

Modeling uncertainty, unawareness, and underspecification among Structural Causal Models¹

Todd SNIDER — *Eberhard Karls Universität Tübingen*

Michael FRANKE — *Eberhard Karls Universität Tübingen*

Abstract. We argue that linguistic communication of information about causal relations frequently involves epistemic states of not only *uncertainty* about, but also *unawareness* of causal facts. We propose a simple model of Causal Beliefs with Unawareness (CBU), which augments Structural Causal Models with additional structure to model both uncertainty and unawareness.

Keywords: causality, communication, structural causal models, uncertainty, awareness.

1. The Puzzle

1.1. Structural Causal Models

Structural Causal Models (SCMs) have been used for a variety of applications of interest to linguists and cognitive scientists, including to provide truth conditions for conditionals (a.o., Pearl, 2000; Hiddleston, 2005; Briggs, 2012), to explain causal selection behavior (Woodward, 2003), and to give semantics for causal verbs like *make* (Nadathur and Lauer, 2020).

All of these different applications, regardless of the packaging, crucially involve (causal) relations between the events/propositions² mentioned; to illustrate, each of the examples (1)–(3) involve a causal dependence ($A > C$) of Charlie’s being upset (C) on Alex’s buying a turtle (A).

- | | | |
|-----|---|------------------|
| (1) | If Alex buys a turtle, Charlie will be upset. | CONDITIONAL |
| (2) | Alex bought a turtle. That’s why Charlie is upset. | CAUSAL SELECTION |
| (3) | Alex’s buying a turtle would make Charlie upset. | CAUSAL VERB |

SCMs are useful across these applications because they represent just those sorts of dependencies, and the different ways they can be instantiated in a model.

Other ongoing work in cognitive science provides even more reasons to find SCMs useful for modeling agents’ behavior. Recent work suggests that causal intuitions arise from people performing mental simulations, modeled as sampling of possibilities based on SCMs (a.o., Gerstenberg, 2022; Quillien and Lucas, 2023). SCMs have also been used as ingredients in the pragmatic reasoning process by which agents interpret the meaning of both causal and non-causal language (Grusdt et al., 2022; Beller and Gerstenberg, 2023).

In short, SCMs are useful for a variety of applications that involve the modeling of causal dependencies. But they have their limitations, as we will explore shortly, which leads us here to aim to supplement them to increase their modeling potential.

¹We would like to thank the audience at Sinn und Bedeutung for useful comments and feedback.

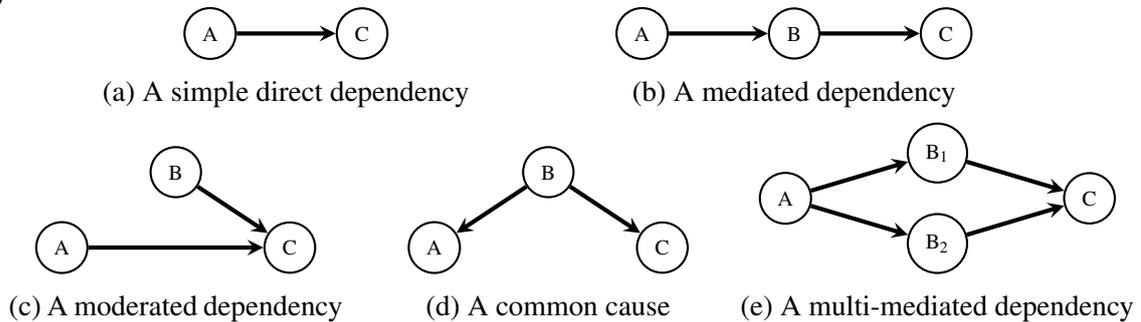
²For the purposes of this paper, it makes no difference whether the nodes of a Structural Causal Model represent events or propositions. The model we advance here is agnostic as to this choice, so we will not expound upon this distinction further. For simplicity, we discuss the model on the version where nodes represent propositions.

1.2. Uncertainty

Many early treatments that used SCMs for the interpretation of conditionals (e.g., Pearl 2000; Hiddleston 2005; Briggs 2012) presume that agents have a single mental model of the world: one SCM that they use to evaluate utterances containing conditionals. When an agent encounters a conditional, they: (i) identify the causal statement conveyed by the utterance, then (ii) perform interventions on their mental model, if necessary (e.g., to set a counterfactual antecedent to ‘true’), and finally (iii) evaluate the truth of the statement on the basis of the (modulated) model—if the SCM validates the statement, then the utterance is judged true, and if the SCM invalidates the statement, then it is judged false. Importantly, there is no representation in such treatments of an agent potentially being uncertain about which SCM is the right model for representing the facts and dependencies of the world.

There have been attempts, however, to incorporate representations of agent uncertainty in SCMs. Inspired by the observation that a single conditional like (1) is compatible with a number of possible explanations (each with a different SCM) like those schematized in (4), Bjorndahl and Snider 2016 incorporated SCMs into a Stalnakerian framework that tracks information gain via a dynamic set of ‘live possibilities’.

- (1) If Alex buys a turtle, Charlie will be upset. [A > C]
 (4)

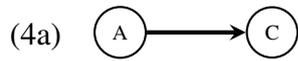


Adopting Starr 2013, each SCM can be treated as a possible world, fixing both the facts and dependencies at that world. In a Stalnakerian turn, then, Bjorndahl and Snider 2016 presents agents as maintaining uncertainty over a set of SCMs: the live candidate worlds in the Context Set (i.e., those worlds still compatible with the information in the Common Ground).

On this treatment, when an agent interprets an utterance containing a conditional, they: (i) identify the set of SCMs compatible with the utterance—on the basis both of values assigned to nodes and the presence/absence of certain dependencies among nodes—, and then if they accept the utterance as true, (ii) intersect that set with the prior Context Set, so that they can (iii) treat that intersection as the new Context Set, the conversational backdrop for subsequent conversational moves. The agent has uncertainty about which of the SCMs in the Context Set is the true representation of the actual world—all are equally possible candidates—but has ruled out of consideration any of the SCMs outside that set.

1.3. Unawareness and inattention

Even with uncertainty over models, there is more that we might want to be able to represent in an agent’s mental model of the world. Consider the schema for a simple direct dependency of C on A , as in (4a):



(4a) does not include a node B as part of the explanation of how C is causally dependent on A ($A > C$). Is this because the agent has an explicit belief that B is unrelated? Or does the agent have an only implicit belief that B is unrelated? In other words, are they not aware that B even exists as a node worth modeling? Put a third way, we could ask: are they not attending to the possibility that B exists? Certainly we can distinguish between the mental states of having a belief in $\neg\varphi$ vs. merely not having any beliefs about φ . But the SCM framework as it currently stands does not represent this distinction.

Awareness and attention are key factors of an agent’s reasoning and decision-making, and they matter for conversation, too (Ciardelli et al., 2009; Franke and de Jager, 2011; Westera, 2022); so, it is not enough to represent an agent’s beliefs, we must also represent their attention, and what things they are (un)aware of. But we can’t read an agent’s awareness or attention off of a SCM (nor from a set of SCMs, as per Bjorndahl and Snider 2016); this framework doesn’t allow us to represent the state of an agent’s awareness or attention.

Awareness and attention affect how agents behave in conversations in subtle ways.

- (5) If Alex buys a turtle, Charlie will be upset.
- a. Skeptical Sam: What do you mean, how so?
 - b. Naïve Nat: Oh, okay. Good to know.
 - c. Guessing Gal: Oh, because Charlie would be jealous?

A single conditional like (1), repeated here in discourse as (5), is compatible with a variety of different types of explanations (as in (4); see Bjorndahl and Snider 2016 for discussion), but agents also bring to bear their prior beliefs and expectations when interpreting (and thus, when responding to) utterances. An agent who has broad reasons to disbelieve or disprefer a direct A -to- C dependency might question the asserted $[A > C]$ covariance, and ask for more detail about how C depends on A ; this sort of agent is exemplified by Sam in (5a). On the other hand, an agent who isn’t even aware that there might (need to) be intervening or moderating factors between A and C —the sorts of factors labeled with B s in (4)—wouldn’t even think to ask follow-up questions about such factors; here this unaware sort of agent is exemplified by Nat in (5b). Alternatively, an agent who is aware of the presence and relevance of some additional factor B but who has uncertainty as to which B is the actual or communicatively-intended B might ask a directed question to try to decide among such candidates; this type of agent is exemplified by Gal in (5c).

At least broadly speaking, the behavior of agents like Sam, Nat, and Gal is driven by—or at least, bounded by—their states of awareness, i.e., what information they are attending to. They maintain uncertainty over SCMs regarding those issues they don’t have enough information to

resolve: either around the truth-values of specific nodes, or about the (non-)existence of edges between nodes. But their (un)awareness determines which SCMs they are ‘even considering’, or at least, which SCMs the agent can distinguish from one another; if an agent isn’t aware that they have uncertainty, they won’t make conversational moves to resolve that uncertainty. So an agent’s awareness plays a role in determining what information they can (and thus, may choose to) focus on, ask about, etc.

Accordingly, what we want is a way to model both the *information* and the *attention* state of an agent, as they consider the possibility space of SCMs which represent the world in order to interpret utterances, reason, and respond in discourse. In the next section, we propose a modification to standard SCM treatments which will allow for just that.

1.4. Related work

There is a strong concern in statistics, machine learning, and neighboring fields about *causal abstraction*, i.e., the question of how to define a formal notion of abstraction for causal models, where a coarser-grained representation of a causal process retains information about dependencies and the effects of interventions, even though the more fine-grained representation is compressed (e.g., Beckers and Halpern, 2019; Beckers et al., 2020). While directly related to our current concerns, the focus of that line of work is on models used as veridical representations, for example, for scientific explanation or applied predictions. In contrast, our concern for the current project is with the nature of human mental representations and, most importantly, highlighting crucial features of compressed or simplified mental representations that affect natural language use and conversation.

While we currently only consider the static case of representation of causal information in a single agent’s mental state, we see this work as being in alignment with prior work attending to dynamics of common ground between interlocutors that goes beyond the standard case of adding information, i.e., eliminating worlds from the set of possibilities under consideration (e.g., Swanson, 2006; Ciardelli et al., 2013; Klecha, 2018; Westera, 2022). The specific way in which the present work adds to this picture of expanding common ground by shifting attention or raising possibilities is by highlighting how this is relevant for a particularly structured and complex, but crucial kind of information: causal knowledge.

We take inspiration from prior work in formal epistemology, logic, and economics on agentive *unawareness*, and build on previous work employing unawareness models for linguistic analysis (de Jager, 2009; Franke, 2014). Formal models of agent awareness can be roughly classified as either syntactic or semantic approaches. The *Logic of General Awareness* (LGA; Fagin and Halpern, 1988) is originally a syntactic approach. The idea is that agents’ explicit representations are separated from their implicit (unaware) assumptions by a set of (propositional) formulas describing what the agents are aware of. In contrast, a prominent semantic approach uses *Subjective State Spaces* (SSS), first proposed for single agents by Modica and Rustichini 1999 and extended to multiple agents in subsequent work (Heifetz et al., 2008; Halpern and Rêgo, 2009). In this paper, we use basic ideas of the semantic approach (borrowing the idea

of subjective indistinguishability between world states), but dispense entirely with higher-order beliefs (so we don't actually give a full logic of unawareness of causal facts) to keep matters relatively simple.

Both of these major approaches to modeling awareness, LGA and SSS, require representations of atomic propositions which agents can be aware or unaware of. To apply either approach to representations of causal knowledge of the kind we consider here would therefore require a representational language to describe SCMs structurally. While there are logics for describing (finite) graphs in general (e.g., Ebbinghaus and Flum, 1995), these are usually more complex languages than propositional logic (e.g., modal or first-order logics). The formal languages developed for reasoning about Structural Equation Models are also quite complex, being a form of dynamic logic, which naturally capture the dynamic nature of interventions as a form of action (e.g., Halpern, 2000). Some propositional, yet powerful representational languages for causal knowledge (e.g., McCain and Turner, 1997) purposefully omit explicitly representing the kind of information about (direct) causation that we would need in the present context. In conclusion, for our current modest purposes, we avoid excessive formal machinery, as we chiefly want to argue that concern about minimal cognitive representations for causal talk is called for, however we might eventually formally model them.

2. The Causal Beliefs with Unawareness (CBU) Model

2.1. Setup

We start with the setup as in Bjorndahl and Snider 2016, using Starr's (2013) Structured Possible Worlds to incorporate SCMs into a Stalnakerian picture of conversational information gain. On that picture, a possible world $w \in W$ is a fully-specified SCM: a directed acyclic graph with nodes for all of the atomic propositional variables being tracked by the theorist; or equivalently, a world is a function from exogenous variables to $\{0, 1\}$ and from endogenous variables to a dependence function, which in turn maps situations (i.e., a node's parents' values) to $\{0, 1\}$.³ Our version of this will involve a slight departure, for reasons we will return to in §2.4, but the core ideas are the same: each possible world tracks both the assignments of values to the propositional variables being modeled, and the dependencies among those variables.

As is standard, we'll take propositions to be sets of worlds, and we'll represent an agent's beliefs via the set of worlds compatible with those beliefs: the agent's belief-set $\beta \subseteq W$. (An agent's desires, and other similar differentiable propositional attitudes may be represented by parallel sets; agents with inconsistent beliefs may find themselves painted into undesirable corners, requiring belief revision or other escape/repair mechanisms.) We'll say that an agent believes some proposition $\varphi \subseteq W$ ($B\varphi$), just in case all of the worlds in the agent's belief-set make that proposition true: $\beta \subseteq \varphi$.

³Endogenous variables are those whose nodes have incoming edges, i.e., whose parent nodes are represented in the graph, and thus, whose truth-values can be determined strictly within the model. Exogenous variables are represented graphically as nodes with no incoming edges; having no parents causally upstream, their truth-values are assigned by the modeler (hence, external to the model itself).

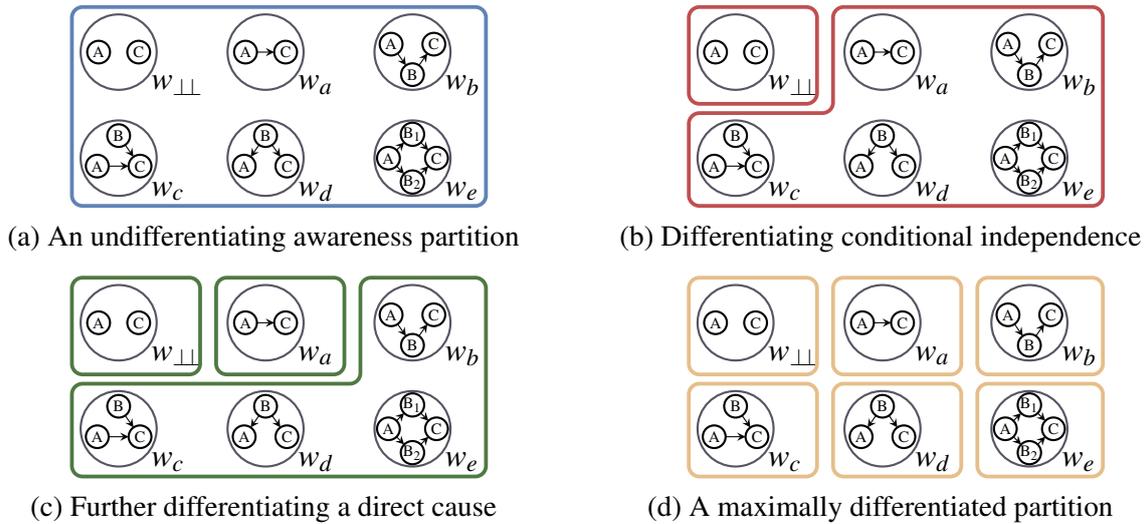


Figure 2: Different possible awareness partitions, each with six possible worlds (of which only the SCM component is shown; assignment of truth-values to variables omitted for readability).

2.2. Adding unawareness

We'll model awareness by using an equivalence relation \sim on W , which induces the partition \mathfrak{X} : our **awareness partition**. Intuitively, \sim and \mathfrak{X} together represent the indistinguishability of worlds: if $w_1 \sim w_2$, then the agent cannot tell w_1 and w_2 apart (or in other words, the agent is unaware of any difference between w_1 and w_2).

This can be illustrated as in Figure 2. As per our setup, each world in our universe of possible worlds is an SCM; here, considering the shapes of possible explanations for (1), each world is one of the explanations from (4), alongside the world $w_{\perp\perp}$ where A and C are causally independent from one another (and so, no edge or path of edges connects the nodes A and C). In such diagrams, our awareness partition \mathfrak{X} is demarcated by the colored boundaries surrounding each cell. Crucially, the worlds within a single cell are (all) related with the equivalence relation \sim . So in Figure 2a for instance, where all worlds are in the same cell, the agent cannot differentiate $w_{\perp\perp}$ from w_e : this (minimal) awareness partition treats all worlds as indistinguishable from one another. In Figure 2b, the agent is represented as being able to distinguish $w_{\perp\perp}$ from the other worlds (where A and C are not conditionally independent), but unable to distinguish w_a from w_b from w_c , and so forth. And only in the maximally differentiated awareness partition illustrated in Figure 2d, where no two worlds share the same cell, are no two worlds related by the equivalence relation \sim ; only under such an awareness partition can the agent distinguish every world individually.

2.3. Explicit and implicit beliefs

Enriching SCMs with this new tool allows us to better model an agent's beliefs relative to their awareness, which in turn allows us to more precisely talk about an agent's beliefs as being explicit or implicit in the CBU model.

As background, let us call two partitions Q_1 and Q_2 **orthogonal** (in the sense of Lewis 1988) if and only if for all $X \in Q_1$ and $Y \in Q_2$, if $X \neq \emptyset$ and $Y \neq \emptyset$, then $X \cap Y = \emptyset$. Then, we'll say that an agent is **unaware of a proposition** φ just in case the partition of W induced by the issue of φ (i.e., the set $\{\varphi, \bar{\varphi}\}$) is orthogonal to \aleph (their prior awareness partition). In other words: if differentiating on the basis of φ subdivides every cell of the prior partition \aleph , then the agent wasn't aware of φ as a factor along which to distinguish the space of possible worlds beforehand; and on the other hand, if any two worlds were already differentiated only by φ , then the agent must already have been aware of φ .

With this in hand, we can say: An agent has an **explicit belief** in φ if and only if they believe φ and are aware of φ ; an agent has an **implicit belief** in φ if and only if they believe φ and are **not** aware of φ .

2.4. (Un)Awareness of different sorts and underspecification

This new awareness partition captures an agent's (in)attention to the distinctions among worlds, where those distinctions might fall along any number of different lines. An agent might not attend to the identity or details of nodes which are known to them (i.e., which they are already modeling): for example, an agent might attend to the possibility of Alex buying a turtle, but might not bother to differentiate among species or individuals being bought (though those are theoretically differentiable possibilities). The same is true for edges, representing the dependencies that may or may obtain between nodes.

The partition also captures what Franke and de Jager 2011 calls the "filtering" of "unmentionables": collapsing over propositional variables an agent is not even aware enough of to include in their model. This includes variables that an agent simply doesn't know about, as well as variables collapsed to a single node, i.e., complexes treated as simplex. For example, one's model might include a node which represents 'the alarm clock goes off'; in the right conversational context, if one's attention were drawn to it, that representation might be 'blown up', expanded to account for the internal mechanisms of the clock whose causal structure might be important. Until such time, though, the nodes we might assign to those internal mechanisms are 'collapsed' into one, and the whole complex is treated as causally simplex. Both such glossed-over complexities, and issues that an agent hasn't even thought to consider, are 'filtered' out in our model insofar as they are treated as indistinguishable to the agent, equivocated within one cell of the awareness partition.

In this sense, and because we are interested in using SCMs to model agents' subjective mental states (insofar as they influence causal reasoning and communicative behavior, including linguistic behavior) rather than the objective mechanics of the world, it is natural to think of each SCM in an illustration like Figure 2 as an only *partial* representation, one which is **underspecified** for the values of variables not modeled explicitly within it. For some conversational purposes (at certain coarse levels of detail), it may be enough to differentiate, for example, worlds which have the shape of a common cause explanation (as in (4d)) for the conditional in (1) from worlds which have the shape of a moderated dependency (as in (4c)). But at other

times, given other conversational purposes (which demand more fine-grained levels of detail, for instance), we might need to ‘expand’ that model: to make differentiations at scales we weren’t previously attending to, by representing propositional variable nodes we weren’t previously including in our models. Rather than treating each SCM in an illustration like Figure 2 as a fully-specified possible world—one which fully settles the issue, for every possible atomic propositional variable, of its truth-value and its dependencies on other variables—, we can treat these ‘small worlds’ as subjective world models, relative to the awareness of the agent and their current representational needs (informed by the goals and norms of the current discourse, at least). These subjective world models have the same function as proper worlds in this sort of causal framework: they represent propositional variables, their truth-values, and the dependencies among them. Only now, instead of treating each diagram as a discrete fully-specified world, we can take them to be representations of those variables which have risen to salience, remaining underspecified with regard to those variables which are not modeled. In a sense, each ‘small world’ model stands in for a ‘family’ of SCMs, the set of models that it could be extended to, without contradiction and without belief revision, namely, those which match all of the nodes and edges which are explicitly represented in the ‘small world’, but which vary along the lines about which it is underspecified, i.e., in terms of the issues not reflected in the current representation at its current level of detail. One can think of a subjective world model as akin to a model of a **situation**, in the sense of situation semantics (Barwise and Perry, 1981; Kratzer, 1989 and much subsequent work).

Treating these SCMs as underspecified reflects the attention and awareness of the agents we’re modeling, and does so in a minimal way, which is useful both conceptually and logistically. As omniscient theorists, we might choose to model at the highest possible level of detail, dealing only with fully-specified possible worlds, each with their zillions of nodes and arrangements of edges among those nodes. The unwieldiness of that approach, though, leads us to instead prefer to treat SCMs in CBU as specified along certain (salient, conversationally relevant) criteria and underspecified along others. Such representations bring us to closer to what individual agents are dealing with: models which are subjective and relative to their state of awareness. Treating these representations as underspecified also makes this approach more aligned with contemporary theories of cognitive psychology, where decades of evidence for finiteness of cognitive resources leads to preferring economical theories over those that rely on infinite processing (Simon, 1955; Tversky and Kahneman, 1974; see Lieder and Griffiths, 2020 for an overview); we will return to this conceptual alignment in §4.1.

3. Applying the CBU Model

In order to demonstrate the usefulness of this expanded CBU model, let us return to our prototypical agents from (5), each of whom respond differently to the assertion of (1), even while they each accept the conditional $[A > C]$ as true.

- (5) If Alex buys a turtle, Charlie will be upset.
 - a. Skeptical Sam: What do you mean, how so?

Skeptical Sam is the sort of agent who is willing to believe that there is *some* sort of dependency between A and C , but seems (i) to have a reason to reject the simple direct dependency of (4a),

and (ii) to know that there are potentially multiple explanations among which to select—hence their request for precisification. In other words, Skeptical Sam is *aware* of their uncertainty.

In order to capture the intuition (i), we can represent Sam’s belief-set as in Figure 3a. On this particular schematization, Sam doesn’t have any (implicit or explicit) beliefs that privilege any of the non-simple models over the others. This is an arbitrary decision on our part; the same skeptical behavior could be compatible with a belief-state which preferred an intermediate cause explanation (thus including w_b and w_e but excluding w_c and w_d), or one which dispreferred a common cause explanation (excluding only w_d). What is crucial in reflecting intuition (i) is only that the models schematized by $w_{\perp\perp}$ and w_a are not included within Sam’s belief-set: $w_{\perp\perp}$ because they are excluded by Sam’s accepting $[A > C]$, and w_a because Sam seems to be unwilling to accept the simplest explanation.

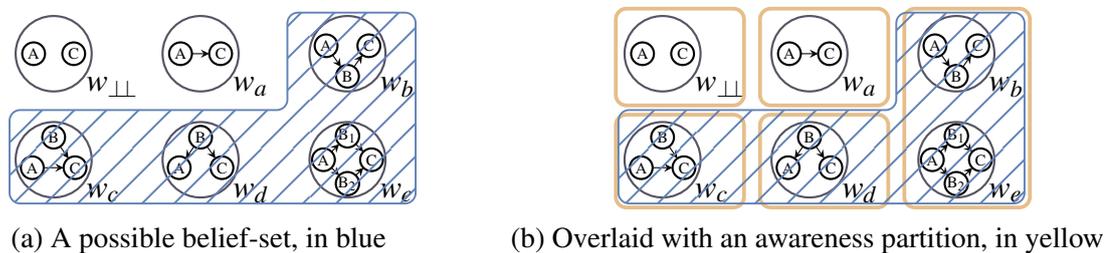


Figure 3: Skeptical Sam’s belief-set and awareness partition

We capture the intuition (ii), meanwhile, via the awareness partition as in Figure 3b. That Sam’s belief-set crosses multiple cells of their awareness partition \mathfrak{X} reflects the state of Sam being able to differentiate among multiple possibilities for the true state of affairs. Sam can tell the difference between the situations schematized as w_c , w_d , and w_e , for instance, but cannot decide among them (because all of those situations are compatible with their beliefs). It is precisely this awareness of uncertainty which leads Skeptical Sam to ask for more information, to help to resolve this manifest uncertainty.⁴

One final aside about Skeptical Sam’s mental state which is made apparent by Figure 3: we might normally associate skepticism with ‘unwillingness to believe’, but on this approach we see that being skeptical doesn’t require having a very restricted belief-set. As illustrated in our particular choice of belief-set in Figure 3a, this version of Skeptical Sam remains open to the possible situations represented by w_b through w_e . What makes Sam skeptical is their unwillingness to accept the simplest explanation, along with not ruling out other possibilities lightly. Other than not accepting the simple w_a , our version of Skeptical Sam is quite open-minded in fact, in terms of which types of explanations are compatible with their beliefs. Our Sam is skeptical, but not narrow-minded.

We can now turn our attention to Naïve Nat, who accepts the assertion of $[A > C]$ without further challenge. We want our model to reflect both (i) that Nat’s belief state should reflect

⁴It is not the case that all uncertainty should necessarily lead to requests for more information, as there is always uncertainty until an agent achieves perfect knowledge of the world (omniscience). Crucially, it is only uncertainty that an agent is aware of, i.e., attending to, which can potentially lead to their deciding to take an action in order to reduce that uncertainty.

this acceptance, and (ii) that they act as though there is no pressing uncertainty to resolve.⁵

- (5) If Alex buys a turtle, Charlie will be upset.
 - b. Naïve Nat: Oh, okay. Good to know.

We can reflect Nat’s acceptance of (1) by ensuring that their belief-set excludes $w_{\perp\perp}$; one way this might look is as illustrated in Figure 4a. In coming to believe $[A > C]$ explicitly, Nat must also (come to) differentiate $w_{\perp\perp}$ from the other types of situations schematized in this toy universe of possible worlds, in order to rule out those worlds where A and C are conditionally independent. So Nat’s awareness partition \aleph must minimally differentiate $w_{\perp\perp}$ from the other situations, as in Figure 4b.

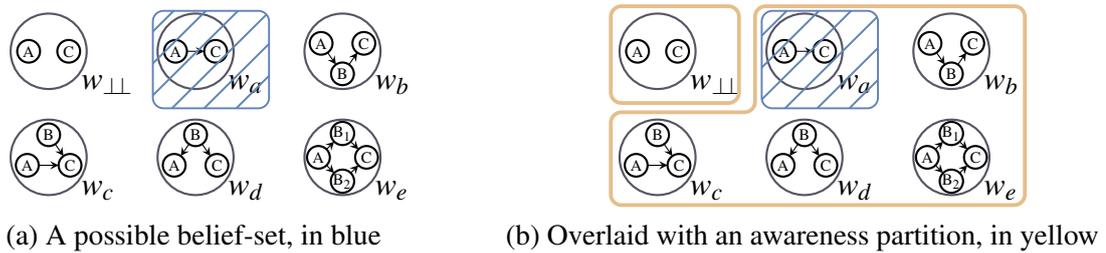


Figure 4: Naïve Nat’s belief-set and awareness partition

As with our representation of Skeptical Sam in Figure 3, there is some arbitrariness in this particular illustration: one could represent an equally-naïve agent with a broader belief-set—just so long as it excludes $w_{\perp\perp}$ and includes at least one situation compatible with $[A > C]$ —or with a more fine-grained awareness partition \aleph than the one depicted in Figure 4b. What is crucial for our characterization, though, is that Nat’s belief-set not extend across multiple cells of \aleph . (Figure 4b’s narrow belief-set and minimal \aleph is merely one way to guarantee that.) Because the subjective worlds of Nat’s belief-set are properly contained within a single cell of \aleph , Nat is unaware of any distinctions to be made among the situations compatible with their beliefs. As such, they are unaware of any uncertainty, and so have no motivation to seek more information in the way that Skeptical Sam or Guessing Gal do. The particular version of Naïve Nat as depicted in Figure 4 has an **implicit belief** in simplicity: they do not currently differentiate between the simple direct dependency situation w_a and the more complex situations, but if this distinction were brought to their awareness, they would then reject any models involving an additional causally relevant node B. But even a more open-minded agent—even one with the same broad belief-set as in Sam’s Figure 3a—would still plausibly behave in the same unquestioning way as Naïve Nat, so long as their awareness partition did not cross-cut that belief-set in any way. They would still be unaware of any uncertainty to resolve, and so would still not be likely to ask follow-up questions along the lines we’ve discussed here. This demonstrates the utility of extending SCMs with awareness partitions: belief-sets alone cannot explain the differing behavior of Skeptical Sam and Naïve Nat.

Finally, our third prototype, Guessing Gal, like Skeptical Sam seems to be aware of some uncertainty to resolve, as evidenced by their making a discourse move to obtain more information.

⁵As hinted at in the previous footnote, there could of course be other reasons for an agent with uncertainty to choose not to raise that as a conversational topic, e.g., for politeness reasons, or to maintain a particular social persona. Accounting for strategic discourse move planning is beyond the scope of the current project; for us it is sufficient to demonstrate how a lack of uncertainty could motivate the sort of behavior exemplified by Naïve Nat.

But Gal seems to have less uncertainty than Sam, at least enough so that they are willing to wager a guess among those differentiable possible explanations for (1).

- (5) If Alex buys a turtle, Charlie will be upset.
 - c. Guessing Gal: Oh, because Charlie would be jealous?

As with our earlier illustrations, we have some leeway in how we might represent the belief-state and awareness partition of an agent who behaves in the way that Gal does in (5), but not total freedom; the choice is not entirely arbitrary. To capture the fact that Gal is aware of some uncertainty to resolve, it must be the case that their belief-set encompasses more than one cell from their awareness partition \mathfrak{K} . And to reflect the intuition that Gal has less uncertainty than Sam, we might want there to be fewer cells of \mathfrak{K} within Gal’s belief-set than in Sam’s (either due to Gal having a narrower set of beliefs or a coarser differentiation in \mathfrak{K}); here we exemplify this version of Guessing Gal with a belief-set as in Figure 5a combined with an awareness partition \mathfrak{K} as in Figure 5b.

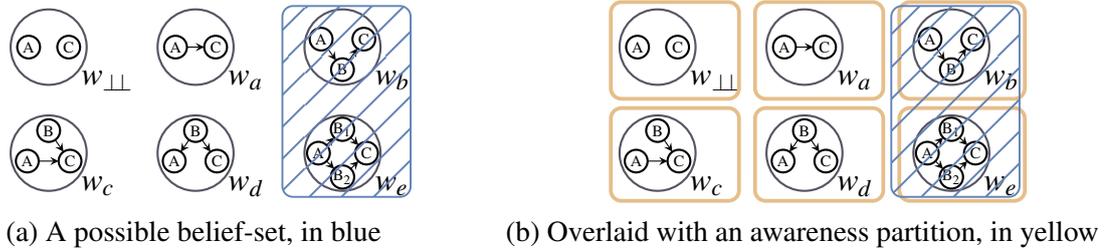


Figure 5: Guessing Gal’s belief-set and awareness partition

This representation of Guessing Gal happens to have a more finely differentiated \mathfrak{K} than Skeptical Sam in Figure 3. But Gal nevertheless has fewer cells of \mathfrak{K} contained within their belief-set due to that belief-set being narrower. There being fewer cells of \mathfrak{K} —possibilities which the agent is differentiating among—within Gal’s belief-set compared to Sam’s explains their willingness to make a guess. For Gal, there are only two⁶ arguably-similar options for scenarios which explain (1). Meanwhile for the Sam represented in Figure 3, there are three such options (which would be four, with Gal’s more fine-grained \mathfrak{K}) reflecting a wider range of causal scenarios, which results in Sam having more uncertainty over the possible explanations for the $[A > C]$ covariance.

As is evident from Figure 5b, this version of Gal has an **explicit belief** that there is some factor B which mediates between A and C —explicit because they are already differentiating between those situations which do and don’t contain an intermediate B —, but maintains some uncertainty about the nature of that mediating factor, and guesses among those situations.

⁶At the current level of detail, Gal is not represented as distinguishing among potentially different intermediate B nodes. Of course, ‘zooming in’ to expand the situation w_b to differentiate among different B s would increase the total number of possibilities; see the discussion of underspecification in §2.4. At this level of detail, though, we still reflect Gal as having relatively less uncertainty than Sam.

4. Discussion

4.1. Minimal representations

One reason to prefer a framework like this, treating Structural Causal Models as underspecified schemas representing subjective world models, is that it keeps our representations tractable. We can model an agent as having a certain stance relative to a specific world (or set of worlds), like belief, without having to model that agent as having a fully-specified mental representation of that world (or set of worlds) and every atomic proposition. This is especially useful in modeling implicit beliefs, those beliefs which agents hold without being aware of them, insofar as having no representation for φ provides an easy way to track an agent's lack of awareness of φ .

This attention to cognitively efficient representations is in line with much like-minded work from cognitive psychology and cognitive science, which emphasizes that not only does the computational cost of processing matter, but so does the cost of mentally representing information. Examples include mental model theories of reasoning (Johnson-Laird, 1986), or work on (probabilistic) Language of Thought (Quilty-Dunn et al., 2022). Or, to take an example of a connection with philosophy, Swanson 2010 argues that the interpretation of causal talk is guided by conversational norms regarding what counts as “good representatives” for causal paths. We aim for our model to offer a formal scaffolding for expressing these sorts of ideas, as we take economical representations—those that attend to relevance, sufficiency, and cost—to be a crucial component in the pragmatics of causal talk. For example, our CBU model allows us to evaluate and compare SCMs in terms of how many nodes they represent and in terms of how they relate to an agent's awareness, two factors that might play a role in judgments about which representations are good, cooperative ones to bring to salience via conversational moves in causal talk.

4.2. Awareness and reasoning

Existing research has established that an agent's state of awareness plays a role in their behavior. Within the domain of causality, Fernbach and Darlow 2010 has illustrated that awareness of alternatives affects how people make judgments about causal consequences, and Grusdt et al. 2022 has shown that what the set of relevant alternatives are determined to be influences pragmatic reasoning about conditionals. Having the ability to model an agent's awareness, and to differentiate between their explicit and implicit beliefs, allows us to better describe—and to better make predictions about—their behavior in these sorts of causal talk and causal prediction paradigms.

4.3. Ease of acceptability

Differentiating between an agent's beliefs on the one hand and attention (or awareness) on the other in the way that our model does also suggests a cline of sorts, an ordering of conversational moves in terms of the **ease of acceptability** that they have in a given context. Consider, as one

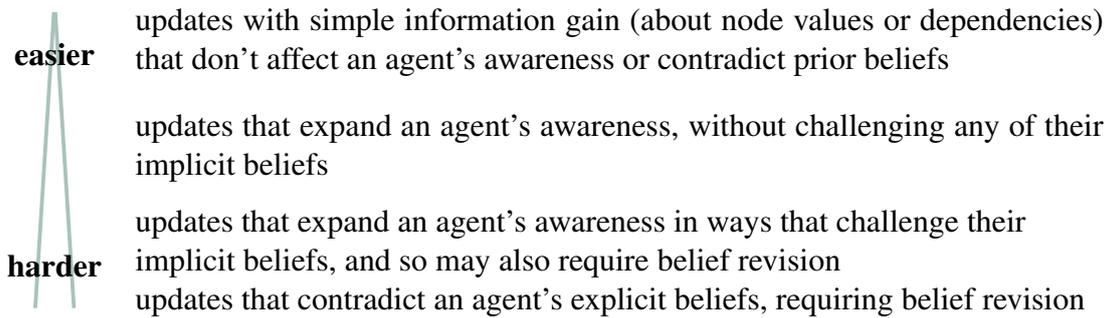


Figure 6: A cline representing the ease of acceptability of various conversational moves

extreme, that some conversational moves would trigger updates involving only information gain, without expanding an agent's awareness or contradicting any of their prior beliefs; these are the sorts of updates that an agent might be most likely to accept without challenge or negotiation. At the other end of the spectrum, of course, would be updates that contradict an agent's explicit beliefs, thus requiring belief revision—a process which is notoriously tricky for theorists to describe formally (Alchourrón et al., 1985 and much subsequent work; see Hansson, 2022 and citations therein), and perhaps even harder for rhetoricians to trigger in their audiences. This other extreme of the cline represents those moves which agents are least likely to accept, as giving up one's previously held beliefs involves not only significant cognitive costs,⁷ but also potentially social costs (in terms of group membership, internal and external identity construction, and more; see Braver et al., 1977; Tavis and Aronson, 2007; Fetterman et al., 2019, among many others). We might plot these extremes in a diagram like Figure 6, with the top representing those conversational moves whose updates are easiest for interpreters to accept, insofar as they carry relatively smaller cognitive costs to incorporate with their prior mental state, and with those conversational moves which are costlier, harder to accept, towards the bottom.

In between these extremes, though, we can consider the other sorts of updates, in terms of both an agent's beliefs and their awareness, that might be triggered by various conversational moves. Updates which expand an agent's awareness, impelling them to adopt a more fine-grained awareness partition, but do so without challenging any of their implicit beliefs, are likely to occupy a medial position on this cline: more involved than a simple information gain update (i.e., requiring more cognitive resources, perhaps taking more time), but less so than updates which trigger a belief revision process. This model also allows us to distinguish those conversational moves whose updates would expand an agent's awareness in ways which challenge their implicit beliefs; whether the belief revision process for implicit beliefs is the same (procedurally, and in terms of cognitive effort) as that for explicit beliefs is an open question—and one that cannot be asked without a model that differentiates implicit and explicit beliefs. It remains for future research to determine whether there are measurable differences in the activity patterns of agents performing these differentiable kinds of updates.

⁷Expanding one's set of possibilities to 'rule back in' things which were already ruled out may potentially open the proverbial floodgates, letting back in all sorts of previously-thought-absurd possibilities; in other words, leading the agent to question, 'If I don't believe this anymore, what *do* I believe?'

4.4. Moving forward

Adding an agent's (un)awareness and (in)attention to a Bjorndahl and Snider 2016-style framework with Structural Causal Models enriches our representation and allows us to distinguish implicit and explicit beliefs, to explain how different prior mental states can lead to different kinds of responses (as exemplified by our prototypical Sam, Nat, & Gal), and to do so in ways we take to be more congruous with current theories of cognitive psychology. That said, this is just one step in what we hope to be the right direction, and this work invites a number of extensions for future exploration.

First, the current framework as laid out is static, and represents an agent's beliefs and awareness at a single moment in a discourse. We can model what an agent's belief-set and awareness partition are, both before and after some discourse move, but we have not yet developed a theory formalizing how individual discourse moves induce specific changes on beliefs or awareness; such a dynamic version of this model is a natural extension of this project, and would lead to a more comprehensive picture of the discourse dynamics of belief and awareness, as agents not only learn new information but also shift their attention, entertain possibilities, revise their implicit and explicit beliefs, and grapple with uncertainty.

Second, the framework as described here only represents the beliefs and awareness of a single agent. An eventual hope for this project, however, would be to extend this system such that we could model the beliefs and awareness of multiple agents simultaneously, in order to capture the conversational dynamics of multiple agents producing and interpreting messages, triggering and performing updates in one interactive scenario. Such an extension to this framework would in theory be able to model joint attention, as agents coordinate in treating certain entities and abstract objects as salient in a discourse—ideally without requiring infinitely-nested mental representations (e.g., awareness of one's awareness of one's awareness, etc.). Tracking the attention and awareness of multiple agents simultaneously would also allow for closer investigation of scenarios in which discrepancies among agents emerge, where agents disagree about what is mutually known or mutually attended to, which would likely be reflected in how subsequent discourse unfolds, perhaps requiring repair or reconciliation techniques.

Third, as described in §4.3, this framework suggests a cline along which conversational moves might be positioned, in terms of the cognitive costs associated with performing the informational/attentional updates they trigger. Along with the proper linking hypotheses, connecting this abstract notion of 'cost' to specific behavioral measures such as reaction times or event-related potentials, this project invites further experimental investigation into the conversational moves at different points along this cline. Can we measure the cognitive effort associated with expanding one's awareness, or that of belief revision? And is there a difference in the 'costs' of revising implicit beliefs as opposed to explicit ones?

In sum, as discussed in §4.1, this project has proceeded with the intention of aligning with work from neighboring fields like cognitive psychology, cognitive science, and philosophy, work which attends to concerns surrounding efficient and minimal representations. Future work could bring these traditions even more closely together, using our framework to formalize

concepts and theories described informally in such fields, or elaborating our model further with additional tools inspired by insights from such fields.

References

- Alchourrón, C. E., P. Gärdenfors, and D. Makinson (1985). On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic* 50(2), 510–530.
- Barwise, J. and J. Perry (1981). *Situations and Attitudes*. The MIT Press.
- Beckers, S., F. Eberhardt, and J. Y. Halpern (2020). Approximate causal abstraction. In *Proceedings of the 35th Uncertainty in Artificial Intelligence Conference*, Volume 115, pp. 606–615.
- Beckers, S. and J. Y. Halpern (2019). Abstracting causal models. *Proceedings of the AAAI Conference on Artificial Intelligence* 33(1), 2678–2685.
- Beller, A. and T. Gerstenberg (2023). A counterfactual simulation model of causal language. PsyArXiv Preprint.
- Bjorndahl, A. and T. Snider (2016). Informative counterfactuals. In S. D’Antonio, M. Moroney, and C. R. Little (Eds.), *Semantics and Linguistic Theory (SALT)*, Volume 25, pp. 1–17. LSA and CLC Publications.
- Braver, S. L., D. E. Linder, T. T. Corwin, and R. B. Cialdini (1977). Some conditions that affect admissions of attitude change. *Journal of Experimental Social Psychology* 13(6), 565–576.
- Briggs, R. (2012). Interventionist counterfactuals. *Philosophical Studies* 160(1), 139–166.
- Ciardelli, I., J. Groenendijk, and F. Roelofsen (2009). Attention! ‘might’ in inquisitive semantics. In E. Cormany, S. Ito, and D. Lutz (Eds.), *Semantics and Linguistic Theory (SALT)*, Volume 19, pp. 91–108. CLC Publications.
- Ciardelli, I., J. Groenendijk, and F. Roelofsen (2013). Inquisitive semantics: A new notion of semantic meaning. *Language and Linguistics Compass* 7(9), 459–476.
- Ebbinghaus, H.-D. and J. Flum (1995). *Finite Model Theory*. Heidelberg: Springer.
- Fagin, R. and J. Y. Halpern (1988). Belief, awareness and limited reasoning. *Artificial Intelligence* 34(1), 39–76.
- Fernbach, P. and A. Darlow (2010). Causal conditional reasoning and conditional likelihood. In S. Ohlsson and R. Catrambone (Eds.), *Cognitive Science Society*, Volume 32, pp. 1088–1093.
- Fetterman, A. K., S. Curtis, J. Carre, and K. Sassenberg (2019). On the willingness to admit wrongness: Validation of a new measure and an exploration of its correlates. *Personality and Individual Differences* 138, 193–202.
- Franke, M. (2014). Pragmatic reasoning about unawareness. *Erkenntnis* 79(4), 729–767.
- Franke, M. and T. de Jager (2011). Now that you mention it: Awareness dynamics in discourse and decisions. In A. Benz, C. Ebert, G. Jäger, and R. Rooij (Eds.), *Language, Games, and Evolution: Trends in Current Research on Language and Game Theory*, pp. 60–91. Springer.
- Gerstenberg, T. (2022). What would have happened? Counterfactuals, hypotheticals and causal judgements. *Philosophical Transactions of the Royal Society B: Biological Sciences* 377(1866).
- Grusdt, B., D. Lassiter, and M. Franke (2022, 10). Probabilistic modeling of rational communication with conditionals. *Semantics and Pragmatics* 15(13).
- Halpern, J. Y. (2000). Axiomatizing causal reasoning. *Journal of Artificial Intelligence Re-*

- search* 12, 317—337.
- Halpern, J. Y. and L. C. Rêgo (2009). Reasoning about knowledge of unawareness. *Games and Economic Behavior* 67(2), 503–525.
- Hansson, S. O. (2022). Logic of Belief Revision. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2022 ed.). Metaphysics Research Lab, Stanford University.
- Heifetz, A., M. Meier, and B. C. Schipper (2008). A canonical model for interactive unawareness. *Games and Economic Behavior* 62(1), 304–324.
- Hiddleston, E. (2005). A causal theory of counterfactuals. *Noûs* 39(4), 632–657.
- de Jager, T. (2009). ‘Now that you mention it I wonder...’: Awareness, Attention, Assumption. Ph. D. thesis, Universiteit van Amsterdam.
- Johnson-Laird, P. N. (1986). Conditionals and mental models. In E. C. Traugott, A. ter Meulen, J. S. Reilly, and C. A. Ferguson (Eds.), *On Conditionals*, pp. 55–75. Cambridge University Press.
- Klecha, P. (2018). On unidirectionality in precisification. *Linguistics and Philosophy* 41, 87–124.
- Kratzer, A. (1989). An investigation of the lumps of thought. *Linguistics and Philosophy* 12, 607–653.
- Lewis, D. (1988). Relevant implication. *Theoria* 54(3), 161–174.
- Lieder, F. and T. L. Griffiths (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences* 43, 1–60.
- McCain, N. and H. Turner (1997). Causal theories of action and change. In *Proceedings of the Fourteenth National Conference on Artificial Intelligence and Ninth Conference on Innovative Applications of Artificial Intelligence*, pp. 460—465. AAAI Press.
- Modica, S. and A. Rustichini (1999). Unawareness and partitioned information structures. *Games and Economic Behavior* 27(2), 265–298.
- Nadathur, P. and S. Lauer (2020, June). Causal necessity, causal sufficiency, and the implications of causative verbs. *Glossa: A Journal of General Linguistics* 5(1).
- Pearl, J. (2000). *Causality: Models, Reasoning and Inference*. Cambridge University Press.
- Quillien, T. and C. G. Lucas (2023). Counterfactuals and the logic of causal selection. *Psychological Review*.
- Quilty-Dunn, J., N. Porot, and E. Mandelbaum (2022). The best game in town: The re-emergence of the language of thought hypothesis across the cognitive sciences. *Behavioral and Brain Sciences*, 1—55.
- Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics*, 99–118.
- Stalnaker, R. (1978). Assertion. *Syntax and Semantics* 9, 315–332.
- Starr, W. (2013). Structured possible worlds. Ms. Cornell University.
- Swanson, E. (2006). *Interactions with Context*. Ph. D. thesis, Massachusetts Institute of Technology.
- Swanson, E. (2010). Lessons from the context sensitivity of causal talk. *The Journal of Philosophy* 107(5), 221–242.
- Tavris, C. and E. Aronson (2007). *Mistakes Were Made (but Not by Me): Why We Justify Foolish Beliefs, Bad Decisions, and Hurtful Acts*. Houghton Mifflin Harcourt.
- Tversky, A. and D. Kahneman (1974). Judgment under uncertainty: Heuristics and biases.

Modeling uncertainty, unawareness, and underspecification among Structural Causal Models

Science 185(4157), 1124–1131.

Westera, M. (2022, September). Attentional pragmatics: A pragmatic approach to exhaustivity. *Semantics and Pragmatics* 15(10), 1–51.

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press.