not a spy.

An important fact about spies is that their trade requires illusion: they act like normal people to avoid arousing suspicion. So the fact of Sasha being a spy can itself cause her to behave in a non-spy way, behavior which could serve as an indicator of the proposition that she is *not* a spy. (In fact, strategic leveraging of indicators in this fashion is the key property of deception.) This indicator can then lead to belief formation in the usual way, as illustrated in the causal fragment below:

$$spy(s) \longrightarrow indic(\neg spy(s)) \longrightarrow acq(m)(i_{\neg spy(s)}) \longrightarrow B(m)(\neg spy(s))$$

In effect, Sasha's non-spy behavior constitutes indeterminate evidence for whether she is a spy or not: she would act that way no matter what, so her behavior can plausibly be construed as an indicator for s or $\neg s$, and thus by our account, serving as evidence for either belief.

This example is problematic for the lexicalization constraint in (11) because it is weak, requiring only constructing that a causal model from π to α be *possible*. Therefore, it incorrectly predicts *contra* to be a possible lexical item on the basis of examples like (17).

What this state of affairs tells us is that we want to capture the intuitive badness of *contraing*—the unrelatedness of $\neg p$ to the belief that p—not in terms of possibility, but rather about general or typical situations. After all, the facts of the world don't *typically* give rise to evidence for their negation, even if in deviant cases like the spy scenario, they do.

3.3 Generalizing the PLC

The PLC is weak, since, as it stands, it simply asks whether we *can* construct a causal model with a chain from presuppositions to at-issue content. What we need to account for the impossibility of *contra* is a version of the PLC which is not satisfied by simply finding one example of a deviant situation, but rather needs to hold for *arbitrary* causal models with a particular structure.

This generalizing of the PLC is needed to allow for certain individual cases in which the facts can initiate a causal chain that results in disbelief of those facts, without such cases being sufficient to lexicalize *contra*. Essentially, what we want to show is that we cannot lexicalize eventualities characterized by causal models of the form below, for arbitrary choices of p (and attitude holder x):

$$p \overset{??}{\dashrightarrow} indic(\neg p) \longrightarrow acq(i_{\neg p})(x) \longrightarrow B(x)(\neg p)$$

In determining whether a particular predicate is lexicalizable, we need to be concerned with constructing **templates** of causal models, rather than considering models with variables valued in a particular way. After all, we can always construct tortured scenarios to establish a causal relation between *any* pair of propositions, apart from perhaps those which would entail a logical contradiction.

Rather, what we want is for only predicates which describe typical causal relations between eventualities to be lexicalizable. To capture this with the PLC, we need to somehow define the kinds of relations that are 'typical'. Formally, we achieve this by means of relativizing the meaning components of the predicate to a 'normative' world w_n, which can be intuitively thought of as 'the world in which causal relations between events proceed as a rational agent believes they would'[7]:

(18) **Predicate Lexicalization Constraint, final**:
A predicate V with at-issue content α can have a presupposition π iff in the normative world w_n a causal chain from $\pi(w_n)$ to $\alpha(w_n)$ can be constructed for any assignment of the variables of the arguments of V.

What (18) amounts to is a requirement that verbal presuppositions are such that under normal circumstances, they can head a causal chain resulting in the at-issue content, regardless of the specific values of the arguments. For example, if our denotation for *know* is as in (19a), then the lexicalization constraint necessitates the truth of the condition in (19b):

(19) a. $\lambda p_{st}.\lambda x_e.\lambda w_s : [p(w)]_\pi.[B(x)(p)(w)]_\alpha$
b. $\forall p, x$: there is a causal chain from $p(w_n)$ to $B(x)(p)(w_n)$

Importantly, a lot in this definition is riding on what exact is meant by a 'normal' world, a notoriously slippery concept in epistemology and philosophy of language (e.g. Alexander, 1973; Yalcin, 2016; Goodman and Salow, 2023). While we cannot adjuciate among the many possible definitions of 'normal' in this paper, for our purposes, we can intuitively think of 'normal' as referring most closely to the actual world in which a sentence including the predicate is uttered, but could also, in the right context, refer to fictional worlds. (We do not, after all, want to exclude predicates from existing because they describe eventualities which are ontologically deficient in fictional universes with bizarro cosmologies.)

Applying this definition, we need to be able to construct causal chains between p and $B(\neg p)$ **for any choice of** p in the normal world in order to lexicalize *contra*. Although in the normal world the proposition *Sasha is a spy* arguably can head a causal chain resulting in a belief that Sasha is not a spy, the same cannot be said about any number of other propositions, like *The moon is made of cheese*. And as we saw in the flat earth example, false beliefs can (and in fact usually are) formed on the basis of information which has nothing to do with the facts *per se*, but rather some kind of mistaken assumption. In such an example, the fact of the earth being round does not in and of itself cause the belief that the earth is flat; rather, some other evidential source does.

[7] An anonymous reviewer points out that 'normal' is not necessarily the right description for this concept. We suggest that a world in which everything is 'normal' from the perspective of a rational agent is one in which the relationship between causes and effects is precisely what that rational agent would expect.

4 Deriving the stativity criterion

Recall, from section 2, Holton's four necessary conditions for being a contrafactive predicate: (a) the presupposition that its complement is false, (b) doxasticity, (c) stativity and (d) unanalyzability. As an anonymous reviewer points out, it is not obvious why stativity should play a critical role. In this section, we would like to show why the stativity of a predicate might preclude it from presupposing the falsity of its complement.

The argument will be based on two case studies of predicates that seem to presuppose the falsity of their complement but fail tests for stativity. One class includes change of state verbs that entail intentional deception, such as Dutch *wijsmaken* (Hoeksema, 1999, 2021) and its German counterpart *weismachen*. The other class includes activity verbs whose subjects are forced to entertain a false belief, such as *hallucinate*.[8]

We will construct causal models corresponding to the meaning of these verbs and show that once a node in the model that accounts for their eventivity is removed, the result is a model that is deviant just like *contra*'s. We will then conclude that adding eventivity into the lexical semantics of a verb seems to license the co-packaging of a falsity presupposition.[9]

4.1 *Wijsmaken* and change of state predicates with the contrafactive inference

The predicate *wijsmaken* describes situations of deception, similar to an English predicate like *fool X (into thinking that)*. It entails that its complement is false, as continuing a *wijsmaken*-report with the assertion of the embedded clause is contradictory, in (20a). Moreover, this inference survives even when the verb is negated, as in (20b), which is consistent with the projection behavior of presuppositions.

(20) *Alice ('me'), talking about Mary ('ze')*:

a. Ze maakte me wijs dat ze rijk was… #en ze was wel rijk.
she made me wise that she rich was and she was PRT rich
'She fooled me into thinking that she was rich #and she *was* rich.'
(modified from Hoeksema 2021: ex. 3a to include the continuation and names)

b. Ze maakte me niet wijs dat ze rijk was… ik wist al dat ze arm was.
she made me not wise dat she rich was I knew already that she poor was
'She didn't fool me into thinking that she was rich (because) I already knew that she was poor.'

The predicate *wijsmaken* comes with three additional inferences that will be relevant, in (21b) through (21d). We have just seen that (21a) is presupposed, so is the prior lack of belief inference in (21b). The ongoing belief inference in (21c) and the deception inference in (21d) are entailments.[10]

[8] Many thanks to a reviewer for suggesting to look at *hallucinate*, as well as some of the related examples below.

[9] We will not similarly motivate the doxasticity requirement, but we speculate that non-doxastic attitude predicates for which the notion of (contra-)factivity is well-defined, e.g., speech predicates like *lie* (though see Anand and Hacquard 2014), might all have to be eventive, which in turn, might allow for subsuming them under the discussion in this section.

[10] Strictly speaking, the sentence also implies that the speaker no longer believes at *speech time* that Mary is rich (cf. *#I falsely believe that she's rich*). This inference is not relevant for present purposes and should not affect the foregoing discussion.

(21) (20a) implies…

a. that Mary isn't rich ($= \neg p$),
b. that Alice does not hold the belief that Mary is rich prior to topic time ($= \neg B(a)(p)(t')$, with $t' < t_0$ and $rightBound(t') = leftBound(t_0)$),
c. that Alice believes, at topic time, that Mary is rich ($= B(a)(p)(t_0)$),
d. that Mary *does something* in order to fool Alice ($= fool(m)(a)$).

The idea that *wijsmaken* encodes an intent to deceive on the part of the subject is substantiated by examples like (22). In this sentence, the subject of *wijsmaken* is inanimate, so the sentence is only felicitous under a humorous/metaphorical reading in which the clock 'intended' to deceive the speaker.

(22) *Context: A broken clock displays the time as 6:00. At 7:00 I observed the clock and erroneously formed the belief that it was 6:00, though later I learned that the clock was wrong.*

#De klok maakte me wijs dat het 6:00 was.
the clock made me wise that it 6:00 was
'The clock fooled me into thinking that it was 6:00.'

The question arises of whether *wijsmaken* should be considered a contrafactive or not given Holton's other criteria. While we'll focus on it violating stativity, we note the following. First, the predicate is decomposable into the morphemes wise+make and could be ruled out as a contrafactive on these grounds as well. But it is unclear to what extent the meaning of the whole can be derived through the meaning of its constituent morphemes. So we do not rely on this reason here.

The second point concerns doxasticity. At least for English and Dutch, predicates typically referred to as doxastic also have their subject as the believer (in *x believes p*, *x* is the believer). With *wijsmaken* or, say, *y fools x into believing p*, the believer is an object, while the subject of the (main) verb performs some action responsible for a *change* in the object's belief state. Holton's criteria for contrafactivity make no prescription for syntax or argument structure, it is perhaps telling that deception verbs might also systematically differ from factives like *know* in this respect as well.

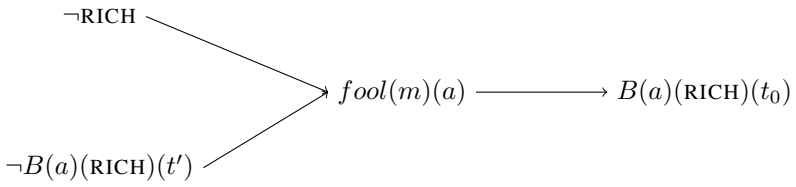
Returning to our main thread, in addition to the inferences in (21), we can diagnose the eventivity of *wijsmaken* by applying standard tests like those of Dowty (1979). For instance, *wijsmaken* is possible as a complement of *forceren* 'force', where stative predicates are not, suggesting that *wijsmaken* is likely to be eventive[11]:

(23) Hij forceerde haar om me wijs te maken dat ze rijk was.
he forced her for me wise to make that she rich was

[11]This argument is not quite sufficient to demonstrate that *wijsmaken* **must** be eventive, but rather that it **can** be. Arguing that *wijsmaken* is never stative is difficult, since standard diagnostics of obligatory eventivity in English, such as a verb in the simple present being interpreted as habitual, do not work in Dutch. Two native Dutch consultants informally report that *wijsmaken* reports only seem to describe actions, rather than states, but these data points should be treated with care.

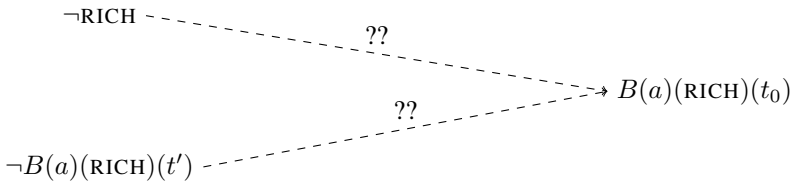
'He forced her to fool me into thinking that she was rich.'

We now propose a causal model representing the meaning of *wijsmaken*, which ties together the inferences just drawn out.



In this model, we suggest that $\neg p$ and $\neg B(a)(\text{RICH})(t')$ (with t' left adjacent to evaluation time t_0) are joint necessary conditions for $fool(m)(a)$, Mary's attempt to fool Alice. This node, on the other hand, is a sufficient condition for $B(a)(\text{RICH})(t_0)$, Alice's belief that Mary is rich at evaluation time. Motivating the first claim, one cannot fool someone into thinking that p if p is true (RICH is necessary) and one cannot fool someone into thinking something that they think already ($\neg B(a)(\text{RICH})(t')$ is necessary). Thus, here, nothing prevents the nodes corresponding to the presuppositions of *wijsmaken* from being causally upstream of the verb's assertions (in fact, they must be).

The question that we would now like to entertain is whether we can construct a coherent causal model in which (a) the node $fool(m)(a)$ is removed, where the reason for targeting this node is that it corresponds to the eventive assertion of the verb, and (b) the rest of the nodes keep their respective status as presupposed or entailed content. Coherent, here, means that the two nodes on the left (presupposed) have to be causally upstream of the node on the right (asserted). Below is one hypothetical way of constructing a model with the specifications given.



The key observation here is that this model is not coherent, as we require $\neg p$ to be causally upstream of Bp, which we've previously argued is not generally possible. While this is not necessary for this particular model (as it already incurs the cost of incoherence), a legitimate question is whether $\neg Bp$ can be causally upstream of Bp, this too, we think, is not possible without an intervening factor (such as the $fool(m)(a)$ node) that causes the change.

There are at least two other hypothetical models to consider, namely $\neg p \mapsto \neg Bp \mapsto Bp$ and $\neg Bp \mapsto \neg p \mapsto Bp$. If the previous observation is on the right track, these two models will also be ruled out by the Predicate Lexicalization Constraint.

In other words, while *wijsmaken* is not an impossible predicate given our proposal, removing one of its meaning components should result in a meaning that is not lexicalizable for the same reason as why *contra* is not lexicalizable.

The meaning component that we removed was what arguably makes the predicate a change of state, hence an eventive predicate. And it is this node that allows the contrafactive inference $\neg p$ to be causally upstream of *wijsmaken*'s assertion.

To complete our argument that stativity is necessary for the impossibility of contrafactives in Holton's sense, we would have to show that the nodes removed (like $fool(m)(a)$) that turn a causal model from a coherent to an incoherent one are in general nodes that encode eventivity.

We do not currently see a way of making this argument in full generality, but hope that an example is sufficient to motivate the relationship between contrafactivity and stativity. We now turn to a second case study about a different kind of eventive predicate, namely an activity, which leads us to the same conclusion as *wijsmaken*.

4.2 *Hallucinate* and activity predicates with the contrafactive inference

A second subclass of predicates which presuppose false belief but fail the stativity requirement includes *hallucinate*.

Unlike *wijsmaken*, *hallucinate* has an argument-role mapping more typical of belief verbs, in that it's the verb's subject that corresponds to (something like) an attitude holder. But whether the verb feels doxastic or eventive depends in part on the nature of its subject. We will consider its uses with human and chat bot subjects, but leave the exploration of anything else for another occasion.

When *hallucinate* occurs with a human subject, it typically describes a kind of nonveridical *sensory experience* which might also imply false belief, in (24).

(24) Consuela hallucinated that the ghost of Frida Kahlo is in her house.
$\rightsquigarrow$The ghost of Frida Kahlo is not in Consuela's house.

An anonymous reviewer points to a recently innovated sense of *hallucinate*, typically predicated of large language models (LLMs), which is linked to their sometimes producing strings of text that are false for no apparent reason, in (25). (We learn, incidentally, that Angela Merkel does not like dogs.)

(25) ChatGPT hallucinates that Angela Merkel loves dogs (#and she does).

In this use, the predicate clearly entails that its complement is false, as evidenced by the contradiction with a follow up like #*and she does*.

Moreover, this inference seems to project through the standard entailment-canceling operators like negation (26a) and polar questions (26b), which indicates that it is presupposed.

(26) a. ChatGPT doesn't hallucinate that Angela Merkel loves dogs.

b. Does ChatGPT hallucinate that Angela Merkel loves dogs?
   ⇝ Angela Merkel does not like dogs.

While it is far from obvious that an LLM could (or should) be construed as having beliefs, we can nevertheless imagine that some speakers treat them as such. To the extent that this sense of *hallucinate* is being treated as a mental state predicate (≈ 'believe hallucinatorily'), it seems to have the right inference patterns we would expect from a contrafactive. This is thus the use of the the predicate that we'll choose to model below.

We can immediately show that the predicate *can* be eventive on both of its uses. The predicate is, for example, acceptable in the present progressive under an ongoing event interpretation, namely, giving rise to the intuition that something is *happening* at speech time. (This is again based on tests from Dowty (1979).)

(27) What's happening over there?
   a. Consuela is hallucinating that the ghost of Frida Kahlo is in her house.
   b. ChatGPT is hallucinating that Angela Merkel loves dogs.

It is also straightforward to show that the predicate *can't be stative on its uses with human subjects*. Indeed, the (28a) is odd, and it only improves with some explicit mention of or contextually recoverable frequency, suggesting that it's a habitual.

(28) a. #Consuela hallucinates that the ghost of Frida Kahlo is in her house.
   b. Every time you prod her, Consuela hallucinates that the ghost of Frida Kahlo is in her house.

The same argument that the predicate can't be stative is more difficult to make for when its subject is an LLM. In environments that bring forth a **recurrence** or **iteration** inference, sentences like (24) imply that the hallucination *happened* several times over, but, for reasons that we do not understand fully, this inference is subdued in present simple sentences with ChatGPT as their subject, like (25).[12]

We can attempt to draw further contrasts between *ChatGPT hallucinates* and run of the mill statives. What we see is a degradation in acceptability or the promotion of the recurrence inference.

The first one of these tests involves pairing *hallucinate* with durative adverbials like *for 6 months* or *for 10 minutes*. Sentences thus constructed lack the reading that a particular hallucinatory state persisted continuously over the mentioned period of time, as in (29). Rather, they describe that the subject experienced a particular recurrent hallucination over it.[13]

[12]This contrast is reminiscent of the observation that inanimate noun subjects are sometimes possible with present simple attitude verbs like *explain* or *prove* (Bondarenko, 2022), or with verbs that take Repository of Information subjects like *the book* or *the sign* (Grimshaw, 2015; Major and Stockwell, 2021).

[13]It is interesting that varying the duration of the *for* adverbial makes a difference in terms of whether one is talking about what happened during a particular ChatGPT session or about properties of its different versions. This difference does not seem to affect the validity of the test.

(29) a. ChatGPT hallucinated for 6 months that Angela Merkel loves dogs.
b. Consuela hallucinated for 6 months that the ghost of Frida Kahlo is in her house.
$\rightsquigarrow$The hallucination started and stopped.

There is no recurrence or iteration inference with run-of-the mill stative belief predicates:

(30) a. ChatGPT knew/believed for 6 months that Angela Merkel loves dogs.
b. Consuela knew/believed for 6 months that the ghost of Frida Kahlo is in her house.
$\not\rightsquigarrow$The subject started and stopped having that belief.

A second test involves a particular property of some English modals like *must* and *can't*, which is that they only give rise to epistemic interpretations when they combine with stative verb phrases (Ramchand, 2014). Now, the pair in (31) shows that epistemic interpretations are available with *know*, but that that they are absent with *hallucinate*—unless the sentences in (31b) are read as habituals.

(31) a. ChatGPT must/can't know that Angela Merkel hates dogs.
b. #ChatGPT must/can't hallucinate that Angela Merkel loves dogs.

We might ultimately not be able to tell whether *ChatGPT hallucinates that p* only has a habitual of an activity reading, or whether it also has a bona fide stative reading. But, based on the foregoing discussion we conclude that simple present sentences like (25), are to be understood as **habituals** formed out of an eventive, rather than involving a lexically stative one. That is to say that *hallucinates that p* in sentences like (25) means something like 'repeatedly has a hallucination that p', as in (32).

(32) Every time you ask ChatGPT about Angela Merkel, it (always) hallucinates that she loves dogs.

Turning now to a third, crucial property of *hallucinate*, we assume that an essential component of hallucination is the distortion of reality. For a human, this distortion might be due to drugs, illness, or other perceptual interference; for an LLM, it is the unpredictable transformation of text inputs. Put another way, hallucination links an input ($\neg p$) and an causes some output (and eventually, some belief that p).

This leads to the following set of inferences for the predicate:

(33) ChatGPT hallucinates that Angela Merkel loves dogs.
$\rightarrow$ Angela Merkel doesn't like dogs. ($\neg p$) (presupposed)
$\rightarrow$ ChatGPT produces the output that Angela Merkel loves dogs ($= p$) (at issue)
$\rightarrow$ Something has gone wrong that led to ChatGPT producing this statement. ($B(chat)(p)$) (at issue)

A causal model that matches the inferences in (33), is given below.

$$\neg\textsc{lovesDogs} \longrightarrow \textsc{distortion} \longrightarrow B(chat)(\textsc{lovesDogs})$$

Here, the node $\neg$LOVESDOGS is a necessary condition for the node DISTORTION, and the node DISTORTION is a necessary condition for the node $B(chat)$(LOVESDOGS).

The node DISTORTION is at least one meaning component that makes the predicate *hallucinate* eventive. (A second one could be a refinement of the meaning of $B(chat)$(LOVESDOGS) into something like *actively entertain* or *actively imagine*, as hallucinations might require a cause and a sustained energy input.) To see why this is the case, it might suffice to imagine prisms and distortion pedals: These require a signal going into them (hence the necessary link between the first node and the second) and produce a signal coming out of them, which, down the line, may cause a certain belief (also a necessary condition in the end). The reason why the node contributes eventivity is that changes in the input will (generally) cause changes in the output, which extend in time.

Thus, it is possible for predicates like *hallucinate* to exist: We can construct coherent causal models in which the presupposition that $\neg p$ is upstream of the predicate's assertion.

We proceed now similarly to what we had for *wijsmaken*. If the node DISTORTION is removed from the model and the requirement that the presupposition $\neg p$ be upstream of the assertion Bp is kept, the model reduces to the incoherent one from *contra*: There is no conceptually feasible way of linking the former and the latter in the appropriate way.

$$\neg\textsc{lovesDogs} \overset{??}{\dashrightarrow} B(\textsc{lovesDogs})$$

What we see across the two case studies is that for a false complement presupposition to be encoded in the meaning of a predicate, the predicate also needs to encode some additional meaning component encoding deception or distortion. The former involves the action of one individual against another (or themself) or the (possibly passive) intervention against one input towards another. Both of these causes introduce eventivity into the meaning of a predicate and removing them from that meaning, all else being equal, results in deviance per the Predicate Lexicalization Constraint. Whence we conclude that stativity is a property that no predicate that has the contrafactive inference can have.

5 Further substantiation of the empirical claim

In §2, we observed that languages provide various ways of ascribing false belief to individuals without using contrafactive predicates, and in §3, we proposed that contrafactives are unlexicalizable because they have causally deficient denotations. We also observed with the cases of *wijsmaken* and *hallucinate* that predicates which look contrafactive on the surface can fail to meet one or more of the criteria for true contrafactivity.

In many contexts, even the simple use of verbs like *believe* or *think* gives rise to the inference that their complement is false:

(34) Mary {believes, thinks} that I have a sister.
⇝ The speaker does not have a sister.

Standard analyses of *believe* and *think* model these verbs as doxastically neutral, without incorporating truth or falsity into their lexical semantics. Neutrality is motivated by the existence of many uses of these verbs that do not give rise to the falsity inference in (34), and by the success of analyses that derive this inference through pragmatic means. According to the 'antipresuppositional' view (Percus, 2006; Chemla, 2008), for example, sentences like (34) compete with presuppositionally stronger alternatives containing *know*. If the speaker did have a sister, they would have said 'Mary knows that I have a sister.' From the fact that they used *believe* instead, we infer that they do not have a sister (and similarly for *think*).

In contrast to English *think*, *believe*, and their similar neutral counterparts, a number of belief-ascribing predicates have been described, across many unrelated languages, that give rise to stronger, more conventionalized, or more context-invariant inferences to the effect that their complement is false. In this section, we review representative ones that data has been published, presented, or collected by us on, and conclude that for none of them do we have sufficient reason to believe it is contrafactive.

5.1 Candidates that imply, but do not presuppose, the falsity of their complement

5.1.1 Candidates with weak falsity inferences

Our first category of putative counterexamples includes predicates that typically give rise to a false belief inference, but which is cancelable, and therefore not presupposed. This category includes predicates from a diverse array of languages, including Mandarin *yǐwéi* (Glass, 2022), Kipsigis *par-* (Bossi, 2023), Tagalog *akala* (Kierstead, 2013), German *wähnen* (Sander, 2025), and Toba Batak *agam* and *rimpu* (Silitonga, 1973; Rosenberg, 1975).

Looking first at Mandarin and Kipsigis, which exhibit many similar properties, Glass and Bossi show that the inferences that they describe (a) do not obtain for all unembedded positive uses of *yǐwéi* and *par-* respectively and (b) are reinforceable and cancelable. This leads Glass and Bossi *not* to propose that these verbs are contrafactive in the sense that they presuppose that their complement is false, and they derive these verbs' false belief inferences indirectly (in different ways to each other).

We illustrate the base case of *yǐwéi*'s falsity inference with (35).

(35) tā **yǐwéi** tā huì yíng
3S YǏWÉI 3S will win
'She is under the impression that she's going to win.'
(adapted from Glass 2022: ex. 5)

Such examples suggest that *yǐwéi* is a bona fide false belief predicate. Glass provides a range of reasons for thinking these falsity inferences are stronger and less context dependent than the ones that might be associated with, e.g., English *think* or *believe*. Among them, for this

sentence, is her comment that "[it] could not be used by an impartial journalist, but only by a highly biased pundit" and that replacing *yǐwéi* with the more neutral *rènweí* in the same sentence produces an alternative that "could be used impartially."

As Glass shows, however, *yǐwéi* may sometimes be used in contexts where the speaker merely *doubts* the truth of the attitude holder's belief, rather than outright believing it to be false. This is unexpected if the predicate directly presupposes the falsity of its complement. Moreover, the falsity inference described for (35) is both reinforceable and cancelable, with the latter case shown in (36). (As the lead in implies that the embedded subject is not a billionaire, to cancel this inference is to assert that she is—a "non-contradictory (though marked) continuation.")

(36) rénmén **yǐwéi** tā shì yìwànfùwēng… ér tā díquè shì.
person.PL YǏWÉI 3S be billionaire and 3S indeed be
'People are under the impression that she's a billionaire… and she actually is.'
(adapted from Glass 2022: ex. 9)

Data parallel to *yǐwéi* are reported by Bossi (2023) for Kipsigis *par*. A *par*-report like (37) implies that the speaker is not sick. (Like Mandarin, Kipsigis features a neutral belief predicate *pwaat*, to which the strength of *par*'s falsity inference can be compared.)

(37) **ø-pɑr-e** kaamɛɛ-nyʊʊn ɑɑ-mnyon-i.
3-PAR-IPFV mother-my 1SG-be.sick-IPFV
'My mother is under the impression that I'm sick.' (Bossi, 2023: ex. 1b)

As with *yǐwéi*, the falsity inference that *par* gives rise to is also reported to be reinforceable and cancellable, although the latter option, illustrated in (38), requires more contextual support and is a marked discourse move.

(38) My friend Lydia invented a famous app, and people think she made millions from it. Actually, although my friend never made any money from her app, she inherited money from her parents.

∅**-pɑr-e** piik mogoriot Lydia oko ɛɛn ɪman ko mogoriot. Lakini moo
3-think-IPFV people rich.person L. and in truth 3 rich.person but NEG
mogoriot kiin ko-ɑldɑ ap. Kii-goo-chi siigiik-chik rabɪɪnɪk.
rich.person when 3-sell app PST-give-APPL parents-3.POSS money
'People are under the impression that Lydia's rich and she actually is. But she's not rich from selling the app. Her parents gave her the money.' (Bossi, 2023: ex. 38)

The fact that the falsity inference associated with *yǐwéi* and *par* alike disappears given the right contexts suggests that it is not a true entailment of the verb–and certainly not a presupposition. Indeed, these facts lead both Glass and Bossi to derive the 'basic' false belief meaning indirectly. Glass (2022) proposes that *yǐwéi* postsupposes that its complement not be added to

the common ground, where Bossi's analysis of *par* imposes a disjunctive requirement on the CG: the output CG must not contain p (like *yǐwéi*), or the input CG must contain $\neg p$.

Turning to Tagalog, we see similar facts. Kierstead (2013) argues that the predicate *akala* gives rise to the conventional implicature that its complement is false. Conventional implicatures are usually thought to be uncancelable inferences (Potts, 2005). The sentence in (39) is reported to be acceptable in a context where the speaker knows that there is no party, and unacceptable in particular (a) when the speaker knows that there *is* a party or (b) when they don't know whether there is one. Thus, the sentence seems to commit the speaker to the falsity of *akala*'s complement.

(39) Akala [ni Jim] [na may party].
falsely.believe IND Jim COMP exist party
'Jim falsely believes there's a party.'
(Kierstead 2013: exx. 9–11; gloss and translation from the original)

Taken in isolation, this data point could suggest that *akala* is more of a candidate for contrafactivity than *yǐwéi* and *par*, but it isn't enough on its own to demonstrate that *akala* is a true contrafactive. One way to test this is to see whether an *x akala p* sentence can be uttered in a context where the speaker is explicitly unaware of whether p. Kierstead does not report such an example, but absent data examining the projectivity of the falsity inference, we cannot be certain whether or not *akala* is contrafactive.

However, we provide here one piece of evidence that the falsity inference may not be as strong as Kierstead suggests. In our constructed example (40), which we have modeled after Glass (2022: ex. 7) the speaker declares their ignorance about the truth of *akala*'s complement, yet the continuation is felicitous. This should only be the case if the falsity inference is defeasible here as well (which is reflected in our gloss and translation below).

(40) Di ko alam kung nakascore siya, peru akala niya siguro nakascore
NEG I know Q was.able.score 3SG, but AKALA 3SG perhaps was.able.score
siya.
3SG
'I don't know whether he was able to score, but he thinks that he might have scored.'
(Jed Pizarro-Guevara, p.c.)

While there is a tension between what Kierstead claims and our observation in (40), we are not entirely sure why that may be, and more data is evidently needed to adjudicate this difference. Given that cancelability was found to be a marked discourse move in Mandarin and Kipsigis as well, perhaps tighter control of discourse context could help settle the debate. At a minimum, in any case, we lack the kind of data to concretely declare *akala* a contrafactive in our sense.

Finally, a note about the German verb *wähnen*. This is perhaps the first documented candidate of a contrafactive predicate (see the discussion of Frege's use of the predicate in Holton 2017) and it is directly analyzed as such more recently by Sander (2025). Pace Sander, in modern

German, we found that *wähnen* does not reliably embed *that*-clauses for many of our German consultants, so it is difficult to ascertain whether or not it is contrafactive.

Wähnen does however reliably occur with small clauses, and in many such cases the false belief inference is present (41). However, this seems to depend in part on the context and content of the small clause, because out-of-the-blue utterances of sentences with *wähnen* sometimes carry only marginal or nonexistent inferences of falsity (42), suggesting that the false belief inference, when present, comes from a source other than inherent contrafactivity of the verb:

(41) Er **wähnt** sich sehr sicher.
he WÄHNEN himself very safe
'He took himself to be very safe.'
⇝ He is not safe. (Doris Penka, p.c.)

(42) Ich **wähne** mich glücklich.
I WÄHNEN myself happy
'I consider myself happy.'
~~⇝~~ I am not happy. (Doris Penka, p.c.)

But, as Sander points out, it could be that the *wähnen* of the late 19th century really was contrafactive, even if the *wähnen* of today is not. Frege's own description of *wähnen* and those of contemporaneous dictionary definitions are certainly suggestive of contrafactivity, ascribing to it definitions like 'wrongly assume' and 'believe erroneously'. Frege offers the following example, in which the speaker's belief is evidently that returning Alsace-Lorraine would not appease France's desire for revenge:

(43) Bebel wähnt, daß durch die Rückgabe Elsaß-Lothringens Frankreichs
Bebel WÄHNEN that through the return of.Alsace-Lorraine France's
Rachegelüste beschwichtigt werden können.
desire.for.revenge appeased become could
'Bebel *wähnt* that by returning Alsace-Lorraine, France's desire for revenge could be appeased.' (Frege 1892: 47-48)

Sander describes the convergent descriptions by various sources of the time as a 'wealth of clear evidence' on the contrafactive meaning of *wähnen* around the turn of the 20th century. However, the type of data we seem to have most abundantly is informal descriptions of the meaning of *wähnen*, which are not quite enough if we want to really put the hypothesis that it is contrafactive to the test. To consider a parallel example, the English idiom *be under the impression*, a clause-embedding predicate, typically conveys falsity of the embedded clause (44). This falsity is not present in all contexts; for instance, present-tense *be under the impression* can be used with first-person subjects to indicate a sense of uncertainty, rather than falsity (45).

(44) Gwyneth is under the impression that her accountant is following the law.
⇝ Her accountant is not following the law.

(45) I am under the impression that Gwyneth's accountant is following the law, but I would need to double-check the statute to be sure.

While (45) provides good reason to believe that *under the impression* is not contrafactive, modern dictionary definitions of *under the impression* (46) frequently include an explicit reference to the subject's beliefs being false:

(46) a. "Thinking, assuming, or believing something...often [suggesting] that the idea or belief one had is mistaken" (dictionary.com)
b. "To believe that something is true when it is not." (Longman Dictionary of Contemporary English)
c. "If you are under the impression that something is the case, you believe that it is the case, usually when it is not actually the case." (Collins English Dictionary)

While these definitions are importantly not unanimously categorical in treating the belief expressed by *under the impression* as false, they demonstrate that dictionary definitions alone are not reliable enough to conclude one way or the other about the fine semantic details of a particular predicate. When it comes to Frege-era *wähnen*, we lack a lot of the kind of information we would want to see in determining whether or not it is contrafactive, such as data about the cancelability, reinforceability, and projectivity of falsity inferences, and we can also not collect them since we no longer have access to speakers of a bygone form of German.

5.1.2 Falsity inferences which can disappear

Predicates have also been described that imply the falsity of their complement more strongly than the ones reviewed in the previous section. An example that receives current attention here is the Spanish predicate *creerse* (Anvari et al. 2019, Maldonado and Percus 2024), which, at least in its present tense uses in positive declaratives, gives rise to a *non-cancelable* inference that its complement is false. This is illustrated in (47) by the infelicity of the continuation given.

(47) Juan se cree que Ana tiene 30 años... #¡y tiene razón!
Juan REFL believe that Ana has 30 years and has right
Intended: Juan believes that Ana is 30 years old... and he's right!
(Anvari et al., 2019: ex. 10)

This inference can also be shown to project in a way that is characteristic of semantic presuppositions. For example, if the falsity inference associated with *creerse*, but not non-reflexive *creer*, projects universally from the scope of a universal quantifier, as Anvari et al. (2019) argue, it is correctly predicted that a *creerse* report like (48b) with a universal subject binding into the embedding clause is infelicitous, given that there is some value of the bound variable that in fact renders the embedded clause true.

(48) *Context: in the aftermath of a swimming match, the only possible outcome of which is that exactly one of the competitors wins.*

a. Cada una de las nadadoras cree que ha ganado la carrera.
'Every swimmer believes that she has won the race.'

b. # Cada una de las nadadoras **se cree** que ha ganado la carrera.
'Every swimmer REFL believes that she has won the race.'

In terms of the stickiness of the falsity inference, *creerse* is a good candidate for being a contrafactive. In addition to its inferential profile illustrated in (48) and (47), it is stative and doxastic. One could say that it isn't technically unanalyzable because it travels with the reflexive morpheme, but this argument would not stand on its own, since the meaning of *creerse* is not obviously composed of the meanings of *creer* and *-se*. Even though there does seem to be a non-accidental connection between reflexive predicates and ones that imply the falsity of their complement (such that we would want to investigate this link further), there is not a clear sense in which the falsity inference is encoded in the reflexive construction (cf. suffixes like *mis-*).

With that in mind, there is nevertheless a more substantial reason for which one might resist accepting *creerse* as contrafactive: Whether or not the predicate gives rise to the falsity inference described above depends on factors like negation on the predicate interacting with indicative or subjunctive in its clausal complement (Anvari et al. 2019, Maldonado and Percus 2024), as well, possibly, as grammatical aspect on the predicate (Maribel Romero, p.c.). As shown in (49), when negated *creerse* combines with an indicative complement, the resulting sentence is ambiguous between giving rise to the inference that its complement is false vs. the inference that its complement is true (a factive inference).[14]

(49) Hilda no se cree que Jirafales es honesto. Indicative
Hilda not REFL believe that Jirafales is.IND honest.

a. → Jirafales is not honest and Hilda doesn't believe he is.

b. → Jirafales is honest and Hilda believes he isn't.

Maldonado and Percus (2024: ex. 5)

In (50), we are shown that the predicate comes with neither inference when its complement is in the subjunctive.

(50) Hilda no se cree que Jirafales sea honesto. Subjunctive
Hilda not REFL believe that Jirafales is.SUBJ honest.
→ Hilda fails to believe that Jirafales is honest.

Maldonado and Percus (2024: ex. 7)

[14]Note that Anvari et al. (2019: ex. 4) originally only describe the predicate as factive under these conditions. This work does adopt an analysis of *creerse* as presupposing the falsity of its complement, but this requires commitment to the syntactic analysis of neg-raising (see e.g. Collins and Postal 2014). Such an analysis is controversial, and it is unclear how such an analysis would deal with the mood alternation facts.

Along with Maldonado and Percus (2024), we take these facts to argue against an analysis where *creerse* is an unanalyzable predicate that encodes the presupposition that $\neg p$. These authors, for example, suggest that it is made up of a (silent) morpheme WRONGLY and a plain belief predicate, which gives them the freedom to have these elements interact with negation and mood. If their analysis is correct, then *creerse* is not a contrafactive anyway: it is semantically complex.

5.2 Candidates that violate at least one of the other criteria

So far, we have seen several predicates which violate the flagship condition of contrafactivity: they fail to **presuppose** falsity, even if they strongly generate false belief inferences, and in the previous section, we discussed candidates which fail to satisfy the stativity criterion. But this does not demonstrate that presupposing falsity is impossible: in fact, some other candidate predicates do seem to presuppose falsity, but as we will show, each of these candidates also appear to fail at least some of the other three criteria for contrafactivity.

5.2.1 Candidates that are not monomorphemic

The contrafactive gap is essentially a constraint on lexicalizability: are contrafactive meanings lexicalizable? As we alluded to with *wijsmaken*, it is not always easy to tell whether some morphologically complex expression is semantically atomic or not. While we think the difficulty of diagnosing lexicalization is a good reason to not reject potential contrafactives purely on the grounds of complex morphology, we should also point out that there are some false belief predicates which do nevertheless seem to clearly divide the labor of falsity and belief, which at the very least casts serious doubt on their unanalyzability. An example of such a predicate is English *disprove*. The falsity inference seems to arise from the prefix *dis-*, since the truth of the embedded clause is implied by the use of *prove* (51b) instead:

(51) a. Recently science disproved that breakfast was the most important meal of the day.
→ Breakfast is not the most important meal of the day
b. Recently science proved that breakfast was the most important meal of the day.
→ Breakfast is the most important meal of the day

We of course do not want to rule out the existence of lexical items that are multimorphemic yet semantically non-compositional (idioms exist, after all). But while *disprove* itself may be ruled out as a contrafactive anyway (it is not clear to us that the falsity inference is projective), we can exclude *misremember* and *disprove* as candidates on the grounds that their falsity inference can transparently be traced to a proper subpart of affixal morphology, rather than the lexical semantics of a verbal root. Such sentences may also be supplemented by modifiers like *wrongly*, *mistakenly* or *falsely* to reinforce (or add) this inference of falsity, but this is also the case for vanilla belief verbs like *believe* and *think* in any event.

5.2.2 Candidates that are not doxastic

The last set of predicates with false belief inferences that we would like to consider are ones like *pretend* and *lie*, which, to the extent that they are compatible with *that*-clause complements, definitely entail that they are false.[15]

(52) a. Alice lied that she liked her father's cooking.
b. Alice pretended that she liked her father's cooking.

Interestingly, this falsity inference also projects. Both positive and negated 'pretend' are compatible with propositions that we know are false but not ones that we know are true.

(53) a. Alice pretended that Stuttgart was in France / #Germany.
b. Alice didn't pretend that Stuttgart was in France / #Germany.

Despite presupposing the falsity of their conmplement, these predicates do not count as instances of contrafactives because they are not doxastic. This is to say that their main point is not to ascribe to their subject a belief—even though they entail something about their subject's beliefs, namely that they think that p is false. We can see that the belief is not at-issue, since it too projects: both the negated and un-negated *pretend*-reports in (53) require that the subject believes the complement of *pretend* to be false.

(These predicates are also not stative, with *lie* an achievement/accomplishment and *pretend* an activity.) The exchange in (54) is used to test whether the belief inference associated with 'pretend' has main point status. If the embedded clause can contribute the main point, then we expect a response which directly addresses the truth of that clause to be felicitous, but it is not:

(54) a. Alice pretended that Stuttgart is in France.
b. No she didn't.
c. #No it isn't.

Like *wijsmaken*, *pretend* describes an event of *acting* as though something false is true. So while both *pretend* and its ilk do seem to entail false belief—thus making them at least superificially plausible contrafactives—they do not presuppose it.

5.3 Predicates with inconclusive data

Finally, we wish to raise a class of possible candidate contrafactives identified in various corners of literature that very well could fit the bill, but do not yet have sufficient evidence to be classed as bonafide contrafactives under our definition. We present here two such examples.

The first example is the Taiwanese Southern Min verb *liah8-tsun2* (Hsiao, 2017). Hsiao describes this predicate as 'counterfactual', and as she shows, it is associated with a typically

[15] Holton (2017) dismisses 'lie that' as he considers the construction ungrammatical.

quite strong falsity inference, which cannot be cancelled. This is shown by the infelicity of following up an utterance of (55a) with (55b):

(55) a. gua2 liah8-tsun2 il tsa1-hng1 e7 lai5.
I think he yesterday will come
I thought he would come yesterday.

b. #kiat4-ko2 il tsa1-hng1 tsin1-tsiann3 u7 lai5.
As.a.result he yesterday really ASP come
As a result, he really came yesterday.
(adapted from Hsiao 2017: exx. 40a-b, glosses and translations from the original)

For reasons that are not (to our knowledge) understood, this predicate cannot be negated (56). This means we cannot test for projectivity of the falsity inference with negation, and Hsiao provides no examples which demonstrate projectivity through other operators.

(56) *i1 bo5 liah8-tsun2 a1-ing1 tsa1-hng1 kah4 ong5-sian1-sinn1 tso3-hue2.
He NEG think A-ing yesterday with Wang-Mr. be.together
'He didn't mistakenly think that A-ing was with Mr. Wang yesterday.'

Notably, given the semantics that Hsiao herself proposes for *liah8-tsun2* (Hsiao 2017: ex. 57), the predicate is also not a contrafactive in our sense: it does not presuppose $\neg p$, in that $\neg p$ is entailed by the input common ground; rather, the verb presupposes that the common ground does not entail p, a weaker claim. So while the data is largely compatible with the hypothesis that *liah8-tsun2* is a contrafactive, more data is need to stress-test the falsity inference.

Our second intriguing but inconclusive example comes from McGregor (2023), who describes grammaticized expressions of 'mistaken belief' as a feature of many indigenous languages of Australia. These expressions take many different potential forms, including particles, enclitics, and (possibly conventionalized) pragmatic implications.

For our purposes, two classes of languages described by McGregor are most relevant: mistaken belief verbs and nominals, since those are the closest possible candidates to our contrafactives. In the end, McGregor identifies only one clear candidate for a true contrafactive: the Yanyuwa (Pama-Nyungan) verb *yudirri*, glossed as 'erroneously/mistakenly think', discussed by Kirton and Charlie (1996). As McGregor notes, in all of Kirton & Charlie's examples involving *yidirri*, it is always accompanied by the 'mistaken belief' particle *katha*.

(57) Katharra-**yudirri** katha babalu ngala kurdardi ngala wakardawakarda.
we.DU.EXCL-YUDIRRI KATHA buffalo but no but bull
'We mistakenly thought a buffalo (was there) but (it was) not, (there was a) bull.'
(Kirton and Charlie 1996: 205)

While examples like (57) are certainly compatible with a contrafactive semantics for *yidirri*, we do not have enough evidence to categorize it one way or the other. Kirton and Charlie

(1996) provide only a handful of examples of its use (p. 205-206), none of which are the sort we might like to find if we were looking for presuppositions, such as embedding the verb under a downward-entailing operator. Moreover, the fact that *yudirri* always co-occurs with *katha* in the reported sentences makes it difficult to assess the contribution of either element on their own.

5.4 Interim summary

After surveying a number of plausible candidates for 'true' contrafactives from the literature, we conclude that they all either fail to exhibit one of the necessary traits for contrafactivity under our (and Holton's) definition, or we lack enough information to be certain that they are contrafactive. From this, we conclude that the contrafactive gap is real: whether this is a categorical ban on predicates with contrafactive meaning or simply a strong typological tendency remains to be determined.

We would like to stress that a major limitation of studies on contrafactivity in general is the sparsity or inconsistency of data. Many authors discuss expressions' 'false belief', but it is not always clear whether this refers to contrafactivity in our sense. While we do not have conclusive evidence for a contrafactive predicate so far, it is nevertheless striking that contrafactives are at the very least hard to come by, if not outright nonexistent. At the very least, we hope the discussion in this section can provide a template for the kind of data we would need to collect in order to determine whether or not something is contrafactive; in particular, we would hope to see tests about the cancellability or reinforceability of falsity inferences.

6 Presuppositions beyond *contra*

If our lexicalization constraint holds, we predict that attested predicates should not violate it. As a proof of concept, in this section we examine several classes of English presuppositional clause-embedding predicates, including several that appear superficially intransigent for our constraint, while ultimately concluding that they do in fact obey our constraint if we dig deeply enough into their lexical semantics.

6.1 Change of state predicates

Roberts and Simons (2024) (henceforth R&S) aim to provide an ontological account of several different kinds of presupposition triggers and their associated projective behavior, including change of state (CoS) predicates.

In R&S's view, presuppositions are not lexically specified, but rather a predictable consequence of the relations of parts of lexical meanings. The idea is this: predicates don't encode their presuppositions as such, but rather our (perhaps extralinguistic) mental event ontology imposes certain requirements on events. We note here that while we have treated presuppositions as lexical, in principle this is not necessary for our core idea to go through; we could well think of our lexicalization constraint as a constraint on ontologically possible events as R&S do.

For instance, R&S consider predicates which denote changes of state, such as *fall off*. Intuitively, the kind of event which is described as 'falling off' involves a very particular kind of transition, namely from being on something (like a ladder) to no longer being on that thing. So while *y fell off x* does **presuppose** that *y was on x*, R&S suggest the not-at-issue status of this inference is is not annotated in the lexicon, but rather derived in virtue of being an inherent precondition of other entailments of the predicate.

(58) Jane fell off the ladder. (modified from R&S ex. 5)
Entailed: Jane transitioned from being on the ladder to being off the ladder by means of falling.
Precondition: Jane was on the ladder.

To put it simply, change of state predicates lexicalize a transition from state s_1 to state s_2, of which the holding of s_1 at some prior point is an indefeasible requirement. Ontological preconditions of this sort adhere quite straightforwardly to our PLC, since if p is a necessary precondition of q, p is causally upstream of q in our sense.

An additional prediction of our account is that predicates which lexicalize only states which coincide with result states of CoS predicates–for example, the state of being off the ladder–but presuppose some prior different state–like being on the ladder–should not be lexicalizable. This would be a hypothetical predicate like *shmell off* which has the following interpretive profile:

(59) Jane shmell off the ladder.
Entailed: Jane was/is off the ladder.
Presupposed: Jane was on the ladder at some point in the past.

To our knowledge, no predicates of this kind exist, which gives some additional credence to R&S's point that the presupposition of CoS predicates is an emergent feature of change of state lexical semantics.[16] In fact, we believe R&S's account and the present paper complement one another nicely: where R&S explain why some inferences are presupposed and projective, we are trying to explain why some conceivable presuppositions don't exist. Both of these are argued to follow from general constraints on the relation between lexicalization and more domain-neutral aspects of cognition.

6.2 Response-stance predicates

With respect to the preconditions of clause-embedding predicates, one coherent class of presuppositionally interesting verbs are so-called response-stance predicates. A response-stance predicate V has the property that a sentence of the form *a V p* (a) presupposes that someone expressed a public commitment to p and (b) asserts that a expresses or comes to believe p **in response to** that public commitment (Cattell, 1978). Predicates of this category include *accept, admit, agree, confirm, deny, reply,* and *verify*, among others. We can reasonably

[16] Notice that the hypothetical verb *shmell* is similar to the *wijsmaken* when we removed it of its meaning component of deception, in section 4: A juxtaposition of different states, with no transition between them.

assume that the prior commitment is presupposed since it projects through presupposition holes like negation (60).

(60) a. Marianne accepted that the Earth was flat.
→ Someone said that the Earth was flat.
→ Marianne came to believe that the Earth was flat.
b. Marianne didn't accept that the Earth was flat.
→ Someone said that the Earth was flat.
→ Marianne did not come to believe that the Earth was flat.

Since response-stance predicates are clearly attested, the PLC indicates there should exist a stimulatory causal link between their presupposed inference *someone commits to p* and the at-issue inference *a responds to someone's commitment to p*. We believe that this is the case. To illustrate, consider the following link for *respond*:

$$committed\text{-}to(x)(p) \longrightarrow respond(a)(p)$$

$$respond(a)(p) := 1 \text{ if } committed\text{-}to(x)(p)$$

The intuition here is that *responding*, like other response-stance events, inherently describes making an utterance *as a reaction to* someone's commitment that p (as expressed, for example, by publicly asserting it). In this sense, response-stance verbs are quite well-behaved, since they presuppose content which is a necessary precondition for their at-issue content.

6.3 *Be wrong*

A closer kin to *contra* that is also attested is *be wrong*, since it implies, in positive assertive contexts, the falsity of its prejacent. However, as we suggested in (3), *wrong* seems to presuppose an attitude holder's commitment to p and **assert** the falsity of p, rather than presupposing $\neg p$ (Anand and Hacquard 2014). This gives us roughly the denotation in (61):

(61) $[\![\text{be wrong}_{\text{A\&H}}]\!] = \lambda p_{st} \lambda x_e : \textbf{committed-to}(x)(p).\neg p$

This denotation seems problematic for our constraint, since the subject's commitment to p obviously does not systematically result in the fact of $\neg p$. (One's commitment to a proposition clearly does not lead to the falsity of that proposition in the general case.) However, not all hope is lost if we change the camera angle. Rather that viewing the at-issue content as the falsity of p full stop, we propose that *be wrong* should be treated more like a response-stance predicate, in that it asserts the falsity of a presupposed **commitment**.[17]

(62) $[\![\text{be wrong}]\!] = \lambda p_{st} \lambda x_e : \textbf{committed-to}(x)(p).\neg \iota q[\textbf{committed-to}(x)(q)]$

[17] The domain of the iota operator in (62) is assumed to be restricted to the relevant situation such that x's commitment in this situation is unique, à la the kind of domain restriction necessary to explain the uniqueness presupposition of definite descriptions in general.

While (62) is truth-conditionally equivalent to (61), it differs in what entailments must be derived (vs. lexically encoded). Anand & Hacquard's *wrong* takes 'x is committed to p' and 'p is false' as basic, deriving that x's commitment is false; our *wrong* takes the conjunction 'x is committed to p' and 'x's unique commitment is false' to derive that p is false. This definition yields a causal configuration that is effectively the same as that of response-stance predicates, whereby x's commitment to some p is a necessary precondition of their unique commitment being false.

$$\textit{committed-to}(x)(p) \longrightarrow \neg\iota q[\textit{committed-to}(x)(q)]$$

$$\neg\iota q[\textit{committed-to}(x)(q)] := 1 \text{ if } \textit{committed-to}(x)(p) = 1$$

However, the truth-conditional equivalence of our definition and Anand & Hacquard's poses a challenge to independently motivating this move. At worst, it could indicate that with enough massaging, any deviant meaning could be whipped into PLC-compliance. We will suggest that some additional evidence points to *wrong*'s at-issue content not merely being that p is false. For instance, *wrong*-reports are natural if what is under discussion is the reliability of a claim (63). However, they are less natural if the topic is the truth/falsity of p itself (64); such cases are acceptable insofar as the use of *wrong* can be understood as parenthetical:

(63) A: How were John's Oscar predictions?
B: Awful. He was wrong that *Tár* would win Best Picture.

(64) A: Did *Tár* get the Best Picture Oscar?
B: ?No, John was wrong that it would win.

6.4 *Forget*

A last interesting case is English *forget*. In the frame *a forgot p*, *forget* is a predicate that is factive and that implies that *a* does not hold the belief that *p* at the time the sentence is asserted. The predicate also implies that *a* did hold the belief that *p* at a past time. These three inferences are exemplified in (65).

(65) Philine forgot that Enschede is in the Netherlands.
⇝ Philine used to believe that Enschede is in the Netherlands.
⇝ Philine does not believe (now) that Enschede is in the Netherlands
⇝ Enschede is in the Netherlands.

Which of these inferences are entailments, which are presuppositions? Examples like (66) suggest that the predicate is factive in that it presupposes its complement. The infelicity of (66b) can be explained if the sentence gives rise to the inference that Enschede is in Germany, which is false at the world of evaluation.

(66) a. Philine didn't forget that Enschede is in the Netherlands.
b. #Philine didn't forget that Enschede is in Germany.

Likewise, the 'a has a prior belief that p' inference also seems to project through typical presupposition holes, suggesting that it is likewise presupposed:

(67) a. She didn't forget that I'm Turkish.
b. Did she forget that I'm Turkish?
c. If she forgot that I'm Turkish, I'll eat my hat.
d. Maybe she forgot that I'm Turkish.
$\rightsquigarrow$ She used to believe that I'm Turkish.

On the other hand, (67a) asserts that she still believes p, (67b) asks whether she still believes p, and so on, indicating that the inference that a does not presently have the belief that p is at issue. Putting these facts together, we may start out by writing (68), where underlined conjuncts are understood to be presupposed.

(68) $[\![\text{forget}]\!]^t(p)(x) = \underline{p \wedge \exists t'[t' < t \wedge B_{t',\,x}p]} \wedge \neg B_{t,\,x}p$ (first pass)

Given this lexical entry, the PLC requires that we should be able to construct a causal chain between $p \wedge \exists t'[t' < t \wedge B_{t',x}p]$ and $\neg B_{t,x}p$ in the general case, for any t. But, at first sight, it is hard to see how *forget* could adhere to the PLC, since having a past true belief that p is neither necessary nor sufficient to cause the current *lack* of a belief that p: Not sufficient because one could (come to) believe p once and then always keep believing it; and not necessary because one could fail to believe p without ever having believed it in the first place.

But, let us examine this second claim more closely. There are two ways in which $\neg B_{t,x}p$ could be true: As mentioned, it could be the case that x never believed p and still doesn't, or, alternatively, that x used to believe p and no longer doesn't. The former possibility is nakedly incompatible with the prior belief presupposition of *forget*, so it will never hold in any situation which verifies *a forgot p*. The latter case, where forgetting implies a change of state resulting in $\neg B_{t,x}p$, must then be part and parcel of the meaning of *forget*.

Despite this result, the relationship between the past belief conjunct and the present lack of belief conjunct is still too loose in (68). To see this, first observe that example in (69b) is not true in the context in (69a) despite the fact that the context makes all three conjuncts in (68) true. This suggests that there is something missing from the denotation in (67).

(69) a. At first, Sam correctly held the belief that I'm Turkish. Then, she came to believe (instead) that I'm French. Then, she came to believe (instead) that I'm Greek. She comes up to me and says "Kalimera!" I tell my friend:
b. ?Sam forgot that I'm Turkish.

Intuitively, what fails in the example in (69) is that the subject's current belief state, which satisfies $\neg Bp$, is not temporally adjacent to the prior belief state, which satisfies Bp, that the sentence presupposes. If the middle belief state were not there and Sam's belief that I'm Turkish was succeeded by the belief that I'm Greek, the sentence in (69b) would have a better shot at becoming acceptable and true again.

The reason we say this is that temporal contiguity between Bp and $\neg Bp$ is also not enough, as can be seen in the oddness of example (70b) in the context given. This example suggests that the belief that p must naturally decay into a belief state that is not a belief that p.[18]

(70) **Context**: In the movie *The Matrix*, people can be connected to a computer and skills loaded into their brains. Assume that beliefs can be loaded and unloaded in the same way, and that Neo held the belief that Link is from Zion, that that belief is unloaded with the help of a computer.

?Neo forgot that Link is from Zion.

Given these observations, we suggest that a more accurate representation of *forget*'s definedness and truth conditions are as stated in (71).

(71) $[\![\text{forget}]\!](p)(x) = \lambda s.\underline{p \wedge \exists s' : B_x(s', p)} \wedge p \wedge \neg B_x(s, p)$
$\wedge \iota s''[B_x(s'', p)]$ changes into into s

This entry presupposes that *forget*'s complement is true and that there is a state of x believing p. In addition to repeating the content of this presupposition in its assertion, it asserts that there is a state that is not a state of x believing p and that x's positive belief that p state changes into the absence of a belief that p state. We have provided the predicate with an eventuality argument to be able to talk about belief states more easily—following, e.g., Hacquard (2006) and other authors' treatment of belief predicates—but nothing crucial hinges on this assumption.

Here, then, for the assertion to be true, it is necessary for the presupposition to be true, which makes it possible (in fact, requires) that there be a causal chain between the former and the latter. This is enough for *forget* to satisfy the PLC. In effect, *forget* functions as a change of state predicate like any other, thus making it amenable to a Roberts and Simons (2024)-style analysis.

6.5 Postsuppositional predicates

Glass (2023) describes the Mandarin verb *yǐwéi*, which is often used to ascribe false beliefs, and analyzes it as a belief verb which is not presuppositional but *post*suppositional: it imposes constraints on output contexts, rather than input contexts. In particular, she argues that *x yiwei p* is felicitously used just in case p remains up for debate after the entire sentence has been added to the common ground. In many discourse contexts, this generates false belief-like inferences, since the speaker explicitly signals that p should not be wholeheartedly adopted in virtue of this postsupposition:

(72) a. $c + x$ *yǐwéi* p = $c + x$ believes p
b. defined only if $\exists w \in (c + x \textit{ believes } p) : p(w) = 0$ (Glass 2023: ex. 22)

[18] Further refinements can be made, which we leave for another occasion. For example, *forget* is also not able to describe situations in which prior knowledge states are intentionally revised into non- or false belief states.

Taking Glass's analysis for granted, our proposal trivially predicts *yiwei* to be lexicalizable, since it imposes constraints on presupposition, not postsupposition. If we also assume that postsuppositions of this kind are not only possible but attested across languages, we might well seek a similar explanation for what kinds of postsuppositions we do(n't) see. Glass (2025) argues that this kind of postsupposition is attested in multiple languages (and that reanalyzing them as presuppositions doesn't capture the facts). If her reasoning is correct, then it gives more grist for the mill of the idea that what is special about presuppositions is their anteriority, connected in a particular way, to at-issue content.

7 Alternative accounts

Finally, we consider some alternative explanations for the contrafactive ban in terms of different grammatical or extragrammatical limitations. We might think that the impossibility of verbs ended speaker-subject belief misalignment is because such situations more marked (and therefore rarer) than alignment.

This intuition is supported by diverse kinds of evidence. For example, false belief reports are often misunderstood by preschool children (de Villiers and Pyers 2002; Lohmann and Tomasello 2003, a.m.o.; see also citations in Hacquard and Lidz 2022). False beliefs are also arguably representationally complex, since agents cannot simply map their model of the world onto another's mental state, but rather need to hold two representations in mind at once–their own and that of the false believer (Phillips and Norby, 2021).

We believe the common denominator among all of these observations is that predicates which encode some kind of negation are somehow strange or deviant. It stands to reason that the contrafactive gap may actually be an instantiation of a much more general phenomenon. Below we outline a couple such possibilities for an explanation, and suggest that neither of them or their own are sufficient to explain the contrafactive gap.

7.1 Anti-presuppositions and mirror languages

One possible alternative reason for the contrafactive gap is that languages do not lexicalize false belief verbs because such beliefs can systematically be expressed by means of implicature generated by attitude reports involving (e.g.) *know*, à la Horn (1989) for the absence of the connective XOR. To put it another way, *contra* doesn't offer anything to a lexicon that cannot already be done with existing grammatical resources. While this alternative is plausible, we argue that it fails to account on its own for the asymmetry between *know* and *contra*.

To sketch the account, recall that attitude reports with unmodified uses of *think* and *believe* often describe false beliefs without any other overt marking, such as (73).

(73) Mary thinks that Rotterdam is in Belgium.

Sentences like (73) can (though need not) *imply* that p is false, rather than merely being assertable in contexts where p is false. In fact, if the speaker believes p, *think* and *believe* often sound odd, compared to *know*.

(74) a. #Mary thinks that Rotterdam is in the Netherlands.
b. Mary knows that Rotterdam is in the Netherlands.

Some have argued that the false belief inference triggered by *believe* and *think* is due to a combination of lexical competition with *know*, a verb with the same assertion but a stronger presupposition, and the pressure to presuppose as much as possible, i.e. the principle of "Maximize Presupposition!", (Percus, 2006; Sauerland, 2007; Chemla, 2008).

The reasoning process behind this falsity inference, familiar from scalar implicatures, goes as follows: In contexts where p is true, the utterance *x thinks/believes p* unassertable, because it is strictly less informative than the stronger alternative *x knows p*. Thus, when a speaker asserts *x thinks/believes p*, the inference is drawn that we are not in a context in which p is true (i.e. the speaker does not believe p), because the speaker would have used *know* if p was the case. In certain contexts (namely those in which the speaker is competent and an authority about p, for Chemla), this inference is strengthened to the speaker belief that not p, rather than just the absence of the belief that p.

This leads us to the following hypothesis, which we will reject:

(75) False belief verbs are not lexicalized in languages because false belief inferences can be systematically derived through pragmatic reasoning.[19]

In other words, false belief verbs would not be lexicalized because they are not needed, as false belief ascriptions are expressible by means of a vanilla belief predicate with an antipresupposition of nonfactivity. Importantly, this reasoning requires the existence of a factive alternative to nonfactive belief verbs.

However, consider the possibility of a hypothetical mirror language Hsilgne, which contains *believe* and *contra* but not *know*–i.e., it has contrafactives but no factives. In Hsilgne, false belief ascriptions can straightforwardly be expressed by *contra*, but the lexicon offers no simple way to express the meaning of *know*, i.e. true belief. However, Hsilgne can still communicate this concept using *believe*-reports, which can generate systematic true belief implicatures via competition between *contra* and *believe*. Using Chemla's algorithm, we can see how a *believe p* statement would give rise to the inference that p is true:

(76) a. Mary contras that Rotterdam is in the Netherlands.
(Assert $B(m)(p)$, presuppose $\neg p$)
b. Mary believes that Rotterdam is in the Netherlands.
(Assert $B(m)(p)$, implicate $\neg$(76a))

By parity of reasoning, (76b) is not assertable in contexts where p is false. Hence, when it is assertable, the inference arises that p is true. This mirror language is equivalently expressive to one whose lexicon contains *know* instead of *contra*, but to our knowledge, no such language exists. This pragmatic calculus in and of itself offers no explanation for why factives

[19] Interestingly, Percus conjectures that it is impossible for a language to have a belief verb that is neutral with respect to the truth of p if it has a factive belief verb like 'know.'

should be common (if not universal) across languages and contrafactives should be rare (if not impossible).

We could augment this hypothesis by saying that contrafactives are somehow more complex or costly than factives, so a language in which contrafactivity is basic and factivity is derived is improbable, but this amounts to a restatement of the original problem—why do we have factives but no contrafactives—to begin with. Thus, we need some additional explanation for the contrafactive gap.

7.2 Contrafactives are not communicatively useful

Another plausible derivation of the contrafactive gap could come from general pressures on communication, an analysis which has been profitably applied to many other lexical domains. In a nutshell, the core intuition is that natural language lexica are attempting to balance competing constraints for expressions to be maximally *informative* and minimally *costly*. This basic principle underlies many explanations for gaps in other lexical domains, including Boolean connectives (Horn, 1972, 1989; Katzir and Singh, 2013; Bar-Lev and Katzir, 2022; Carcassi and Sbardolini, 2022: a.m.o.), quantifiers (Horn, 1972; Keenan and Stavi, 1986; Enguehard and Spector, 2021; Sbardolini, 2023), and modals (Imel and Steinert-Threlkeld, 2022; Steinert-Threlkeld et al., 2023).

Applied to lexicalization, we can conceive of cost straightforwardly: it is more costly to have an item in the lexicon than not to have it, so only we should only lexicalize concepts whose communicative usefulness provides a sufficient benefit to outweigh this cost. So if contrafactives are unlexicalizable, it must be that they are not informative enough to justify their existence. But informative by what metric? One way for a verbal predicate to be uninformative is to describe situations that are either statistically marginal—they either occur so rarely that there is no need to dedicate a specific lexeme to them, or are so commonplace that using language to refer to them has little communicative import. In other words, situations of use for lexical items need to occupy a 'Goldilocks zone' of frequency: not too common and not too rare (Enguehard and Spector, 2021).

The intuition that contrafactives are not utilitarian enough to lexicalize is attractive, particularly in light of our analysis that contrafactives describe an 'abnormal' configuration of information. That being said, we are less inclined to argue that utility itself is the source of the contrafactive gap for both empirical and pragmatic reasons. First, *prima facie*, false belief situations don't seem inherently outlandish (and thus statistically remote) nor exceptionally commonplace. In fact, we use explicit language to ascribe false beliefs to others quite often, as a cursory news search of the string *falsely believes* reveals, suggesting that the usefulness of *contra* is not so thin as to discourage lexicalization:

(77) a. They concede that he **falsely believes** U.S. Supreme Court Justice Amy Coney Barrett visited him on death row.[20]

[20] https://www.texastribune.org/2022/10/28/texas-execution-scott-panetti/

b. The former Mesa County clerk faces criminal charges for using her position overseeing local elections to empower a movement that **falsely believes** U.S. voting systems are rigged to steal elections.[21]
c. The effects of the medication mirror the diseases Chloe Sherman **falsely believes** she has.[22]
d. One of the earliest operas to feature the idea of space travel, Haydn's *Il Mondo della Luna*, written in 1777, is about an astronomer who **falsely believes** that he's transported to the moon.[23]

Second, it is a major practical challenge to calculate the putative utility of *contra* versus alternatives like *know*. If utility of a verb is calculated in relation to the frequency of situations involving the eventuality that verb describes, we need some way of quantifying false belief situations relative to true belief ones. To put it differently, what proportion of beliefs that we want to talk about are false, and how could we know?

A utility-based approach may well provide a suitable explanation in the end for the contrafactive gap, or indeed provide a more general explanation for the PLC. We leave operationalizing this hypothesis to future work.

8 Conclusion

This paper started out with a puzzle: we have antiveridical and factive predicates, so why are there seemingly no (or at least few) contrafactives?

Our proposal was that there are restrictions on the internal configuration of information packaged within a single lexical entry: we cannot simply slap multiple entailments into one lexeme; they must be **related** in some way, which we cashed out in causal terms: within a single lexical predicate, the presupposed content must be causally prior to the at-issue content. This makes rather strong predictions for the kinds of predicates we find across languages, which will necessitate strong and careful cross-linguistic investigation to validate.

If the current proposal is on the right track, a major question which remains yet unanswered is the extent to which the PLC can be extended to other kinds of presuppositions, besides those encoded by eventuality-denoting predicates. For instance, additive presupposition triggers like *too* do not seem to require a close conceptual link between their at-issue content and presupposition content in the same way as *know*. This could suggest that if the PLC really does hold cross-linguistically, it is not encoded 'as such' in the human blueprint for language, but rather follows from a more general constraint on event cognition.

References

Abrusán M (2011) Predicting the presuppositions of soft triggers. Linguistics and Philosophy 34(6):491–535. https://doi.org/10.1007/s10988-012-9108-y

[21] https://www.cpr.org/podcast-episode/the-colorado-clerk-on-trial-for-the-big-lie-and-what-it-means-for-the-2024-election/
[22] https://screenrant.com/run-movie-chloe-sickness-poisoning-medicine-accuracy-explained/
[23] https://www.wqxr.org/story/11-composers-outer-space-pluto/

Abrusán M (2016) Presupposition cancellation: Explaining the 'soft–hard' trigger distinction. Natural Language Semantics 24(2):165–202. https://doi.org/10.1007/s11050-016-9122-7
Abusch D (2002) Lexical alternatives as a source of pragmatic presuppositions. In: Jackson B (ed) Proceedings of Semantics and Linguistic Theory (SALT), pp 1–19
Abusch D (2010) Presupposition triggering from alternatives. Journal of Semantics 27:37–80
Alexander P (1973) Normality. Philosophy 48(184):137–151. https://doi.org/10.1017/S0031819100060605
Anand P, Hacquard V (2014) Factivity, belief and discourse. In: Crnič L, Sauerland U (eds) The Art and Craft of Semantics: A Festschrift for Irene Heim, vol 1. MIT Working Papers in Linguistics, Cambridge, MA, p 69–90
Anvari A, Maldonado M, Soria Ruiz A (2019) The puzzle of reflexive belief construction. In: et al. MTE (ed) Proceedings of Sinn und Bedeutung 23, pp 57–74
Baglini R, Bar-Asher Siegal E (2020) Direct Causation: A new approach to an old question. Proceedings of the 43rd Annual Penn Linguistics Conference pp 19–28
Bar-Lev ME, Katzir R (2022) Communicative stability and the typology of logical operators. Linguistic Inquiry pp 1–42
Bernard T, Champollion L (2018) Negative events in compositional semantics. In: Semantics and Linguistic Theory, pp 512–532
Bledin J (2022) Composing negative individuals in truthmaker semantics. Amsterdam Colloquium presentation
Bondarenko T (2022) Anatomy of an Attitude. PhD thesis, MIT
Bossi M (2023) Negative bias, reminding, and pragmatic reasoning in Kipsigis belief reports, Ms., University of California, Berkeley
Carcassi F, Sbardolini G (2022) Assertion, denial, and the evolution of Boolean operators. Mind and Language pp 1–21
Cattell R (1978) On the source of interrogative adverbs. Language 54:61–77
Chemla E (2008) An epistemic step for anti-presuppositions. Journal of Semantics 25(2):141–173
Chierchia G, McConnell-Ginet S (1990) Meaning and Grammar: An Introduction to Semantics. MIT Press
Collins C, Postal P (2014) Classical NEG Raising: An essay on the syntax of negation. Cambridge, MA: MIT Press
Copley B, Mari A (2022) *Forbid* is not *order not*. In: Proceedings of NELS 52
Degen J, Tonhauser J (2022) Are there factive predicates? An empirical investigation. Language 98(3):552–591
Djärv K, Bacovcin HA (2020) Prosodic effects on factive presupposition projection. Journal of Pragmatics 169:61–85
Dowty D (1979) Word Meaning and Montague Grammar. Reidel, Dordrecht
Egré P (2008) Question-embedding and factivity. In: Lihoreau F (ed) Grazer Philosophische Studien 77. p 85–125
Enguehard E, Spector B (2021) Explaining gaps in the logical lexicon of natural languages: A decision-theoretic perspective on the square of Aristotle. Semantics and Pragmatics 14
Frege G (1892) Über Sinn und Bedeutung (On Sense and Reference). Zeitschrift für Philosophie und philosophische Kritik 100:25–50
Glass L (2022) The negatively biased Mandarin belief verb *yǐwéi*. Studia Linguistica 77(1):1–46

Glass L (2023) Using the Anna Karenina Principle to explain why *cause* favors negative-sentiment complements. Semantics & Pragmatics 16(6). https://doi.org/10.3765/sp.16.6
Glass L (2025) Where Common Ground and truth come apart: Explaining the factive/contrafactive asymmetry, ms.
Goodman J, Salow B (2023) Epistemology normalized. Philosophical Review 132(1)
Grimshaw J (2015) The light verbs *Say* and say. In: Toivonen I, Csúri P, van der Zee E (eds) Structures in the mind: Essays on Language, Music, and Cognition. MIT Press
Gärdenfors P (2014) The Geometry of Meaning: Semantics Based on Conceptual Spaces. MIT Press, https://doi.org/10.7551/mitpress/9629.001.0001
Hacquard V (2006) Aspects of modality. PhD thesis, Massachusetts Institute of Technology
Hacquard V, Lidz J (2022) On the acquisition of attitude verbs. Annual Review of Linguistics 8(1):193–212. https://doi.org/10.1146/annurev-linguistics-032521-053009
Hoeksema J (1999) Iemand iets wijsmaken: irrealis en negatieve polariteit. TABU: Bulletin voor Taalwetenschap Groningen 1:37–42
Hoeksema J (2021) Verbs of deception, point of view and polarity. In: Proceedings of the 28th International Conference on Head-Driven Phrase Structure Grammar, pp 26–46, https://doi.org/10.21248/hpsg.2021.2
Holton R (2017) Facts, factives, and contrafactives. Aristotelian Society Supplementary Volume 91(1):245–266
Horn LR (1972) On the Semantic Properties of Logical Operators in English. PhD thesis, UCLA
Horn LR (1989) A natural history of negation. CSLI
Hsiao PYK (2017) On counterfactual attitudes: a case study of Taiwanese Southern Min. Lingua Sinica 3(4)
Ichikawa J, Steup M (2017) The Analysis of Knowledge. *Stanford Encyclopedia of Philosophy*: http://plato.stanford.edu/entries/knowledge-analysis
Imel N, Steinert-Threlkeld S (2022) Modal semantic universals optimize the simplicity/informativeness trade-off. In: Starr JR, Kim J, Öney B (eds) Proceedings of SALT 32, pp 227–248, https://doi.org/10.3765/salt.v1i0.5346
Kadmon N (2001) Formal Pragmatics. Malden, MA: Blackwell
Karttunen L, Peters S (1979) Conventional implicature. In: Oh CK, Dineen D (eds) Syntax and Semantics 11: Presupposition. New York: Academic Press, p 1–56
Katzir R, Singh R (2013) Constraints on the lexicalization of logical operators. Linguistics and Philosophy 36. https://doi.org/10.1007/s10988-013-9130-8
Keenan EL, Stavi J (1986) A semantic characterization of natural language determiners. Linguistics and philosophy 9:253–326
Kierstead G (2013) Shifted indexicals and conventional implicature: Tagalog *akala* 'falsely believe'. Slides from SALT 23
Kirton J, Charlie B (1996) Further Aspects of the Grammar of Yanyuwa, Northern Australia. Canberra: Pacific Linguistics
Klein EH (1975) Two sorts of factive predicate. Pragmatics Microfiche 1.1:B5–C14
Lohmann H, Tomasello M (2003) The role of language in the development of false belief understanding: A training study. Child Development 74(4):1130–1144. https://doi.org/10.1111/1467-8624.00597
Major T, Stockwell R (2021) "say"-ing without a Voice. In: Proceedings of NELS 51

Maldonado M, Percus O (2024) Another look at contrafactive predicates: the case of Spanish *creerse*. Talk presented at Sinn und Bedeutung 29, September 17-19, 2024
McGregor W (2023) On the expression of mistaken beliefs in Australian languages. Linguistic Typology
McHugh D (2023) Causation and Modality. PhD thesis, ILLC, University of Amsterdam
Nadathur P (2023) Causal semantics for implicative verbs. Journal of Semantics https://doi.org/doi.org/10.1093/jos/ffad009
Nadathur P, Lauer S (2020) Causal necessity, causal sufficiency, and the implications of causative verbs. Glossa: a journal of general linguistics 5(1):1–37
Pearl J (2000) Causality: Models, Reasoning, and Inference. Cambridge University Press
Percus O (2006) Antipresuppositions. In: Ueyama A (ed) Theoretical and empirical studies of reference and anaphora: Toward the establishment of generative grammar as an empirical science. Japan Society for the promotion of science
Phillips J, Norby A (2021) Factive theory of mind. Mind & Language 36:3–26. https://doi.org/10.1111/mila.12267
Potts C (2005) The Logic of Conventional Implicature. Oxford: Oxford University Press
von Prince K (2015) A Grammar of Daakaka. De Gruyter Mouton
Radvansky GA, Zacks JM (2014) Event Cognition. Oxford University Press
Ramchand G (2014) Stativity and present tense epistemics. In: Proceedings of SALT 24, pp 102–121
Roberts C, Simons M (2024) Preconditions and projection: Explaining non-anaphoric presupposition. Linguistics and Philosophy (47):703–748. https://doi.org/10.1007/s10988-024-09413-9
Rosenberg MS (1975) Counterfactives: A pragmatic analysis of presupposition. PhD thesis, University of Illinois, Urbana-Champaign
Sander T (2023) The K's evil twin. Unpublished manuscript, Uni Duisburg
Sander T (2025) A puzzle about anti-factives. Journal of the American Philosophical Association pp 1–20. https://doi.org/10.1017/apa.2024.24
Sauerland U (2007) Implicated presuppositions. In: Steube A (ed) Sentence and Context. Mounton de Gruyter, Berlin, Germany
Sbardolini G (2023) The logic of lexical connectives. Journal of Philosophical Logic
Schlenker P (2021) Triggering Presuppositions. Glossa: a journal of general linguistics 6(1)
Schulz P (2002) The interaction of lexical-semantics, syntax, and discourse in the acquisition of factivity. In: Skarabela B, Fish S, Do AHJ (eds) BUCLD 26 Proceedings, pp 584–595
Silitonga M (1973) Some rules reordering constituents and their constraints in Batak. PhD thesis, University of Illinois Urbana-Champaign
Simons M, Tonhauser J, Beaver D, et al (2010) What projects and why. In: Li N, Lutz D (eds) Proceedings of SALT 20. CLC., pp 309–327
Simons M, Beaver D, Roberts C, et al (2016) The best question: Explaining the projection behavior of factives. Discourse Processes 54(3):187–206. https://doi.org/10.1080/0163853X.2016.1150660
Stalnaker R (1970) Pragmatics. Synthese 22(1/2):272–289. https://doi.org/10.1007/BF00413603
Stalnaker R (1973) Presuppositions. Journal of Philosophical Logic 2(4):447–457. https://doi.org/10.1007/BF00262951

Stalnaker R (1974) Pragmatic presuppositions. In: Munitz MK, Unger P (eds) Semantics and Philosophy. New York: NYU Press, p 197–213
Steinert-Threlkeld S, Imel N, Guo Q (2023) A semantic universal for modality. Semantics and Pragmatics 16
Strohmaier D, Wimmer S (2022) Contrafactives and learnability. In: Degano M, Roberts T, Sbardolini G, et al (eds) Proceedings of the Amsterdam Colloquium 2022, pp 298–305
Tonhauser J, Beaver D, Degen J (2018) How projective is projective content? Gradience in projectivity and at-issueness. Journal of Semantics 35(3):495–542. https://doi.org/10.1093/jos/ffy007
Vendler Z (1967) Linguistics in Philosophy. Cornell University Press. Ithaca, New York
de Villiers JG, Pyers JE (2002) Complements to cognition: a longitudinal study of the relationship between complex syntax and false-belief-understanding. Cognitive Development 17(1):1037–1060. https://doi.org/10.1016/S0885-2014(02)00073-4
White AS (2014) Factive-implicatives and modalized complements. In: Iyer J, Kusmer L (eds) Proceedings of the 44th Annual Meeting of the North East Linguistic Society, pp 267–278
White AS, Rawlins K (2018) The role of veridicality and factivity in clause selection. In: Hucklebridge S, Nelson M (eds) Proceedings of the 58th Annual Meeting of the North East Linguistic Society. Amherst, MA: GLSA Publications, p 221–234
Williamson T (2000) Knowledge and its Limits. Oxford University Press
Yalcin S (2016) Modalities of normality. In: Charlow N, Chrisman M (eds) Deontic Modality. Oxford University Press, p 230–255