Semantics & Pragmatics Volume 16, Article 6: 1–49, 2023
https://doi.org/10.3765/sp.16.6

Using the Anna Karenina Principle to explain why *cause* favors negative-sentiment complements*

Lelia Glass
Georgia Institute of Technology

Submitted 2023-03-23 / First decision 2023-06-18 / Revision received 2023-07-31 / Second decision 2023-09-10 / Revision received 2023-09-29 / Accepted 2023-10-06 / Final version received 2023-10-18 / Published 2023-10-18 / Final typesetting 2024-03-11

Abstract This paper sets out to explain why the verb *cause* tends to occur with negative-sentiment complements (*cause damage, cause problems*), as observed by Stubbs (1995). Formalized using causal models (Pearl 2000, Halpern & Pearl 2005, Schulz 2011), the analysis hinges on the asymmetric inference patterns licensed by necessary versus sufficient causes in the common scenario where some variables in a causal model remain uncertain. States of certainty/uncertainty are captured by subdividing the traditional definitions of necessity and sufficiency into a local version (all other variables fixed at particular values) and a global version (all other variables unsettled). *C causes E* is argued to entail that *C* is locally sufficient for *E*, and to implicate that *C* is at least possibly locally necessary for *E*. With this definition, it is shown that *C causes E* can be truthfully applied to more uncertain contexts when *C* is a globally sufficient cause of *E* rather than a globally necessary one. *Cause* thus tends to occur with outcomes depending on a single globally sufficient cause — outcomes which are moreover shown to be negative in sentiment, reflecting the independently motivated "Anna Karenina Principle" that bad outcomes tend to require single sufficient causes, thus indirectly explaining why *cause* prefers negative-sentiment complements. The meaning and collocational sentiment of *cause* are used to illuminate one another.

* For illuminating conversations, I thank Rebekah Baglini, Elitzur Bar-Asher Siegal, Cleo Condoravdi, Bridget Copley, Christopher Potts, and Prerna Nadathur. This work benefitted from the insights of audiences at the COCOA (Converging on Causal Ontology Analyses) workshop organized by Bridget Copley; the University of Texas at Austin; the Construction of Meaning Workshop at Stanford; and Sinn und Bedeutung 28. I am indebted to three anonymous reviewers at *Semantics and Pragmatics* (as well as two more at *Journal of Semantics*) for constructive comments, and to the Editor, Louise McNally, for detailed and thoughtful feedback. I gratefully acknowledge the support of the National Science Foundation Collaborative Research Grant BCS-2040820 ("Computational modeling of the internal structure of events") awarded to Aaron Steven White, Scott Grimm, and me. Errors are mine.

Keywords: causation, causal models, sentiment, corpus linguistics, necessity, sufficiency, subjectivity, Anna Karenina Principle

1 Introduction

The verb *cause* — debated for centuries[1] — might seem in theory to have a purely logical and emotionally neutral meaning. Relatively frequent and difficult to paraphrase, *cause* straddles the border between a content word and a function word, with no obvious affective dimension. But in corpus usage, *cause* shows a striking, unexplained tendency to combine with emotionally negative complements such as *damage* and *problems* (Stubbs 1995, Stefanowitsch & Gries 2003, Xiao & McEnery 2006, Childers 2016, Hauser & Schwarz 2018). Aiming to use formal and emotional elements of meaning to illuminate one another, this paper explains the negative collocation of *cause* by leveraging the asymmetric inference patterns licensed by necessary versus sufficient causes, combined with the insight (the Anna Karenina Principle of Diamond 1997) that good outcomes involve many individually necessary-but-insufficient factors, such that the absence of any one of them suffices for a bad outcome.

First, the paper confirms the negative-sentiment corpus distribution of *cause* (Section 2). Next (Section 3), I use causal models (Pearl 2000, Halpern & Pearl 2005, Sloman, Barbey & Hotaling 2009, Schulz 2011, Baglini & Francez 2016, Nadathur 2016, Nadathur & Lauer 2020, Hitchcock 2020) to cross-cut the traditional definitions of necessity and sufficiency into a logically stronger *global* version (other variables unsettled) and a logically weaker *local* version (other variables fixed at particular values). When people make causal claims, it is often not clear what causal model they are entertaining; these definitions allow us to reason not just about fully determined models, but also about those involving uncertainty. In this framework, *C causes E* is argued to entail that *C* is locally sufficient for *E*, and to implicate that *C* is at least possibly locally necessary for *E*.

In situations where the values of some causally relevant variables are uncertain, this state of affairs is automatically satisfied if *C* is a globally sufficient cause of *E* (global sufficiency entails local sufficiency), but is not determined if *C* is a globally necessary cause of *E* (Section 4). As a result, *C*

1 See, among others, the historical references Hume 1748, Mill 1843, Mackie 1965, Lewis 1973, Dowty 1979, and from this century, Hobbs 2005, Wolff 2007, Sloman, Barbey & Hotaling 2009, Schulz 2011, Neeleman & Van de Koot 2012, Copley & Wolff 2014, Nadathur & Lauer 2020, Hitchcock 2020.

causes E is true in a wider variety of uncertain situations when *C* is a globally sufficient cause of *E* versus a globally necessary one.

The final puzzle piece links the necessity/sufficiency asymmetry back to sentiment (Section 5). The Anna Karenina Principle of Diamond (1997) — encapsulated in Tolstoy's novel by that name — states that, in general, many different factors are individually necessary-but-insufficient for success, while the absence of any one of them is sufficient for failure. Using experimental data, the Anna Karenina Principle is derived from the deeper insight that causal models are subjective. When a person desires a given outcome, they tend to assign a model with multiple individually necessary-but-insufficient causes; when they wish to avoid that outcome, they tend to assign a model requiring a single sufficient cause. I argue that the Anna Karenina Principle arises because people assign different causal models to the outcomes that they view as desirable versus undesirable.

Recall that *C causes E* is true in a wider variety of uncertain situations when *C* is a globally sufficient cause of *E* — which the Anna Karenina Principle links to bad outcomes. *C causes E* is true in a more limited class of situations when *C* is a globally necessary cause of *E*, which this principle links to good outcomes. Thus, *C causes E* is more often true of bad outcomes than good outcomes, explaining why the complements of *cause* tend to be negative (Section 6). This explanation is defended (Section 7) over an alternative whereby *cause* is used to blame the violators of norms (Hart & Honoré 1959, Hilton & Slugoski 1986, Hitchcock & Knobe 2009, Alicke, Rose & Bloom 2011). The same analysis can be extended to *because*, also shown (Section 8) to involve negative collocational sentiment.

Stepping back (Section 9), this exploration illuminates how the causal models constructed for a given situation are subjective and uncertain; how necessary versus sufficient causes give rise to asymmetric inferences under uncertainty; and how affective dimensions of logical/functional language can be derived from its semantic core (Potts 2011, Acton & Potts 2014, Beltrama 2016, Acton 2019).

2 *Cause* occurs with negative-sentiment complements

As first observed by Stubbs (1995), the verb *cause* occurs mainly in unpleasant collocations, such as *cause damage/problems/confusion*, with the effect that when an event is described using *cause* (versus other verbs such as *make* or *produce*), people choose negative-sentiment adjectives to complete the sentence in a fill-in-the-blank task (Childers 2016) and evaluate the event

negatively (Hauser & Schwarz 2018). Here, "negative" refers to the emotional valence typically assigned to a word and/or its referent, drawing on the insight (Osgood, Suci & Tannenbaum 1957) that content words carry emotional as well as denotational meaning; *poor, death,* and *destroy* are negative in sentiment, while *fun, holiday,* and *celebrate* are positive; *geographic, car,* and *pull* are relatively neutral.

2.1 Replicating the negative-sentiment distribution of *cause*

This finding is further replicated using tools built by the natural language processing community (popularized by Pang & Lee 2008) to automatically annotate textual sentiment. The data were drawn from Reddit, a public United States-based web discussion platform whose contents were (up until 2023)[2] made available to researchers by Baumgartner et al. (2020). I used ten million words of comments from January 2018 in the AskReddit forum (a large forum dedicated to general-interest topics); I excluded repeated comments, those written by self-identified bots, and those containing non-ASCII characters.

A dependency parser (Honnibal & Johnson 2015) was used to extract all clauses with a verb as their root.[3] For each clause, each lemmatized noun subject (if any), each lemmatized direct object (if any) or sentential complement clause (if any) of every verb, sentiment was labeled using the Hedonometer of Dodds et al. (2011, 2015), which aggregates human Likert ratings for the emotional valence of words (chosen over other sentiment tools because it covers more word types).[4]

2 After the famous ChatGPT language model was trained in part on Reddit data with no compensation to Reddit, Reddit made its data harder to access. I still have access to the January 2018 portion of Baumgartner et al. (2020).

3 All data and code are available through the Open Science Framework at the link https://osf.io/mv5gk/.

4 To calculate the sentiment of a multi-word clause, I gathered the average sentiment of all its component words, excluding those with middling sentiment ratings between 4 and 6 on a 1–9 scale (Dodds et al. 2011: p. 5). Dodds et al. (2011) use the term "stop words" — normally reserved for highly frequent function words such as *the* and *of*, which are sometimes excluded in certain language processing applications — for words with a middling rating between 4 and 6. This terminology is somewhat confusing, because those middling-sentiment words include content words such as *bottle* which are not normally considered "stop words." In any case, Dodds et al. (2011) suggest that such words should be excluded from the calculation of the mean sentiment of a string. While Dodds et al. (2011) weight a word's sentiment by its inverse frequency, I simply take the unweighted average after excluding those with middling ratings.

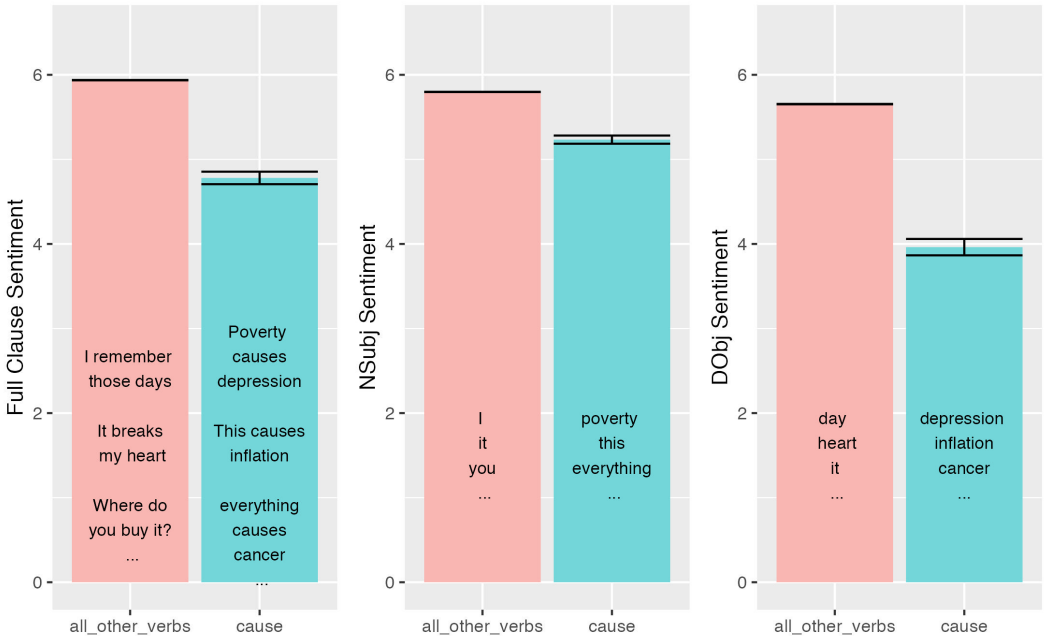


Figure 1 Mean Hedonometer sentiment for all clauses in Reddit data containing a verb; all lemmatized subject nouns (NSubjects); and all lemmatized direct object nouns (DObjects), for all other verbs versus the verb *cause*. Error bars represent the standard error of the mean.

The Hedonometer gives *cause* (not part-of-speech-tagged) a middling rating of 5.22 on a 1–9 scale, which falls close to the average across all verbs of 5.38, and within the range that Dodds et al. (2011) see as sentiment-neutral. But comparing *cause* to all other verbs, I find that the verb *cause* (which serves as the root of 512 unique clauses) occurs in more-negative clauses (mean = 4.78 versus 5.94), with somewhat more-negative lemmatized subject nouns (mean = 5.23 versus 5.80), and especially with much more negative lemmatized direct objects (mean = 3.96 versus 5.66).

(1) ***Cause***

a. Poverty$_{\text{NSUBJ}}$ **causes**$_{\text{VERB}}$ depression$_{\text{DOBJ}}$.
b. This$_{\text{NSUBJ}}$ **causes**$_{\text{VERB}}$ inflation$_{\text{DOBJ}}$.
c. Everything$_{\text{NSUBJ}}$ **causes**$_{\text{VERB}}$ cancer$_{\text{DOBJ}}$.

(2) **All other verbs**

a. I_{NSUBJ} **remember**$_{\text{VERB}}$ those days$_{\text{DOBJ}}$.
b. It_{NSUBJ} **breaks**$_{\text{VERB}}$ my heart$_{\text{DOBJ}}$.
c. Where do you$_{\text{NSUBJ}}$ **buy**$_{\text{VERB}}$ it$_{\text{DOBJ}}$?

These effects are all highly significant ($p < 0.001$) in unpaired, two-sided t tests. Figure 1 shows the mean Hedonometer rating for full clauses, noun subjects, and direct objects of all other verbs compared to *cause*.

The Hedonometer (Dodds et al. 2011) uses a nine-point scale, with a midpoint of five. For all clauses containing verbs other than *cause*, 18% of them have a Hedonometer rating less than five, compared to 53% of clauses containing *cause*. For all direct object tokens of verbs other than *cause*, 16% of them have a Hedonometer rating less than five, compared to 65% of direct object tokens of *cause*. In other words, at least half of *cause*'s tokens are associated with negative sentiment.

Most of the direct object nouns of *cause* — *depression, inflation,* and so on (1) — denote events or states according to the WordNet ontology (Miller et al. 1990), but (in an unpaired t test) are on average significantly more negative in sentiment (mean = 2.96) even compared to other eventuality-denoting direct objects (mean = 5.52).

As for clausal complements (3), those of *cause* are also more negative in sentiment (mean = 5.33) than those of other verbs (mean = 5.91), again significant by the t test. These data are less reliable, though, because the dependency parser often mistakes the discourse connective *(be)cause* (discussed below in Section 8) for a verb with a clausal complement (*cause there's no other option.*). I filtered the 71 instances of *cause* with a clausal complement by hand to remove such cases, finding only 25 true instances (full clauses such as (3) as well as ditransitive cases such as *caused me harm*).[5] These 25 true clausal complements of *cause* are also more negative in sentiment (mean = 5.04) than those of other verbs (mean = 5.91).

(3) a. ***Cause*: causing** [less and less jobs to be available]$_{\text{CCOMP}}$.
b. **Other clause-embedding verbs:** I **thought** [it was cool back then]$_{\text{CCOMP}}$.

Echoing Stubbs (1995), these results show that *cause* is distributionally associated with negative sentiment.

5 Stefanowitsch & Gries (2003) find that *cause* has negative-sentiment direct objects in its ditransitive form as well as its transitive one (*cause you inconvenience*).

2.2 Towards an analysis

For content words such as *poor* or *destroy*, negative sentiment is clearly based in their meaning. But the meaning of *cause* might seem in principle to be purely logical, and the word itself is annotated as neutral. Thus, while it is clear that *cause* occurs in negative-sentiment contexts, it is much less clear why.

Of course, it would be circular to argue that *cause* takes on negative sentiment because it appears in negative-sentiment contexts, or vice versa (Stubbs 1995). Perhaps newsworthy events tend to be negative (Stubbs 1995) — but that would not explain why *cause* patterns more negatively than any other word in the corpus. Perhaps humans prefer stasis, so that caused changes tend to be negative (Louw & Chateau 2010). Perhaps *cause* describes events where an agent overpowers the will of a patient, which the patient would view negatively (Childers 2016). Perhaps, as discussed in Section 7, *cause* is used (for some reason) to profile norm violations and assign blame. Perhaps *cause* lexicalizes negative sentiment in some way — for Childers (2016), via a conventional implicature in the sense of Grice (1989) and Potts (2004) — but this meaning would have to be either stipulated or somehow explained, and one would have to further explain why *cause* can also appear in positive-sentiment contexts without contradiction, as in the web-attested (4):

(4) What **causes** happiness?[6]

As (4) illustrates, the negative collocation of *cause* is a tendency rather than an absolute, and thus cannot be explained absolutely. To reconcile (4) with the claim that *cause* lexicalizes a negative-sentiment conventional implicature, Childers (2016) proposes that *cause* polysemously encodes two meanings: $cause_1$, which is unmarked with respect to formality and conventionally implicates negative sentiment from the speaker, as in (1); and $cause_2$ which is restricted to formal/academic registers and is unmarked with respect to sentiment, as in (4). Here, Childers (2016) echoes Hunston (2007), who suggests that *cause* is neutral in scientific writing when it describes events that do not directly involve humans, but this claim conflicts with the finding from Louw & Chateau (2010) that *cause* favors negative complements even in the

6 From an article in the *Greater Good Magazine* based at the University of California, Berkeley by Kira Newman, July 28, 2015: https://greatergood.berkeley.edu/article/item/six_ways_happiness_is_good_for_your_health

Academic genre of the Corpus of Contemporary American English (Davies 2008-).

In this paper, in contrast, I pursue an analysis whereby *cause* has a unified meaning across registers, consistent with both negative and positive uses (1)–(4). Rather than positing negative sentiment in the lexical meaning of *cause*, I propose to derive it from a sentiment-neutral core semantics, which interacts with independent facts about the types of situations that *cause* describes.

This approach takes inspiration from Potts (2011), who observes that negation (*not*) would seem *a priori* to have a neutral, logical meaning, and yet disproportionately occurs in negative-sentiment movie reviews. Potts proposes that negation comes to be associated with negative sentiment because it is used in dispreferred discourse moves such as disagreeing, rejecting, and uninformative statements. Inspiring a larger exploration of the affective meanings of function words (Potts 2011, Acton & Potts 2014, Beltrama 2016, Acton 2019), the negative sentiment of negation is built pragmatically atop a sentiment-neutral semantic core. This paper applies the same framework to *cause*.

3 Definitions: Local versus global necessity versus sufficiency; *cause* itself

The analysis begins with a sentiment-neutral basic meaning for *cause*. Causal relations can be systematized using causal models (Pearl 2000, Halpern & Pearl 2005, Sloman, Barbey & Hotaling 2009, Schulz 2011, Baglini & Francez 2016, Nadathur & Lauer 2020, Hitchcock 2020), functions determining how a variable depends on those represented as causally upstream in a directed acyclic graph (a causal structure). Causal models do not reduce causation to any deeper primitive, but explicate causal relations so that they can be studied formally.

Causal models are a powerful framework which can represent elaborate structures: variables might take on continuous values or probabilities, might mitigate or moderate one another, or might trigger a cascading chain of effects. But this paper focuses on a simple, deterministic model, inspired by Sloman, Barbey & Hotaling (2009), with three binary variables: a light which is (always and only) on when its two switches are both on (Figure 3), and off if any of its switches is off.

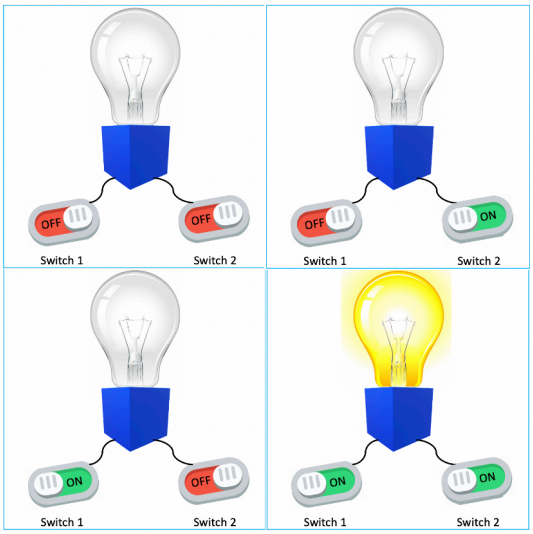


Figure 2 The light is (always, only) on when both switches are on.

3.1 Necessity versus sufficiency

In this example, the two switches represent exogenous variables which get their values (on or off) by stipulation; the light represents an endogenous variable whose value is determined by the switches. With two switches, the light is just complex enough to distinguish necessity versus sufficiency (5)–(6) (these concepts date at least to Hume 1748; my formulation takes inspiration from Pearl 1999). For the light to be on, it is necessary that Switch 1 is on (counterfactually, if Switch 1 were not on, the light would not be on either). For the light to be off, it is sufficient that Switch 1 is off.

(5) *C* is **necessary** for *E* iff *E* guarantees *C*.
If not *C*, then not *E*.

(6) *C* is **sufficient** for *E* iff *C* guarantees *E*.
If *C*, then *E*.

In this example, the light being on represents a conjunctive, "multiple-necessary" scenario (Kelley 1973, Kun & Weiner 1973, Halpern & Pearl 2005): the two switches must both be on for the light to be on, so each one is individually necessary-but-insufficient for that outcome. Together, the two switches being on represent a "sufficient set" (Mackie 1965, Wright 1985, 2013, Baglini & Bar-Asher Siegal 2020) for the light to be on — a conjunction

of multiple factors that are jointly sufficient for the result. Conversely, the light being off represents a disjunctive, "single-sufficient" scenario: any one switch being off is enough for the light to be off, so each one is individually sufficient-but-unnecessary. Of course, the conjunctive and disjunctive scenarios are mirror images in general (Kun & Weiner 1973, von Wright 1974, Bar-Asher Siegal & Boneh 2019): if two conditions are both necessary for a result, then the absence of either one is sufficient for the absence of that result.

Here, the arguments *C* and *E* of *cause* are technically propositions, because they denote truth values: it can be true or false that the switch or the light is on/off. On the other hand, we normally think of causation as a relation between events (or, to be more precise, eventualities — events and states in the terminology of Bach 1986): one event causes another. To unify the boolean character of propositions with the intuition that causation involves events, we follow Lewis (1973) in taking the arguments of the verb *cause* as propositions that certain events occur. The syntactic form does not matter; its arguments can be event-denoting noun phrases; individual-denoting noun phrases understood metonymically to refer to that individual's actions; clauses; and so on — all that matters is that *C* and *E* can be understood the propositions that certain eventualities are instantiated.

While the definitions of necessity and sufficiency are straightforward, it is much less clear how they play a role in the lexical semantics of words such as *cause*. Some researchers (Lewis 1973, Hobbs 2005, Neeleman & Van de Koot 2012, Nadathur & Lauer 2020) have defined *cause* primarily in terns of necessity, others in terms of sufficiency (used by Ikuta et al. 2014 for *cause*, and by Nadathur & Lauer 2020 for the periphrastic causative *make*), and others (Mackie 1965, Halpern & Pearl 2005, Wright 1985, 2013, Baglini & Bar-Asher Siegal 2020, Beller, Bennett & Gerstenberg 2020) using more elaborate combinations of the two.[7] In prior literature, the roles of necessity and sufficiency in the meaning of *cause* remain unresolved.

3.2 Causal models are subjective and uncertain

Contributing to the debate about the meaning of *cause*, it is often not clear what causal models are entertained by speakers or hearers in a conversation

7 Another line of work (Talmy 1988, Wolff 2007) does not use logical relations such as (5)–(6) but instead defines causation by analogy to the physical world using concepts such as force and energy.

where *cause* is used (Hobbs 2005, Halpern & Pearl 2005, Hitchcock 2020, Menzies & Beebee 2020). The lightbulb in Figure 3 represents the rare case where the model is fully explicit. In contrast, in a conversation in which (7) is used, it is not clear whether the speaker is imagining a model with only one upstream variable (the circuit), or one with many upstream variables including the electricity being connected, the presence of oxygen, and the house being made of dry wood.

(7) The short circuit caused the house fire. (adapted from Mackie 1965)

For any real-world situation, there is no single correct causal model and no objective way to decide which contributing factors to include or leave out. In creating a model of (7), Halpern & Pearl (2005) explain, we could represent the house's dry wood as an exogenous variable (stipulated to be true), as an endogenous variable (dependent on some other upstream factors), or could leave it out entirely as an unstated background fact. Such choices are subjective; "It is not always straightforward to decide what the 'right' causal model is in a given situation, nor is it always obvious which of two causal models is 'better' in some sense" (Halpern & Pearl 2005: Section 2). It depends whether the dry wood is taken for granted or in question, which depends on the situation as well as one's view of it.

Along the same lines, Hobbs (2005) argues that we focus on a subset of relevant factors in causal reasoning, typically those that can be altered by human action, while backgrounding others as "presumable;" the choice of which factors are considered changeable or presumable is "dependent upon […] the situation or context." Kun & Weiner (1973) describe scenarios and then elicit the structure of the causal model imagined by each experimental participant, leveraging the assumption that a person's model is not fully determined by the scenario, nor necessarily shared by others. Empirically, Glymour & Wimberly (2007: Section 8) lament that causal intuitions "may vary considerably from person to person" in experiments, and Bethard et al. (2008) abandon an attempt to elicit corpus annotations of necessary versus sufficient causes due to low inter-annotator agreement. Exploring judgments of causal responsibility, Lagnado, Gerstenberg & Zultan (2014: p. 1065) observe that "even when presented with an identical scenario[,] people might construct different models and hence legitimately differ in their responsibility judgments."

Just as it is a subjective and uncertain task to construct a model for a given situation, it is also subjective and uncertain to choose one of many

contributing factors as "the" cause of a result — a question known as "causal selection" (Hart & Honoré 1959, Lewis 1973, Cheng & Novick 1991, Hobbs 2005, Neeleman & Van de Koot 2012, Baglini & Bar-Asher Siegal 2020, Bar-Asher Siegal, Bassel & Hagmayer 2021). For example, how is a single cause (the circuit) chosen as the syntactic subject of *cause* among many contributing factors (the circuit, electricity, oxygen, dry wood) which might or might not be represented in a model of the situation? This question is attributed to Mill (1843), who writes:

> "Causation is seldom if ever between a consequent and a single antecedent …but usually between a consequent and the sum of several antecedents, the occurrence of all of them being requisite to produce …the consequent. In such cases it is very common to single out only one of the antecedents under the denomination of CAUSE, calling the others merely Conditions." (*A System of Logic*, Chapter 5, Section 3; cited by Baglini & Bar-Asher Siegal 2020)

In other words, just as interlocutors might be uncertain about the structure of the model, they might also be uncertain about whether various variables in the model are in question or fixed at one value or another as background conditions. Both types of uncertainty arise because one's causal model depends on subjective judgment. When the Anna Karenina Principle is discussed later on (Section 5), we expand the idea that causal models are subjective; here, we focus on the result that it can be difficult to recover the model imagined by a person who makes a causal claim.

3.3 Local versus global necessity versus sufficiency

This paper attempts to clarify the uncertainty of causal models both visually and conceptually. On a visual level, the lightbulb illustrations explicate the model under discussion. As for which variables are fixed or changeable, the illustrations (Table 1) use a blue lock icon to indicate variables fixed at particular values, and a gray box to hide those whose values are undetermined.

On a conceptual level, the paper proposes to subdivide the traditional definitions of necessity and sufficiency (5)–(6) into a global version (all other variables left open) and a local version (all other variables fixed at specific values) — using terminology from econometrics (Forni & Gambetti 2014) and

ideas drawn from Mackie (1965), Halpern & Pearl (2005), Baglini & Francez (2016), Martin (2018), Nadathur & Lauer (2020), and Baglini & Bar-Asher Siegal (2020). The idea of "fixed" variables does not imply temporal order (though see Martin 2018, Baglini & Bar-Asher Siegal 2020 for discussion of temporality); it just means that these variables are held constant where others may be unknown or changeable. These definitions help us understand how we reason about causal models in situations where different variables are uncertain versus fixed.[8]

As shown in the top row of Table 1, a variable *C* is locally necessary for an outcome *E* if it is necessary for *E* given a particular fixed setting of other relevant variables. Here, S1=OFF (the state where Switch 1 is off) is not globally necessary for L=OFF (we can imagine situations where L=OFF even when S1=ON, namely when S2=OFF), but is locally necessary for L=OFF if we take S2=ON as fixed. Similarly, *C* is locally sufficient for *E* if it is sufficient given a particular fixed setting of other variables (echoing the idea of a "sufficient set" from Mackie 1965, Wright 1985, 2013, Baglini & Bar-Asher Siegal 2020 and the idea from Nadathur & Lauer 2020 of sufficiency with respect to a background situation in which other variables are fixed at the correct values). S1=ON is not globally sufficient for L=ON (if S2=OFF, the light would still be off even if S1=ON), but is locally sufficient for L=ON if we take S2=ON as fixed. These local definitions represent situations where interlocutors know that other variables are fixed at specified values.

Turning to the second row of Table 1, a variable *C* is globally necessary for an outcome *E* if it is necessary for *E* regardless of what happens to any other variable: S1=ON is globally necessary for L=ON because, no matter what happens to any other variable, the light is only on if S1 is on. Similarly, *C* is globally sufficient for *E* if it is sufficient for *E* regardless of what happens to any other variable: S1=OFF is globally sufficient for L=OFF because, no matter what happens to any other variable, the light is off if S1 is off. These global definitions represent situations where the relation between *C* and *E* is logically strong enough that it holds even when interlocutors may be uncertain about the values of other variables in the model, or uncertain about whether other variables are included in the model or not.

8 A note on terminology: The distinction between local and global necessity/sufficiency is independent of the distinction between "type" and "token" causation (generic causal statements versus statements about particular instances). All of this paper's lightbulb illustrations represent particular instances ("token" causation), but some of them represent local necessity/sufficiency, and others represent global versions thereof.

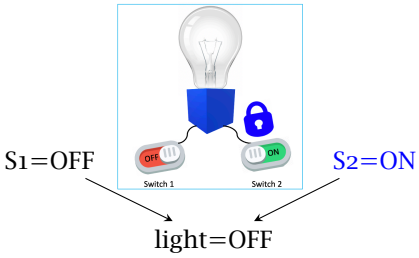
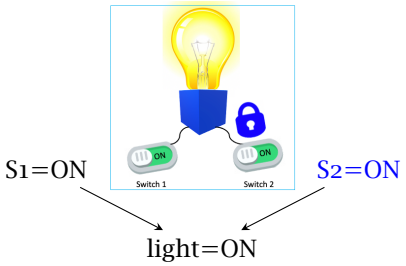
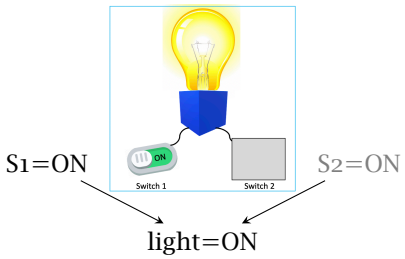
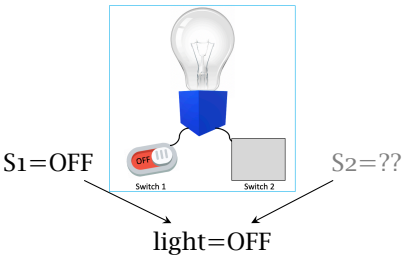
	Necessary	**Sufficient**
Locally	⋆ Given some fixed setting of all other variables, E guarantees C ⋆ Given S2=ON, L=OFF guarantees S1=OFF 	⋆ Given some fixed setting of all other variables, C guarantees E ⋆ Given S2=ON, S1=ON guarantees L=ON 
Globally	⋆ No matter how other variables are set, E guarantees C ⋆ No matter what happens to any other switch, L=ON guarantees S1=ON 	⋆ No matter how other variables are set, C guarantees E ⋆ No matter what happens to any other switch, S1=OFF guarantees L=OFF 

Table 1 Illustrated definitions of local versus global necessity versus sufficiency.

Having sketched these definitions with examples, (8)–(11) make them more explicit. Throughout, these definitions are interpreted relative to some causal model M in which C is causally upstream of E, meaning that there is at least *some* possible setting of other variables in the model M such that toggling the truth of C changes the value of E: to use the terminology of Hobbs (2005), C must be "change-relevant" for E. As a general principle of constructing useful causal models (Hobbs 2005), I assume that every variable must be change-relevant for all its downstream variables: if the sun is going to come up whether or not I set my alarm, then we cannot construct a perverse model where my alarm is placed upstream of the sun.

(8) C is **locally necessary** for E iff, given some fixed setting of all other upstream variables in M, E would not happen but for C.

(9) C is **globally necessary** for E iff, no matter what happens to any other upstream variables in M, E would not happen but for C.

(10) C is **locally sufficient** for E iff, given some fixed setting of all other upstream variables in M, C guarantees E.

(11) C is **globally sufficient** for E iff, no matter what happens to any other upstream variables in M, C guarantees E.

3.4 Proposed satisfaction conditions for *C causes E*

These definitions underlie this paper's proposed satisfaction conditions for *C causes E*, where C and E are propositions that certain eventualities occur:

(12) C *causes* E in world w at time t with respect to a causal model M:
   a. Entails that: C and E **both hold** in w at or prior to t.
   b. Entails that: C is **locally sufficient** for E in M.
   c. Implicates that: C is at least possibly **locally necessary** for E in M.

On this definition, C *causes* E entails that C and E both hold in the actual world w, and further that C — perhaps in combination with other fixed variables — guarantees E in the associated causal model M. As for the relation between the actual world w and the causal model M, I assume that for M to be useful, it should align with w on the values of C and E as well as any other upstream variables taken to determine them. To align with w, M must include the fact that C and E both occur, which thus entails that C is

locally sufficient for E given the correct setting of other variables in M that affect E. The two conjuncts (12a) and (12b) ensure that w and M align in the desired manner. Finally, *C causes E* conversationally implicates (Grice 1989, Potts 2015) that it's at least possible, given what's known/fixed in M, that E would not happen but for C (12c).

This definition (12) only makes reference to the local versions of necessity and sufficiency introduced above, where all other variables in a causal model are fixed; but we will see below (Section 4) that its consequences extend to the cases of global necessity and sufficiency, where some variables in a model remain uncertain.

To situate (12) within the literature, it is clear that *C causes E* entails that C and E both hold in the actual world (Mackie 1965, Lewis 1973). Moreover, if C is taken to be causally upstream of E in a model M, then the occurrence of both C and E is actually equivalent to the claim that C is locally sufficient for E in M: if C and E both hold, then other variables in the model must be set to allow E, so that C combined with those other variables jointly guarantee E (Baglini & Bar-Asher Siegal 2020). Thus, the first conjunct (12a) (that C and E both occur in w) is uncontroversial, and the second (12b) (that C is locally sufficient for E in M) is synonymous with the first when we assume that M must align with w on the values of C and E.

The implicature of necessity (12c) takes inspiration from Nadathur & Lauer (2020), who propose that the causative *make* construction *C makes E* entails that C is (locally) sufficient for E and implicates that it is also necessary; but (12c) is formulated to account for situations where some variables in the model may remain unsettled. To derive this implicature conversationally, I suggest that it would be uninformative to say that *C caused E* if E was going to happen regardless.

In other words, apart from using the new definitions of local necessity and sufficiency proposed above, (12) echoes longstanding ideas from the literature. Turning to an example, we may not know exactly what causal model M is entertained by a person who utters (13) (please see Section 3.2), but (12) predicts a causal model where the short circuit — in combination with other fixed variables, such as the presence of oxygen and the house being made of dry wood — sufficed for the fire, and where, without the short circuit, the fire might not have happened.

(13) The short circuit caused the house fire (in w, with respect to a model M).
   a. Entails that: The short circuit and house fire both occur in w.
   b. Entails that: The short circuit is locally sufficient for the fire in M.
   c. Implicates that: Without the short circuit, it's at least possible in M that the fire would not have happened.

On this analysis, it is worth clarifying which elements of a causal claim are subjective versus objective. I argue that in constructing a model, an agent subjectively decides which factors are backgrounded, fixed, or actionable. The speaker's imagined model is not a presupposition of the verb *cause* (though see Bar-Asher Siegal & Boneh 2020 for discussion of causal relations in the presupposition-diagnosing context of linguistic negation), but part of their beliefs about the world.

Thus, two people could reasonably hold different models of the same situation, which might lead to miscommunication or faultless disagreement (Kölbel 2004), for example about whether a fire was caused by a short circuit, faulty electrical wiring, the failure of a negligent landlord to fix the wiring, and so on. On the other hand, when there is a single, explicit, shared causal model (as in this paper's lightbulb illustrations), then a causal claim about such a model can be objectively true or false, depending on whether the model fits the proposed definition (12) for the verb *cause*.

In support of this proposed definition, I argue below (Section 4) that it correctly predicts the situations in which *C causes E* is judged true, and ultimately (Section 6) grounds an explanation of why *cause* favors negative-sentiment complements.

3.5 Challenges to the proposed definition

Before proceeding, it is worth clarifying how this proposed definition of *cause* (12) handles various challenges raised for all analyses thereof.

3.5.1 Episodic versus generic causal statements

The proposed definition (12) follows Lewis (1973: p. 558) in that it is "meant to apply to causation in particular cases" (also known as "token causation") but is "not an analysis of causal generalizations" ("type causation") such as those made by generic sentences (14), often associated with bare nouns and

simple present tense (Carlson 1977, Krifka et al. 1995; please see Leslie & Lerner 2021 for a recent review). Without going too far afield into this rich literature, I suggest that the proposed definition (12) could be combined with an analysis of generic sentences (Krifka et al. 1995, Leslie & Lerner 2021) to handle such cases.

(14) Smoking causes cancer. (McCawley 1976)

As a sketch, (14) might be analyzed to mean that in general, in situations of the relevant type, smoking and cancer both occur, meaning that smoking is (perhaps in combination with background variables such as a person's age and genetic disposition) locally sufficient for cancer, and to implicate that without the smoking, it's at least possible that the cancer would not have occurred.

3.5.2 Causal (in)directness

It is often suggested that *cause* can describe a causal chain whereby C causes an intermediate event C' which in turn causes E (Fodor 1970). In this respect, *cause* is contrasted with causative verbs such as *kill*, which are claimed (Fodor 1970, Shibatani 1976) — debatably (Neeleman & Van de Koot 2012) — to only allow direct causation, without such intermediaries (please see, e.g., Wolff 2003, Martin 2018, Bar-Asher Siegal & Boneh 2019 for more recent discussion).

For example, in the novel *Anna Karenina*, the title character's husband Karenin will not agree to a divorce, meaning that Anna cannot marry her lover nor legitimize their extramarital child. This situation leads Anna to desperation, which leads her to throw herself in front of a train, which leads to a deadly collision. This tragedy can be described by (15) — an indirect chain of causation:[9]

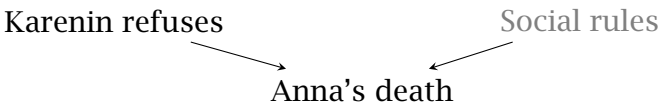
(15) Karenin's refusal of a divorce caused Anna's death.

To handle this statement on the proposed definition (12), one option (16) is to simply construct a model that glosses over intermediate steps (Anna's desperation, her collision with the train, and so on). Other contributing fac-

9 Of course, different readers may be more or less willing to endorse (15), because they may disagree about which factors contributing to Anna's death should be seen as fixed versus changeable — instantiating the "causal selection" question mentioned above.

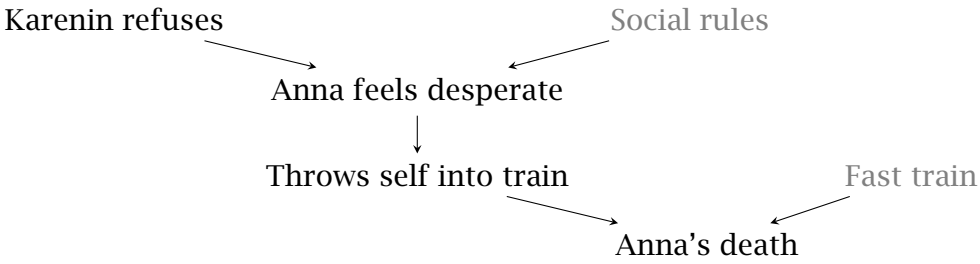
tors (social rules disallowing unilateral divorce and discriminating against extramarital children; the speed of the train; etc) might be represented as factors presumed fixed (in gray), or might be left out entirely. There is no objective fact about what must be included in a causal model, so we are free to construct models where the relation between C and E represents a zoomed-out granularity (Pinker 1989: p. 102), eliding intermediaries. Here, Karenin's refusal is locally sufficient for Anna's death (sufficient in combination with the rules of their society, which we hold fixed here), as well as locally necessary (she would not have died otherwise) — all consistent with (12).

(16)



Another option is to assume that Karenin's refusal is locally sufficient for Anna's death in light of its ensuing consequences, which are taken to unfold deterministically when combined with other factors (the rules of their society; the speed of the train) presumed fixed. Here too, (12) is satisfied.

(17)



In other words, by leveraging the idea that causal models are subjective, we can easily construct models (16)–(17) where the proposed definition (12) holds for indirect causation.

3.5.3 Overdetermination

All definitions of *cause* must famously handle overdetermination and preemption (Lewis 1973): situations where C appears to cause E, but where E was going to happen anyway. In the hallmark example (Lewis 2000), Suzy and Billy both throw rocks at a bottle; Suzy's rock hits first and the bottle breaks; but Billy's rock would no doubt have broken it otherwise. On a necessity-based definition (Lewis 1973) of *cause* requiring that E (the breaking of the bottle) would not happen but for C (Suzy's throw), it is puzzling

that (18a) seems true when the bottle would have broken in any case. On a sufficiency-based definition like the one proposed here, it is perhaps surprising that (18b) seems false — when Billy's throw and the break both occur, and when Billy's throw seems sufficient for breaking the bottle, apparently consistent with (12).

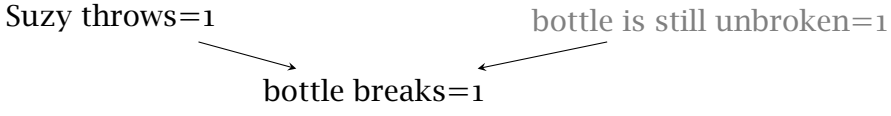
(18) a. Suzy's throw caused the bottle to break.
b. Billy's throw caused the bottle to break.

To handle such issues, researchers have invoked additional notions, such as "how"-causation (Lewis 2000, Beller, Bennett & Gerstenberg 2020) or spatiotemporally-specific causal "production" (Hall 2004) — the idea that Suzy's throw caused the bottle to break in the particular way, at the particular time that it did, following a process set in motion by her throw — to make (18a) true and (18b) false.

Here, in contrast, I suggest that such cases can be handled by (12) as it stands. In brief, as argued by Halpern (2016), the rock-throwing scenario can be represented via two distinct models: one where the bottle is not yet broken, and one where it is.

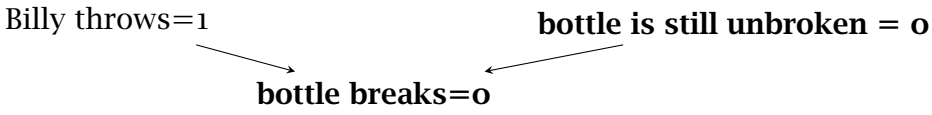
In Suzy's model, Suzy throws her rock, at which point the bottle is unbroken; the bottle breaks as a result. Here, Suzy's throw is locally sufficient (combined with the assumption that the bottle is still whole) for the bottle to break, so (12) correctly makes (18a) true. It's stipulated (though not depicted in this model) that Billy would have broken the bottle otherwise, violating the implicature of (12) that the bottle might not have broken if not for Suzy's throw; but it is not surprising that such a contrived situation may result in slight pragmatic oddity. Crucially, (18a) is still strictly true, which I take as a success for the proposed analysis.

(19)



In Billy's model, Billy throws his rock; but the bottle is already broken, so his rock does not cause the bottle to break. Thus (18b) is correctly predicted to be false.

(20)



Of course, a full account should grapple in more detail with the decades of philosophical research on such cases. But with this brief sketch, I suggest that the proposed definition (12) can handle overdetermination along with other challenges.

Analyzing both (in)directness and overdetermination, I have proposed to leverage the idea that causal models are subjective and uncertain (Section 3.2) to build models that verify our desired intuitions. Such flexibility is not a wily hack, but derives from the deeper insight that there is no single correct model of a given situation.

In any case, the skeptical reader is welcome to favor their own definition of *cause*. To explain why *cause* favors negative-sentiment complements, all we need is an analysis where *C causes E* entails that *C* is locally sufficient for *E* in *M*, which is equivalent to the truism that *C* and *E* both occur in *w*. That is enough to trigger the asymmetric inference patterns licensed by necessity versus sufficiency under uncertainty, explored in the next section, which underlie this paper's proposed analysis.

4 The proposed analysis

Having defined local versus global necessity and sufficiency as well as *cause*, the payoff comes from exploring the inferential consequences of these definitions. Ultimately, it is argued that *C causes E* is true in a wider variety of uncertain situations when *C* is a globally sufficient cause of *E* rather than a globally necessary one, which in turn is used to explain why *cause* favors negative-sentiment complements.

4.1 Logical consequences of necessity versus sufficiency

Logically, the global versions of necessity and sufficiency asymmetrically entail the local versions; in other words:

(21) a. If *C* is globally necessary for *E*, it is certainly locally necessary for *E*.
b. If *C* is locally necessary for *E*, it might or might not be globally necessary for *E*.

(22) a. If *C* is globally sufficient for *E*, it is certainly locally sufficient for *E*.
b. If *C* is locally sufficient for *E*, it might or might not be globally sufficient for *E*.

While the global-to-local entailment (21)–(22) applies equally to necessity and sufficiency, other inferential consequences distinguish them in important ways. Namely, in models where some variables are uncertain, a globally sufficient cause licenses stronger inferences than a globally necessary one. This asymmetry constitutes the key to the proposed analysis of why *cause* favors negative-sentiment complements.

As shown in the right column of Table 2, if C is a globally sufficient cause of E, then — even when other variables are uncertain — knowing C is sufficient for inferring E. Because S1=OFF is globally sufficient for L=OFF, knowing S1=OFF licenses the inference that L=OFF in a model where all variables are specified, as well as in a model where some variables (S2, L) are not given.[10] Even if L=OFF is not explicitly provided (bottom right of Table 2), it can be inferred from the globally sufficient fact that S1=OFF.

In contrast (shown in the left column of Table 2), if C is a globally necessary cause of E, then knowing C does not license an inference that E. S1=ON is globally necessary for L=ON, and there are fully specified models (the top left corner of Table 2) where S1=ON and L=ON are both given. But in a model where some variables (S2, L) are not given (bottom left of Table 2), the globally necessary fact that S1=ON does not provide enough information to infer whether L=ON.

4.2 The models where *C causes E* is true

The next step is to connect these inferential properties to the truth of sentences built from the verb *cause*. On my proposed analysis (12), *C causes E* entails that that C and E both occur, and thus that C is locally sufficient for E. If C occurs, then E can be inferred automatically when C is a globally sufficient cause of E (because global sufficiency entails local sufficiency); but need not follow when C is a globally necessary cause of E. As a result, a sentence of the form *C causes E* is true in a wider range of circumstances when C is a globally sufficient cause of E than when it is a globally necessary cause of E.

The analysis is illustrated by comparing the four quadrants of Table 2. The top row represents the situation where all variables in the model (both switches and the light) are known, so that we can use our local definitions

10 Such inferences are laid out in an experiment by Kun & Weiner (1973), who determine two nodes of a ternary causal structure like this paper's lightbulbs, then ask participants about the third one.

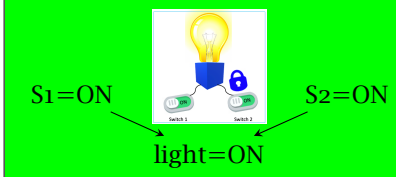
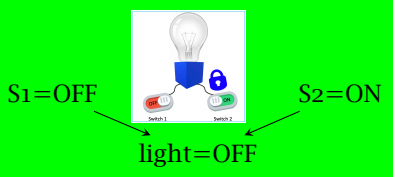
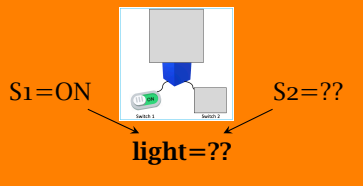
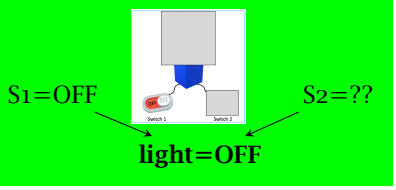
	Globally necessary *C*	**Globally sufficient** *C*
all vars certain	*S1=ON causes L=ON* is true **Locally sufficient** 	*S1=OFF causes L=OFF* is true **Locally sufficient** 
some vars uncertain	*S1=ON causes L=ON* is false **Not sure if locally sufficient!** **Not sure if L=ON!** 	*S1=OFF causes L=OFF* is true **Globally, thus locally sufficient!** **Can surely infer that L=OFF!** 

Table 2 Asymmetric inferences from necessary versus sufficient causes under uncertainty.

of necessity and sufficiency. If we assume that S2=ON (top row), then it's true that *S1=ON causes L=ON* (top left), and that *S1=OFF causes L=OFF* (top right): in both cases, holding S2=ON constant, the state of S1 is locally sufficient for the light to be on/off, as required by (12).

The bottom row of Table 2 represents situations of uncertainty, where we are not sure of the value of some variables and thus must rely on global notions of necessity and sufficiency. If we don't know S2, then it's not true that *S1=ON causes L=ON* (bottom left, in orange) — because, without knowing S2, we don't know whether L=ON, thus we don't know whether S1=ON is locally sufficient for L=ON, as required by (12). In contrast, if we know that S1=OFF, then — even without being given the state of S2 or the light — it is true that *S1=OFF causes L=OFF* (bottom right, in green), because S1=OFF is globally and thus locally sufficient for L=OFF, consistent with (12).

In sum, Table 2 — particularly the orange and green cells in the bottom row — show that necessity and sufficiency license asymmetric inferences under uncertainty. When the states of the second switch and the light are not given, the state of the light is left open when S1=ON is globally necessary for L=ON (bottom left in orange), but can be inferred with certainty when

S1=OFF is globally sufficient for L=OFF (bottom right in green). As a result, *C causes E* is true in a wider range of circumstances when *C* is globally sufficient for *E* than when it is a globally necessary. For a globally necessary *C*, *C causes E* is only true when all relevant upstream variables are known to be fixed at the right setting (top left, in green), but not under uncertainty (bottom left, in orange). In contrast, for a globally sufficient *C*, it is true when *C* alone is fixed at the right setting, regardless of whether other variables are fixed (top right, in green) or unsettled (bottom right, also in green). This fact, I argue, is key to explaining why *cause* tends to combine with negative-sentiment complements.

5 Linking sufficiency to sentiment

The next step is to explain why this logical asymmetry has emotional consequences, namely by leveraging the Anna Karenina Principle (23a)–(23b) that good outcomes tend to have many individually necessary-but-insufficient causes, such that the absence of any of them suffices for a bad outcome: you have to do everything right to succeed; you only have to do one thing wrong to fail.

(23) a. "All happy families are alike; each unhappy family is unhappy in its own way." — novelist Leo Tolstoy, *Anna Karenina*, 1878

b. "It is possible to fail in many ways …while to succeed is possible only in one way (for which reason also one is easy and the other difficult)." — philosopher Aristotle, *The Nichomachean Ethics*, 350 B.C.E.

This principle was named by Diamond (1997), a human geographer, to explain why indigenous animals of many continents were not suitable for domestication; it is invoked[11] in economics, ecology, mathematics, and elsewhere. In linguistics, the Anna Karenina Principle underlies the finding from Sassoon (2013) that the positive multidimensional adjective *healthy* means "healthy along *every* relevant dimension," whereas its negative antonym *unhealthy* means "unhealthy on *some* relevant dimension" — linking desirable states to multiple individually-necessary-but-insufficient factors, such that the absence of any of them suffices for an undesirable state. In social psychology, the Anna Karenina Principle emerges from the idea (Kanouse 1984)

11 Please see https://en.wikipedia.org/wiki/Anna_Karenina_principle.

that "one rancid ingredient can spoil the finest soup" (a single bad attribute overrides countless good ones). The Anna Karenina Principle also captures the finding of Liu, Karasawa & Weiner (1992) that positive emotions are more likely to be ascribed to multiple conjunctive (individually necessary, jointly-sufficient) factors while negative emotions are ascribed to single sufficient factors. Towards the goal of explaining why *cause* favors negative-sentiment complements, the Anna Karenina Principle links the logical necessity/sufficiency asymmetry to an emotional distinction between good and bad outcomes.

In the lightbulb models above (Section 4), L=ON requires two conjunctive conditions (S1=ON, S2=ON) that are individually necessary and jointly sufficient for the light to be on; while L=OFF requires only one of two disjunctive conditions (S1=OFF, S2=OFF) which are individually sufficient for the light to be off. People may be biased to view light as metaphorically preferable over darkness, but these models do not specify which state of the light is desirable; if the structure of the light were reversed so that L=OFF required both switches to be off, our judgments should reflect the logical pattern that globally sufficient conditions license stronger inferences than globally necessary conditions under uncertainty, rather than anything about the desired state of the light. But the Anna Karenina Principle posits that when we move beyond the constrained, fully explicit lightbulb model to the uncertain and subjective models used in the real world, the logical structure of a causal model aligns with the desirability of its outcome.

5.1 Deriving the Anna Karenina Principle

In the social psychology literature, Alves, Koch & Unkelbach (2017) use the Anna Karenina Principle to explain why, across dozens of studies of memory and processing, positively-valenced information is treated as homogenous while negatively-valenced information is more diverse. They derive the Anna Karenina Principle from a deeper "range" principle (named the "Goldilocks principle" by Nouwen 2021), used in astrophysics to characterize habitable planets, that continuous properties such as temperature are only survivable or pleasant at an intermediate range. There is one pleasant range ("just right") and two unpleasant ones (too hot, too cold); thus desirable temperatures are similar to one another, while undesirable ones are heterogeneous.

Alves, Koch & Unkelbach (2017) succeed in explaining why positive information is processed homogenously, which constitutes one facet of the Anna

Karenina Principle ("all happy families are alike"). But their explanation is limited to continuous properties such as temperature, without addressing boolean causal structure: that many individually necessary-but-insufficient factors contribute to a good outcome, while a single factor suffices for a bad one. The same authors (Unkelbach, Alves & Koch 2020) mention that an ideal food should be the right temperature *and* the right level of spice, yielding one way to succeed (perfectly warm, perfectly spicy) and eight ways to fail (every possible combination of too hot, too cold; too spicy, too bland). But they do not explain — leveraging von Wright's insight (von Wright 1974) that the absence of a necessary C is sufficient for not-E — why good food needs a conjunction of desirable properties while a bad food needs only a disjunction of undesirable ones. Thus, there is still room to derive the boolean dimensions of the Anna Karenina Principle from a deeper mechanism.

What makes an outcome "good" or "bad" (according to whom?) and why would such outcomes be assigned different types of causal models? Clarifying the Anna Karenina Principle is valuable not just for its own sake but for what it can tell us about how the subjective endeavor of constructing a causal model (discussed in Section 3.2) is shaped by the desirability of an outcome.

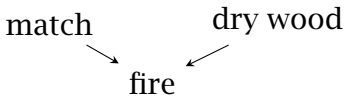
The explanation presented here is grounded in the insight that there is no single correct causal model of a given situation. I propose that when an agent desires an outcome, they view it as an uphill battle requiring many factors that are each necessary but only jointly sufficient for the desired outcome, whereas its alternative is seen as a downhill default requiring any single sufficient factor (the absence of any of the factors necessary for success). In contrast, when an agent wants to avoid an outcome, they view it as a downhill default requiring any single sufficient factor — while viewing its prevention as the uphill battle requiring multiple necessary-but-individually-insufficient factors. The Anna Karenina Principle is derived from the idea that people construct different causal models, uphill battles versus downhill defaults, for the situations that they view as desirable versus undesirable.

Imagine, for example, that a person wants to bring about a fire. I suggest that they will tend to construct a model representing fire as an uphill battle requiring multiple factors (dry wood, matches) which are individually necessary but only jointly sufficient for fire (24a). In contrast, they will tend to view the absence of fire as a downhill default, for which the absence of any necessary factor suffices. Indeed, fire-making instructions mention several

necessary conditions for a fire (24b)–(24c); the absence of any of them would be sufficient for failure.

(24) **Goal: Bring about a fire**

a. *Fire (desirable) happens when* …(both necessary)



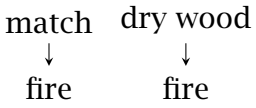
b. "You will need a fire starter [lighter or matches], tinder, small kindling, large kindling, and fuelwood …Wood that is aged, dry and brittle will burn best."[12]

c. "The three requirements for fire — heat, air, and fuel — must come together in the right ratio to burn properly."[13]

On the other hand, imagine that a person wants to prevent a fire. I suggest that they will tend to view fire as a downhill default (25a), for which a single factor suffices. In websites dedicated to fire prevention (25b)–(25c), fire is attributed to a single factor (children playing with matches; lint left in one's clothing dryer), in contrast to the multiple factors discussed (24b)–(24c) in literature for people who want to start fires.

(25) **Goal: Prevent a fire**

a. *Fire (undesirable) happens when* …(each sufficient)



b. "Many fires have been caused by children playing with matches or lighters."[14]

c. "The most common cause of dryer fires is the result of lint build-up in the dryer."[15]

12 "How to make a fire," by Heather Swift, Alderleaf Wilderness College website, https://www.wildernesscollege.com/how-to-make-a-fire.html, accessed June 2022.

13 "How to start a fire: The ultimate guide to modern fire building," by Kevin Estela, *Outdoor Life* magazine, published November 2021.

14 "Matches and lighters," City of Phoenix fire safety webpage, https://www.phoenix.gov/fire/safety-information/fire-safety/matches, accessed June 2022.

15 "Dryer fires: Common causes and prevention tips," by T. David Harlow, *Envista Forensics*, https://www.envistaforensics.com/knowledge-center/insights/articles/dryer-fires-common-causes-and-prevention-tips/, published March 2022.

In other words, I argue that even the same event (fire) is given different causal models depending on whether a person views it as desirable or undesirable. There is no single correct model (Section 3.2); instead, the best model is the one that best guides action. Taking more (rather than less) action towards one's goal is always helpful, so if a person's causal model leads them to act on multiple factors each viewed as necessary-but-insufficient, they are probably more likely to succeed than if they act on only a single factor viewed as sufficient. A person who wants a fire is more likely to succeed if they act on a model like (24a); a person who wants to prevent fire is better served by acting on a model like (25a).[16]

To illustrate, imagine that two camp counselors each want to start a fire. Multi-Factor Maya constructs a model for a fire requiring both matches and dry wood (24a). Single-Factor Sam imagines that just matches are enough (25a), assuming without question that dry wood will be present just as oxygen will. In fact, there is a rainstorm and the wood gets wet. Multi-Factor Maya planned ahead and brought a bag of dry wood along with matches. Single-Factor Sam only brought matches. Multi-Factor Maya succeeds in lighting a fire because her model led her to take more action towards her goal (bringing extra wood). Single-Factor Sam fails because his model ignored a key factor, leading him to take less action towards this goal.

Now, imagine that the two counselors want to prevent fire in a campground where fires are banned. Again, Multi-Factor Maya imagines that fire requires both matches and dry wood (24a); Single-Factor Sam imagines that just matches are enough (25a), again assuming without modeling it that dry wood may be present also. Multi-Factor Maya believes that the wood will be wet from Monday's rain, so she leaves her matches out for children to find. Single-Factor Sam sees the matches as a fire hazard, so he hides them away. In fact, the wood has dried up since Monday. Single-Factor Sam's campers stay safe because his model led him to take more action towards his goal (hiding the matches). Multi-Factor Maya's campers are in danger because her model assumed the wrong value for a key factor, leading her to take less action towards this goal.

In other words, in case we are wrong about some factor that we ignored or mis-valued, we are better served by causal models that lead us to take more

16 As another example, a would-be parent may construct a model of pregnancy with many necessary-but-insufficient factors (ovulation timing, health of sperm and eggs, ideal balance of hormones, and so on), while a high-schooler is better off using a model in which pregnancy is caused by a single sufficient factor (unprotected sex).

rather than less action towards our goals. To do so, we should imagine multiple necessary-but-individually-insufficient factors for desirable outcomes, such that the absence of any of them suffices for an undesirable one. We should think like Multi-Factor Maya for good outcomes, and Single-Factor Sam for bad ones.

This analysis derives the Anna Karenina Principle that good events have many necessary-but-individually-insufficient causes while bad events have single sufficient causes. Rather than requiring events to be classified objectively as good or bad or stipulating that good versus bad events require different causal models, I suggest that people construct different causal models for the outcomes that they *view* as good versus bad. The Anna Karenina Principle emerges because, in the subjective and goal-dependent task of constructing a causal model, it is strategic to treat one's desired outcome as an uphill battle requiring many necessary-but-individually-insufficient causes. Even without invoking the Goldilocks Principle that intermediate ranges are more desirable than extremes, this analysis predicts the key fact from Alves, Koch & Unkelbach (2017) that bad outcomes are heterogeneous: each one might involve the absence of a different necessary factor for the desired result.

5.2 Experiment: Causal models for desirable versus undesirable outcomes

On this proposed explanation for the Anna Karenina Principle, the same outcome should be assigned different causal models depending on whether it is viewed as good versus bad. This prediction was tested in an experiment. In each item, the same outcome (here, *a fire in your living room*) was randomly assigned to one of two conditions, one where you "want" it (26) and one where you "don't want" it (27). The outcomes in the six experimental items (*fire in your living room, wildflowers in your yard,* and so on; (28)) were chosen so that it is plausible for a person to both want or not want that outcome, as rationalized by the "because" clauses (*because fires are cheerful/dangerous, because wildflowers are beautiful/weeds*, and so on).

In each item, it is then stated that the outcome ("indeed/nevertheless") occurs, and then participants are given a binary choice as to whether "you had to do everything {right/wrong}" for this outcome, or whether "you only had to do one thing {right/wrong}" — in essence, asking whether they con-

struct a causal model in which this outcome required many necessary-but-individually-insufficient factors or a single sufficient factor.

(26) You want *a fire in your living room*, because fires are cheerful. After some effort on your part, you indeed end up with *a fire in your living room*. Which statement is more true?

a. You had to do EVERYTHING **right** for *a fire in your living room* to happen.
b. You only had to do ONE thing **right** for *a fire in your living room* to happen.

(27) You don't want *a fire in your living room*, because fires are dangerous. After some effort on your part, you nevertheless end up with *a fire in your living room*. Which statement is more true?

a. You had to do EVERYTHING **wrong** for *a fire in your living room* to happen.
b. You only had to do ONE thing **wrong** for *a fire in your living room* to happen.

(28) **Experimental items**: You {want/don't want} …

a. a fire in your living room, because fires are {cheerful/dangerous}.
b. smoke in your oven, because smoke makes {delicious flavor/foul odor}.
c. wildflowers in your yard, because wildflowers are {beautiful/weeds}.
d. moss on your statue, because moss is {rustic/ugly}.
e. a loud party in your basement, because loud parties are {fun/obnoxious}.
f. a romance in your office, because romances are {heart-warming/unprofessional}.

(29) **Fillers**: You try to hasten …

a. a reconciliation between friends, because you are making peace.
b. a pot of water boiling, because you are making pasta.
c. a dough rising, because you are making bread.

The hypothesis is that the "want" condition should elicit more "you had to do EVERYTHING right" responses, while the "don't want" condition should elicit more "you only had to do ONE thing wrong" responses. Such a result

would show that people construct different causal models for an outcome depending on whether they view it as desirable or undesirable.

After each item, a follow-up question (30) was asked as an attention check.

(30) Did you want *a fire in your living room* to happen? Yes No

At the end of the experiment, a final attention check asked participants to select the number equal to "half of ten" (i.e., 5).

The six items were presented in a random order on the Qualtrics platform and interspersed with three fillers (29) which used the verb phrase *try to hasten* and the adverb *fast* rather than *want/right* or *don't want/wrong*—thus, a two-to-one ratio of items to fillers. Following a pre-registered plan, the experiment was presented on the paid Prolific service to 112[17] self-identified native English speakers geolocated in the United States; 11 were excluded for failing more than one attention check or for giving the same answer to every single question, leaving 101 participants for analysis. Data were analyzed in R (R Core Team 2012) using a binary logistic regression predicting "response" ("EVERYTHING" versus "ONE thing right/wrong") as a function of Condition ("want" versus "don't want"), with random intercepts for participants and items. The regression finds a 22.8% chance of an "EVERYTHING right/wrong" response for the "don't want" condition (27), versus a 65.7% chance for the "want" condition (26), a highly significant effect ($\beta = 1.87, z = 9.98, p < 0.001$). The same result is found in a χ^2 test on the response counts in a 2x2 ("want" versus "don't want," "EVERYTHING" versus "ONE thing right/wrong") contingency table ($\chi^2 = 106.21, p < 0.001$). As predicted (Figure 3), the "want" condition (26) elicits more "you had to do EVERYTHING right" responses, while the "don't want" condition elicits more "you only had to do ONE thing wrong" responses.

Grounding the Anna Karenina Principle, these results are consistent with the claim that desirable outcomes are assigned different causal models than undesirable ones. This principle arises from the deeper claim that people favor causal models that guide them to act on multiple fronts to achieve their goal. While a fire can be desirable or undesirable depending on one's perspective, I suggest that the events described by positively-valenced words (*success, celebration*) will typically be assigned the causal structure of desir-

17 The intention was to run 110 participants, but 112 completed the study because two did so after the survey had timed out and Prolific had already recruited replacements.

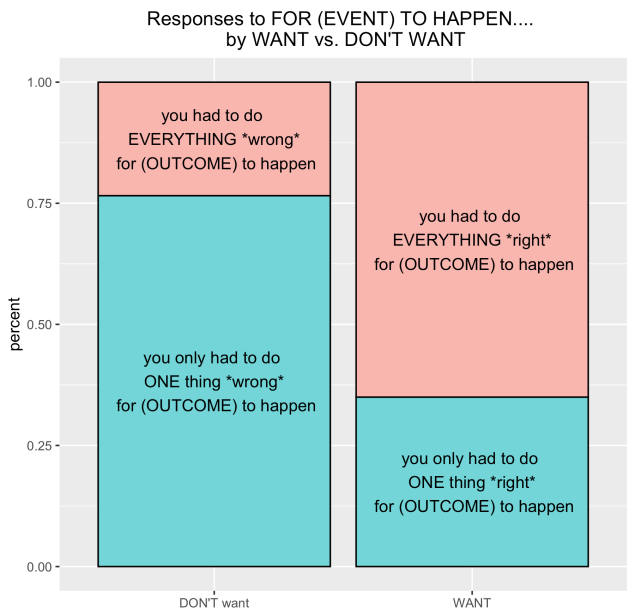


Figure 3 Experimental results. Participants tend to say that if an outcome happens that they *want*, they "had to do EVERYTHING right" — while if an outcome happens that they *don't want*, they "had to do only ONE thing wrong."

able outcomes, while those described by negatively-valenced words (*damage, problems*) will typically be assigned the structure of undesirable outcomes. Thus, as a step towards to explaining why *cause* favors negative-sentiment complements, the Anna Karenina Principle links the logical structure of a causal model to the desirability of its outcome.

6 Why *cause* favors negative-sentiment complements

Putting all the pieces together, it is argued that a sentence of the form *C causes E* is true in more distinct contexts when *E* is a bad outcome with a single sufficient cause. The analysis is illustrated using *success* and *failure* to represent any good/bad outcomes. Imagine a simple causal model (inspired by Kun & Weiner 1973) where, following the Anna Karenina Principle, both work and luck are necessary for success; their conjunction is sufficient for success; and the absence of either is sufficient for failure.

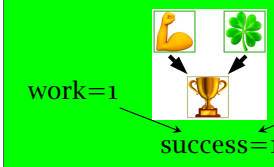
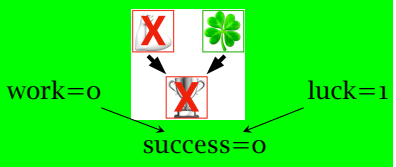
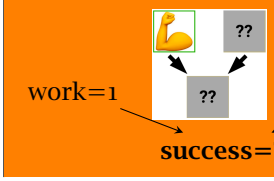
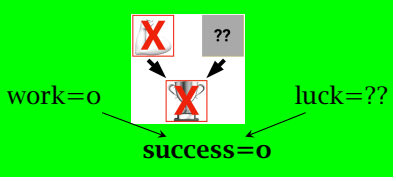
	Globally necessary *C* (success)	**Globally sufficient *C* (failure)**
all vars certain	'Work causes Success' is true **Locally sufficient**  	'Laziness causes Failure' is true **Locally sufficient** 
some vars uncertain	'Work causes Success' is false! **Not sure if locally sufficient!** **Not sure if Success!**  	'Laziness causes Failure' is true **Globally, thus locally sufficient!** **Can surely infer Failure!** 

Table 3 The analysis, combining the logical asymmetry between necessity and sufficiency with the Anna Karenina Principle that good outcomes have necessary-but-individually-insufficient causes while bad outcomes have sufficient causes. *C causes failure* (globally sufficient) is true in a wider range of situations than *C causes success* (globally necessary).

As shown in Table 3, *work causes success* is true only in a state of full certainty about all relevant variables — namely, when we know both *work=1* and *success=1*, which also requires us to know *luck=1*. *Work causes success* is false in a state of uncertainty, when we know only *work=1*, because without knowing *luck*, we cannot infer *success=1*. In contrast, *laziness causes failure* is true both in the state of full certainty about all variables, and in a state where all we know is *work=0* — because even without knowing whether the person is lucky, *success=0* can be safely inferred.

Combining the logical necessity/sufficiency asymmetry and the Anna Karenina Principle, the proposal is that these ingredients shape language usage as observed in corpora. For example, since depression is an outcome that people generally aim to avoid, people may tend to assign it a causal model with single sufficient factors — complementary to happiness, which may be given a model with many factors that are each individually necessary but only

jointly sufficient. If depression is given a model with one or more individually globally sufficient causes, then a statement describing such a cause (*poverty causes depression*) can be true even when other factors are unsettled. In contrast, a complementary claim about a necessary-but-individually-insufficient cause — *wealth causes happiness* — is not true unless all other factors relevant to happiness (health, community, purpose) are fixed at the right values. If *C causes E* is more often true of bad outcomes, that explains why *cause* favors negative-sentiment complements in usage.

7 An alternative analysis: Causes as violations of norms

This paper's proposed analysis is defended over an alternative whereby *cause* favors negative-sentiment complements because it is used to profile norm violations and assign blame (Hart & Honoré 1959, Hilton & Slugoski 1986, Cheng & Novick 1991, Hitchcock & Knobe 2009, Alicke, Rose & Bloom 2011).

7.1 Insight: Cause as blame

In reality (highlighted by Mill 1843), most events depend on multiple factors: a house fire depends not just on a short circuit but also on the presence of electricity, oxygen, and dry wood (Mackie 1965). And yet in language, most tokens of the verb *cause* occur with a single noun as its subject, raising the "causal selection" question (Hart & Honoré 1959, Hilton & Slugoski 1986, Cheng & Novick 1991, Hitchcock & Knobe 2009, Neeleman & Van de Koot 2012, Baglini & Bar-Asher Siegal 2020, Bar-Asher Siegal, Bassel & Hagmayer 2021): how is one factor profiled over the others as "the" cause of the effect?

This literature has found that people are most likely to profile actions over inactions, unusual events over usual ones, and — most relevant here — norm-violating events over norm-conforming ones. Following Reuter et al. (2014) and Kominsky et al. (2015), imagine a computer security system which locks if two people are logged in at once. Alice and Bob both log in (a conjunctive scenario; both necessary) and the system locks. Experiments show that people are more willing to endorse Alice (31a) over Bob (31b) as the cause if Bob is supposed to be logged in while Alice is not, profiling the norm violator as the locus of blame (Hart & Honoré 1959, Hilton & Slugoski 1986, Alicke 1992, Hitchcock & Knobe 2009, Alicke, Rose & Bloom 2011).

(31) a. Alice caused the system to lock.
b. Bob caused the system to lock.

To explain this pattern, Alicke, Rose & Bloom (2011) suggest that judgments of causation are moral judgments of blame: *C* is more strongly profiled as the cause of *E* when *C* is blamed for *E*, because *C* violates norms and/or because *E* is undesirable. Here, it is not necessary for the system locking to be cast as undesirable; but it may be perceived as such because it is attributed to Alice's norm-violating behavior. Such an explanation does not rely on a specific semantics for *cause*, but rather on the pragmatic process of selecting one among many contributing factors as "the" cause. Assuming that norm violators are pinpointed as "the" cause, then the consequences of their norm-violating behavior may tend to be negative, offering an alternative explanation of why *cause* patterns as it does.

7.2 Critique: Doesn't work for sufficient causes

Icard, Kominsky & Knobe (2017), formalizing a mathematical account of causal selection, offer an alternative explanation for why the norm-violating Alice (31a) is profiled as the cause: that people consider counterfactual situations in proportion how normal they are, statistically or morally. People focus on the counterfactual where only normative events occur (where only Bob logs in); there, the system doesn't lock, so Alice's norm-violating contribution is chosen as the cause. Whereas Alicke, Rose & Bloom (2011) directly claim that *cause* is used to blame norm violators, Icard, Kominsky & Knobe (2017) derive that effect from the counterfactual situations that people consider most relevant.

These proposals come apart when applied to sufficient causes. Imagine instead that the computer security system locks if any *one* person logs in; Alice and Bob both log in (a disjunctive scenario; each sufficient) and the system locks. Now people are asked about the extent to which they endorse Alice (31a) or Bob (31b) as the cause of the system locking — again, a question of causal selection, in that we must select which of two individually sufficient factors to profile as "the" cause.

Here, in contrast to the conjunctive scenario above, Icard, Kominsky & Knobe (2017) show experimentally that people are more willing to endorse Bob (31b) over Alice (31a) when Bob is *supposed to be* logged in.

For Alicke, Rose & Bloom (2011), it is surprising that the norm-conformer is chosen as the cause, contrary to their claim that *cause* seeks to blame norm violators. But in the framework of Icard, Kominsky & Knobe (2017), people again consider the counterfactual where only normative events occur (where

only Bob logs in); there, the system still locks, so Bob's norm-conforming contribution is chosen as the cause.

Whether or not one accepts the counterfactual sampling framework of Icard, Kominsky & Knobe (2017), the important point is their empirical finding: for sufficient causes, the factor that people tend to profile as "the" cause is actually the norm-conforming one, thus complicating the attempt to explain the negative collocation of *cause* using the idea that *cause* profiles norm violations. In fact, *cause* does not always profile norm violators, so an explanation built on that assumption would stand on shaky ground.

The norm-violating explanation for the negative collocation of *cause* faces even more trouble if it is connected to the Anna Karenina Principle. The Anna Karenina Principle states that good outcomes tend to have multiple individually necessary-but-insufficient causes whereas bad outcomes tend to have single sufficient causes (Section 5; which arises in turn, I argue, because people construct different causal models for the situations that they see as desirable versus undesirable). In disjunctive scenarios with single sufficient causes, Icard, Kominsky & Knobe (2017) find experimentally that it is actually the norm-conformer who is chosen as "the" cause. Putting these pieces together, one might arrive at the counter-intuitive prediction that bad events with single sufficient causes should be brought about by norm conformers. One might expect that normative behavior should lead to a good outcome, so it is surprising that norm conformers are connected to bad outcomes when we combine the causal selection theory of Alicke, Rose & Bloom (2011) with the Anna Karenina Principle.

In sum, although one might think that *cause* favors negative-sentiment complements because *cause* is used to profile norm violations, that explanation breaks down when it is applied to sufficient causes, for which Icard, Kominsky & Knobe (2017) show that norm-conforming actions are chosen as causes. But on this paper's proposal, sufficiency is leveraged in the explanation rather than confounding it, which is a point in its favor.

8 Beyond *cause*

One might ask whether this analysis extends to *cause*'s more frequent cousin, *because*. If *E because C* is analyzed theoretically in the same way as *C caused E* (McHugh 2023), then perhaps *because* should also empirically be associated with negative sentiment.

To test this prediction, I explored *because* in the same ten million words of data from AskReddit introduced above (Section 2). Using the SpaCy dependency parser (Honnibal & Johnson 2015), I identified all tokens part-of-speech-tagged as sentential conjunctions (SCONJ): *if, as, because, than, that*, and so on. I standardized all spellings of *because* (*cause, 'cause, cuz*). I excluded cases where the SCONJ introduces the complement to a clause-embedding verb (32).

(32) a. And I know **that**$_{\text{SCONJ}}$ millions more are just like me. (excluded)
b. I wonder **if**$_{\text{SCONJ}}$ that's a thing in Korea? (excluded)

For each SCONJ token, I identified its main clause as well as the subordinate clause or prepositional object introduced by the SCONJ (33)–(34); then I gathered the sentiment of each of these according to the Hedonometer (Dodds et al. 2011).

(33) a. [I once failed a semester of college]$_{\text{MAIN}}$ [**because**$_{\text{SCONJ}}$ I spent every waking moment trying to get the golden sniper rifle in COD-MW]$_{\text{SUBORD}}$.
b. [It's really **because**$_{\text{SCONJ}}$ of [the hospital beds]$_{\text{POBJ}}$]$_{\text{MAIN}}$.

(34) a. [So it would cost $1 for a dozen eggs]$_{\text{MAIN}}$ [**if**$_{\text{SCONJ}}$ you raise the hens yourself]$_{\text{SUBORD}}$.
b. [Also, I've had a cold]$_{\text{MAIN}}$ **since**$_{\text{SCONJ}}$ [November]$_{\text{POBJ}}$.

Whereas *cause* occurs 226 times per million in the Reddit data and serves as the main verb in 512 clauses, *because* occurs more than ten times as often, at 2899 per million; it appears as SCONJ in 20,557 sentences. Comparing *because* to all other SCONJ (33)–(34), I find that *because* combines with slightly more negative main clauses (mean = 5.70 versus 5.89) and slightly more negative subordinate clauses (mean = 5.81 versus 5.97) — small effect sizes, but still highly significant ($p < 0.001$) in unpaired, two-sided t tests. In contrast, its prepositional objects (33b) are no different in sentiment from those introduced by any other SCONJ (mean = 4.55 versus 4.49, not significantly different in a t test). Figure 4 shows the mean Hedonometer rating for main clauses, subordinate clauses, and prepositional objects of all other SCONJ compared to *because*.

In comparing *cause* to other verbs above (Section 2), the most striking finding was that the direct objects of *cause* — the words describing its effect; *cause cancer, cause depression,* and so on — are more negative in sentiment

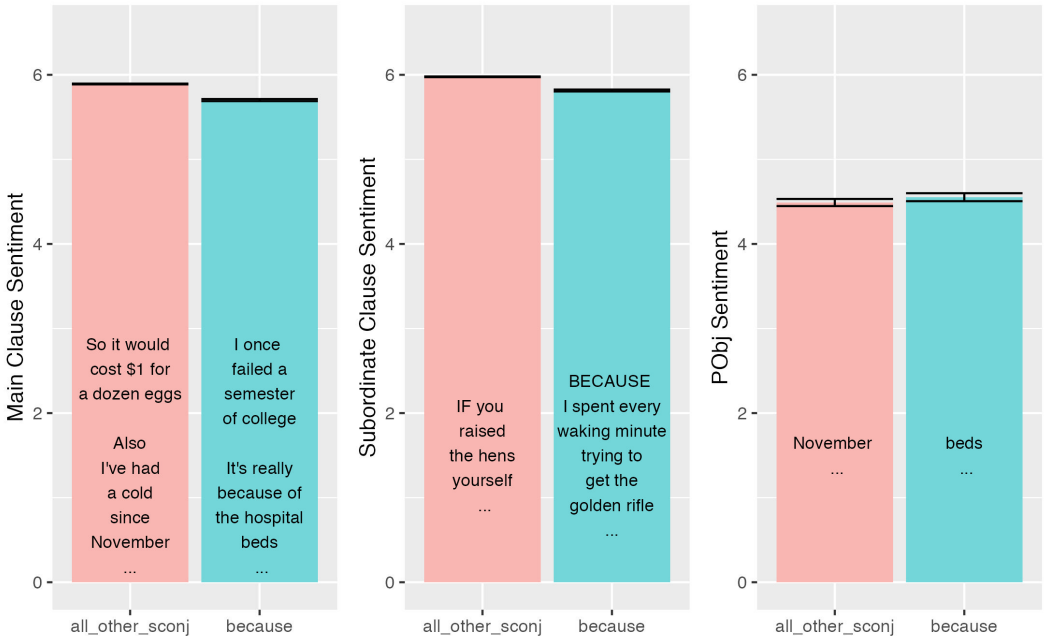


Figure 4 Mean Hedonometer sentiment for all main clauses of SCONJ in Reddit data; all subordinate clauses; and all prepositional objects for all other SCONJ versus *because*. Error bars represent the standard error of the mean.

compared to those of other verbs (mean = 3.96 versus 5.66). Turning to *because*, the most striking finding is that its main clauses — *I failed a semester of college because* (33) — are more negative in sentiment than those of other SCONJ (mean = 5.70 versus 5.89): again, the part of the sentence describing its effect. The difference is smaller for *because*, but the overall picture is the same: *cause* and *because* are both associated with effects (encoded as direct objects and main clauses, respectively) that are negative in sentiment. Thus, I suggest, *because* can also be included in this paper's proposed explanation of why *cause* is associated with negative collocational sentiment.

Further widening the lens, the same research program can be expanded to other cases in which the sentiment associated with a lexical item can be grounded in the causal model proposed to handle its meaning. As inspiration, Baglini & Francez (2016) observe that *manage* suggests that its complement is desirable from the perspective of the sentential subject (*I managed*

to win), and sounds ironic when combined with undesirable complements (*I managed to miss my flight*). They argue that *manage* presupposes the occurrence of a catalyst which is normally necessary-but-insufficient for the result, and asserts that on this occasion the catalyst was in fact (in my terms, "locally") sufficient for the result. From the perspective of a party who instigated the catalyst, it is fortunate that other favorable conditions allowed the catalyst to be locally sufficient for the result, so that — just as I propose for *cause* — the sentiment of *manage* is derived from the causal model attributed to it.

Future work might explore *cause*'s near-synonyms, such as *make, produce*, and *bring about*, which seem to differ from *cause* in both meaning (Nadathur & Lauer 2020) and sentiment (Childers 2016, Hauser & Schwarz 2018). Work in the same tradition could also consider "adversative" (Chappell 1980) *get*-passives and causatives (*he got arrested, we got them to dance*), which can connote negative consequences for the affected party. Beyond English, this research program might illuminate why French *à cause de* 'because of' is famously identified with negative sentiment (Goosse & Grevisse 2008: p. 1383), in contrast to *grâce à* 'thanks to' which is positive. Even though many of these words seem to share a unified function of describing causation, it may be productive to adopt the assumption of "causal pluralism" (Copley & Wolff 2014, Bar-Asher Siegal & Boneh 2020): the idea that each word might have a different meaning to be defined using causal models, rather than a shared atomic causal core. Therefore, this paper opens up a research program of exploring how the sentiment of many different causal words can derived from the unique meaning of each one.

9 Conclusion

Aiming to explain why the verb *cause* tends to occur with negative-sentiment complements, it is argued that *C causes E* is true in a wider variety of uncertain contexts when *C* is a globally sufficient cause of *E* than when it is globally necessary. This inferential asymmetry is combined with the Anna Karenina Principle that good outcomes tend to have multiple necessary-but-individually-insufficient causes while bad outcomes tend to have single sufficient causes, so that *C causes E* is true in more contexts for bad outcomes than good outcomes.

By connecting formal semantics with the idea of sentiment used in natural language processing, this paper takes inspiration from Potts (2011), who

derives the negative collocation of *not* from its core semantics and pragmatics; and more recently from Nouwen (2021). Nouwen observes that when an adverb is derived from a positive-sentiment gradable adjective (*pretty, pleasant*), it denotes a moderate degree of the property it combines with (*pretty/pleasantly hot*), whereas when it is derived from a negative-sentiment adjective (*terrible, painful*), it denotes an extreme degree (*terribly/painfully hot*). Nouwen suggests that positive-sentiment modifiers pick out the most desirable, moderate "Goldilocks zone" — not too much, not too little (Section 5.1) — along a property's associated scale, whereas negative-sentiment modifiers describe unpleasant extremes. Thus, Nouwen brings together textual sentiment with an independently motivated principle about what counts as desirable in order to explain the interpretation of adverbial modifiers, just as I do here with the Anna Karenina Principle to explain the sentiment of *cause*'s complements.

Using tools from philosophy and statistics, this paper builds on recent work leveraging causal models in lexical semantics (Sloman, Barbey & Hotaling 2009, Baglini & Francez 2016, Nadathur 2016, Martin 2018, Nadathur & Lauer 2020, Baglini & Bar-Asher Siegal 2020, Nadathur & Bar-Asher Siegal 2022). With the proposed distinctions between global and local necessity and sufficiency, the paper contributes to that literature a framework for thinking about the common scenario where interlocutors may be uncertain about the causal model under discussion.

The subjective uncertainty of causal models is also used to ground an explanation of the Anna Karenina Principle, which transcends linguistics: I propose that people construct different models for the events that they see as desirable versus undesirable, strategically favoring causal models that guide them to take action on multiple fronts to achieve their goal. I leave it to future work to assess whether this explanation of the Anna Karenina Principle can be extended to other phenomena for which that principle is invoked.

Although it is just one word, *cause* deserves attention because the concept that it denotes is crucial to many domains of inquiry. This paper leverages tools from natural language processing and philosophy to explore how the formal, emotional, and distributional dimensions of *cause* are linked.

References

Acton, Eric K. 2019. Pragmatics and the social life of the English definite article. *Language* 95(1). 37–65. https://doi.org/10.1353/lan.2019.0010.

Acton, Eric K. & Christopher Potts. 2014. That straight talk: Sarah Palin and the sociolinguistics of demonstratives. *Journal of Sociolinguistics* 18(1). 3–31. https://doi.org/10.1111/josl.12062.

Alicke, Mark D. 1992. Culpable causation. *Journal of Personality and Social Psychology* 63(3). 368–378. https://doi.org/10.1037/0022-3514.63.3.368.

Alicke, Mark D., David Rose & Dori Bloom. 2011. Causation, norm violation, and culpable control. *The Journal of Philosophy* 108(12). 670–696. https://doi.org/10.5840/jphil20111081238.

Alves, Hans, Alex Koch & Christian Unkelbach. 2017. Why good is more alike than bad: Processing implications. *Trends in Cognitive Sciences* 21(2). 69–79. https://doi.org/10.1016/j.tics.2016.12.006.

Bach, Emmon. 1986. The algebra of events. *Linguistics and Philosophy* 9(1). 5–16. https://doi.org/10.1002/9780470758335.ch13.

Baglini, Rebekah & Elitzur A. Bar-Asher Siegal. 2020. Direct causation: A new approach to an old question. In Alexandros Kalomoiros & Lefteris Paparounas (eds.), *Penn Linguistics Colloquium*, vol. 26, 19–28. Philadelphia, PA: University of Pennsylvania Working Papers in Linguistics.

Baglini, Rebekah & Itamar Francez. 2016. The implications of managing. *Journal of Semantics* 33(3). 541–560. https://doi.org/10.1093/jos/ffv007.

Bar-Asher Siegal, Elitzur A., Noa Bassel & York Hagmayer. 2021. Causal selection–the linguistic take. *Experiments in Linguistic Meaning (ELM)* 1. 27–38. https://doi.org/10.3765/elm.1.4887.

Bar-Asher Siegal, Elitzur A. & Nora Boneh. 2019. Sufficient and necessary conditions for a non-unified analysis of causation. 36. 55–60. https://www.lingref.com/cpp/wccfl/36/paper3446.pdf.

Bar-Asher Siegal, Elitzur A. & Nora Boneh. 2020. Causation: From metaphysics to semantics and back. In *Perspectives on causation: Jerusalem studies in philosophy and history of science*, 3–51. Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-030-34308-8_1.

Baumgartner, Jason, Savvas Zannettou, Brian Keegan, Megan Squire & Jeremy Blackburn. 2020. The PushShift Reddit dataset. *International Association for the Advancement of Artificial Intelligence (AAAI) conference on web and social media* 14. 830–839. https://doi.org/10.1609/icwsm.v14i1.7347.

Beller, Ari, Erin Bennett & Tobias Gerstenberg. 2020. The language of causation. In *The 42nd annual conference of the Cognitive Science Society (CogSci)*. Cognitive Science Society. https://www.cognitivesciencesociety.org/cogsci20/papers/0783/.

Beltrama, Andrea. 2016. *Bridging the gap: Intensifiers between semantic and social meaning*. Chicago, IL: University of Chicago dissertation.

Bethard, Steven, William J. Corvey, Sara Klingenstein & James H. Martin. 2008. Building a corpus of temporal-causal structure. *International Conference on Language Resources and Evaluation (LREC)* 6. http://lrec-conf.org/proceedings/lrec2008/pdf/229_paper.pdf.

Carlson, Gregory N. 1977. *Reference to kinds in English*. Amherst, MA: University of Massachusetts dissertation. https://semanticsarchive.net/Archive/jk3NzRlY/carlson.diss.pdf.

Chappell, Hilary. 1980. Is the get-passive adversative? *Research on Language and Social Interaction* 13(3). 411–452. https://doi.org/10.1080/08351818009370504.

Cheng, Patricia W. & Laura R. Novick. 1991. Causes versus enabling conditions. *Cognition* 40(1-2). 83–120. https://doi.org/10.1016/0010-0277(91)90047-8.

Childers, Zachary. 2016. *Cause and affect: Evaluative and emotive parameters of meaning among the periphrastic causative verbs in English*. Austin, TX: University of Texas, Austin dissertation. https://doi.org/2152/46919.

Copley, Bridget & Phillip Wolff. 2014. Theories of causation should inform linguistic theory and vice versa. In Bridget Copley & Fabienne Martin (eds.), *Causation in grammatical structures*, 11–57. Oxford: Oxford University Press.

Davies, Mark. 2008-. *The Corpus of Contemporary American English: One billion words, 1990-present*. https://www.english-corpora.org/coca/ [accessed May 2020].

Diamond, Jared M. 1997. *Guns, germs, and steel: A short history of everybody for the last 13,000 years*. New York & London: W. W. Norton & Company.

Dodds, Peter Sheridan, Eric M. Clark, Suma Desu, Morgan R. Frank, Andrew J. Reagan, Jake Ryland Williams, Lewis Mitchell, Kameron Decker Harris, Isabel M. Kloumann, James P. Bagrow, Karine Megerdoomian, Matthew McMahon, Brian Tivnan & Christopher Danforth. 2015. Human language reveals a universal positivity bias. *National Academy of Sciences* 112(8). 2389–2394. https://doi.org/10.1073/pnas.1411678112.

Dodds, Peter Sheridan, Kameron Decker Harris, Isabel M. Kloumann, Catherine A. Bliss & Christopher M. Danforth. 2011. Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *PloS one* 6(12). e26752. https://doi.org/10.1371/journal.pone.0026752.

Dowty, David R. 1979. *Word meaning and Montague grammar*. Dordrecht: Reidel. https://doi.org/10.1007/978-94-009-9473-7.

Fodor, Jerry A. 1970. Three reasons for not deriving 'kill' from 'cause to die'. *Linguistic Inquiry* 1(4). 429–438. http://www.jstor.org/stable/4177587.

Forni, Mario & Luca Gambetti. 2014. Sufficient information in structural VARs. *Journal of Monetary Economics* 66. 124–136. https://doi.org/10.1016/j.jmoneco.2014.04.005.

Glymour, Clark & Frank Wimberly. 2007. Actual causes and thought experiments. In Joseph Keim Campbell, Michael O'Rourke & Harry Silverstein (eds.), *Causation and explanation*, 43–68. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/1753.003.0005.

Goosse, André & Maurice Grevisse. 2008. *Le bon usage: Grammaire française [Proper usage: French grammar]*. 14th edn. Brussels: De Boeck & Larcier.

Grice, H. Paul. 1975. Logic and conversation. In Peter Cole & Jerry Morgan (eds.), *Syntax and semantics 3: speech acts*, 41–58. Academic Press. Republished as Grice 1989. https://doi.org/10.1163/9789004368811_003.

Grice, H. Paul. 1989. Logic and conversation. In *Studies in the way of words*. Cambridge: Harvard University Press. Republished version of Grice 1975.

Hall, Ned. 2004. Two concepts of causation. In John Collins, Ned Hall & Paul Laurie (eds.), *Causation and counterfactuals*, 225–276. Cambridge, M.A.: MIT Press. https://doi.org/10.7551/mitpress/1752.003.0010.

Halpern, Joseph Y. 2016. Appropriate causal models and the stability of causation. *The Review of Symbolic Logic* 9(1). 76–102. https://doi.org/10.1017/s1755020315000246.

Halpern, Joseph Y. & Judea Pearl. 2005. Causes and explanations: A structural-model approach – part I: causes. *British Journal for the Philosophy of Science* 56(4). 843–887. https://doi.org/10.1093/bjps/axi147.

Hart, Herbert Lionel Adolphus & Tony Honoré. 1959. *Causation in the law*. Oxford: Clarendon Press.

Hauser, David J. & Norbert Schwarz. 2018. How seemingly innocuous words can bias judgment: Semantic prosody and impression formation. *Journal of Experimental Social Psychology* 75. 11–18. https://doi.org/10.1016/j.jesp.2017.10.012.

Hilton, Denis J. & Ben R. Slugoski. 1986. Knowledge-based causal attribution: The abnormal conditions focus model. *Psychological Review* 93(1). 75. https://doi.org/10.1037/0033-295x.93.1.75.

Hitchcock, Christopher. 2020. Causal models. In Edward N. Zalta (ed.), *The Stanford encyclopedia of philosophy*, Summer 2020. Metaphysics Re-

search Lab, Stanford University. https://plato.stanford.edu/archives/sum2020/entries/causal-models/.

Hitchcock, Christopher & Joshua Knobe. 2009. Cause and norm. *The Journal of Philosophy* 106(11). 587–612. https://doi.org/10.5840/jphil2009106112 8.

Hobbs, Jerry R. 2005. Toward a useful concept of causality for lexical semantics. *Journal of Semantics* 22(2). 181–209. https://doi.org/10.1093/jos/ffh024.

Honnibal, Matthew & Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. *Empirical Methods in Natural Language Processing (EMNLP)*. 1373–1378. https://doi.org/10.18653/v1/d15-1162.

Hume, David. 1748. *Enquiries concerning the human understanding: And concerning the principles of morals*. Oxford: Clarendon Press (1902). http://doi.org/10.1093/oseo/instance.00046349.

Hunston, Susan. 2007. Semantic prosody revisited. *International journal of corpus linguistics* 12(2). 249–268. https://doi.org/10.1075/ijcl.12.2.09hun.

Icard, Thomas, Jonathan Kominsky & Joshua Knobe. 2017. Normality and actual causal strength. *Cognition* 161. 80–93. https://doi.org/10.1016/j.cognition.2017.01.010.

Ikuta, Rei, Will Styler, Mariah Hamang, Tim O'Gorman & Martha Palmer. 2014. Challenges of adding causation to Richer Event Descriptions. *Workshop on Events: Definition, Detection, Coreference, and Representation* 2. 12–20. https://doi.org/10.3115/v1/W14-2903.

Kanouse, David E. 1984. Explaining negativity biases in evaluation and choice behavior: Theory and research. In Thomas C. Kinnear (ed.), *North American advances consumer research*, vol. 11, 703–708. Provo, UT: Association for Consumer Research.

Kelley, Harold H. 1973. The processes of causal attribution. *American psychologist* 28(2). 107–128. https://doi.org/10.1037/h0034225.

Kölbel, Max. 2004. Faultless disagreement. 104(1). 53–73. https://doi.org/10.1111/j.0066-7373.2004.00081.x.

Kominsky, Jonathan F., Jonathan Phillips, Tobias Gerstenberg, David Lagnado & Joshua Knobe. 2015. Causal superseding. *Cognition* 137. 196–209. https://doi.org/10.1016/j.cognition.2015.01.013.

Krifka, Manfred, Francis Jeffrey Pelletier, Gregory N. Carlson, Alice ter Meulen, Godehard Link & Gennaro Chierchia. 1995. Genericity: An in-

troduction. In Gregory N. Carlson & Francis Jeffrey Pelletier (eds.), *The generic book*, 1–124. Chicago: University of Chicago Press.

Kun, Anna & Bernard Weiner. 1973. Necessary versus sufficient causal schemata for success and failure. *Journal of Research in Personality* 7(3). 197–207. https://doi.org/10.1016/0092-6566(73)90036-6.

Lagnado, David A., Tobias Gerstenberg & Ro'i Zultan. 2014. Causal responsibility and counterfactuals. *Cognitive Science* 37(6). 1036–1073. https://doi.org/10.1111/cogs.12054.

Leslie, Sarah-Jane & Adam Lerner. 2021. Generic generalizations. In Edward N. Zalta (ed.), *The Stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2021/entries/generics/.

Lewis, David. 1973. Causation. *The Journal of Philosophy* 70(17). 556–567. http://doi.org/10.2307/2025310.

Lewis, David. 2000. Causation as influence. *The Journal of Philosophy* 97(4). 182–197. https://doi.org/10.2307/2678389.

Liu, James H., Kaori Karasawa & Bernard Weiner. 1992. Inferences about the causes of positive and negative emotions. *Personality and Social Psychology Bulletin* 18(5). 603–615. https://doi.org/10.1177/0146167292185011.

Louw, Bill & Carmela Chateau. 2010. Semantic prosody for the 21st century: Are prosodies smoothed in academic contexts? A contextual prosodic theoretical perspective. In Isabella Chiari, Luca Giuliano & Sergio Bolasco (eds.), *Journés d'analyse statistique des données textuelles (JADT)*, vol. 10, 755–764. Rome, Italy: LED Edizioni. https://www.ledonline.it/ledonline/JADT-2010/allegati/JADT-2010-0755-0764_129-Louw.pdf.

Mackie, John L. 1965. Causes and conditions. *American Philosophical Quarterly* 2(4). 245–264. http://www.jstor.org/stable/20009173.

Martin, Fabienne. 2018. Time in probabilistic causation: Direct vs. indirect uses of lexical causative verbs. *Sinn und Bedeutung* 22(2). 107–124. https://doi.org/10.21248/zaspil.61.2018.487.

McCawley, James D. 1976. Remarks on what can cause what. In M. Shibatani (ed.), *The grammar of causative constructions*, vol. 6. New York: Academic Press. https://doi.org/10.1163/9789004368842_004.

McHugh, Dean. 2023. *Causation and modality*. Amsterdam, The Netherlands: Institute for Logic, Language & Computation at the University of Amsterdam dissertation.

Menzies, Peter & Helen Beebee. 2020. Counterfactual theories of causation. In Edward N. Zalta (ed.), *The Stanford encyclopedia of philosophy*, Winter

2020. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/entries/causation-counterfactual/.

Mill, John Stuart. 1843. *A system of logic, ratiocinative and inductive, being a connected view of the principles of evidence and the methods of scientific investigation*. New York: Harper & Brothers. Available via Project Gutenberg. https://doi.org/10.1037/11967-000.

Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross & Katherine J. Miller. 1990. Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography* 3(4). 235–244. https://doi.org/10.1093/ijl/3.4.235.

Nadathur, Prerna. 2016. Causal necessity and sufficiency in implicativity. *Semantics and Linguistic Theory (SALT)* 26. 1002–1021. https://doi.org/10.3765/salt.v26i0.3863.

Nadathur, Prerna & Elitzur A. Bar-Asher Siegal. 2022. Modeling progress: Causal models, event types, and the imperfective paradox. *West Coast Conference in Formal Linguistics (WCCFL)* 40. https://ling.auf.net/lingbuzz/006736.

Nadathur, Prerna & Sven Lauer. 2020. Causal necessity, causal sufficiency, and the implications of causative verbs. *Glossa: a journal of general linguistics* 5(1). https://doi.org/10.5334/gjgl.497.

Neeleman, Ad & Hans Van de Koot. 2012. The linguistic expression of causation. In Martin Everaert & Marijana Marelj (eds.), *The theta system: Argument structure at the interface*, 20–51. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199602513.003.0002.

Nouwen, Rick. 2021. Evaluation, extent, and Goldilocks. Manuscript, University of Utrecht. https://ricknouwen.org/d/goldi.pdf.

Osgood, Charles Egerton, George J. Suci & Percy H. Tannenbaum. 1957. *The measurement of meaning*. Urbana-Champaign, IL: University of Illinois Press.

Pang, Bo & Lillian Lee. 2008. Opinion mining and sentiment analysis. In *Foundations and trends in information retrieval*, vol. 2, 1–135. NOW Publishers. https://doi.org/10.1561/1500000011.

Pearl, Judea. 1999. Probabilities of causation: Three counterfactual interpretations and their identification. *Synthese* 121(1-2). 93–149. https://doi.org/10.1023/A:1005233831499.

Pearl, Judea. 2000. *Causality*. Cambridge: Cambridge University Press. https://doi.org/10.1017/cbo9780511803161.

Pinker, Steven. 1989. *Learnability and cognition: The acquisition of argument structure*. Cambridge, MA: MIT Press. New edition published 2013. https://doi.org/10.7551/mitpress/4158.001.0001.

Potts, Christopher. 2004. *The logic of conventional implicatures*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199273829.001.0001.

Potts, Christopher. 2011. On the negativity of negation. *Semantics and Linguistic Theory (SALT)* 20. 636–659. https://doi.org/10.3765/salt.v0i20.2565.

Potts, Christopher. 2015. Presupposition and implicature. In Shalom Lappin & Chris Fox (eds.), *Handbook of contemporary semantics*. Malden, MA: Wiley-Blackwell. https://doi.org/10.1002/9781118882139.ch6.

R Core Team. 2012. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. http://www.R-project.org/.

Reuter, Kevin, Lara Kirfel, Raphael Van Riel & Luca Barlassina. 2014. The good, the bad, and the timely: How temporal order and moral judgment influence causal selection. *Frontiers in Psychology* 5. https://doi.org/10.3389/fpsyg.2014.01336.

Sassoon, Galit W. 2013. A typology of multidimensional adjectives. *Journal of Semantics* 30(3). 335–380. https://doi.org/10.1093/jos/ffs012.

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

Shibatani, Masayoshi. 1976. The grammar of causative constructions: A conspectus. In Masayoshi Shibatani (ed.), *The grammar of causative constructions*, vol. 6 (Syntax and semantics), 1–40. New York: Academic Press. https://doi.org/10.1163/9789004368842_002.

Sloman, Steven, Aron K. Barbey & Jared M Hotaling. 2009. A causal model theory of the meaning of *cause, enable*, and *prevent*. *Cognitive Science* 33(1). 21–50. https://doi.org/10.1111/j.1551-6709.2008.01002.x.

Stefanowitsch, Anatol & Stefan Th. Gries. 2003. Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics* 8(2). 209–243. https://doi.org/10.1075/ijcl.8.2.03ste.

Stubbs, Michael. 1995. Collocations and semantic profiles: On the cause of the trouble with quantitative studies. *Functions of language* 2(1). 23–55. https://doi.org/10.1075/fol.2.1.03stu.

Talmy, Leonard. 1988. Force dynamics in language and cognition. *Cognitive Science* 12(1). 49–100. https://doi.org/10.1207/s15516709cog1201_2.

Unkelbach, Christian, Hans Alves & Alex Koch. 2020. Negativity bias, positivity bias, and valence asymmetries: Explaining the differential processing of positive and negative information. In Bertram Gawronski (ed.), *Advances in experimental social psychology*, vol. 62, 115–187. Amsterdam, The Netherlands: Elsevier. https://doi.org/10.1016/bs.aesp.2020.04.005.

Wolff, Phillip. 2003. Direct causation in the linguistic coding and individuation of causal events. *Cognition* 88(1). 1–48. https://doi.org/10.1016/s0010-0277(03)00004-0.

Wolff, Phillip. 2007. Representing causation. *Journal of Experimental Psychology: General* 136(1). 82. https://doi.org/10.1037/0096-3445.136.1.82.

von Wright, Georg Henrik. 1974. *Causality and determinism*. New York: Columbia University Press. https://doi.org/10.7312/wrig90574.

Wright, Richard W. 1985. Causation in tort law. *California Law Review* 73(6). 1735–1828. https://doi.org/10.2307/3480373.

Wright, Richard W. 2013. The NESS account of natural causation: A response to criticisms. In Benedikt Kahmen & Markus Stepanians (eds.), *Critical essays on 'causation and responsibility'*, chap. 14, 285–322. Berlin: De Gruyter. https://doi.org/10.5040/9781472561022.ch-014.

Xiao, Richard & Tony McEnery. 2006. Collocation, semantic prosody, and near synonymy: A cross-linguistic perspective. *Applied linguistics* 27(1). 103–129. https://doi.org/10.1093/applin/ami045.

Lelia Glass
School of Modern Languages
Georgia Institute of Technology
Swann Building
613 Cherry Street NW
Atlanta, Georgia 30332
lelia.glass@modlangs.gatech.edu