Building Compressed Causal Models of the World

David Kinney[1] and Tania Lombrozo[2]

[1]Department of Philosophy and Program in Philosophy-Neuroscience Psychology,

Washington University in St. Louis

[2]Department of Psychology, Princeton University

Forthcoming in Cognitive Psychology

Author Note

Abstract

A given causal system can be represented in a variety of ways. How do agents determine which variables to include in their causal representations, and at what level of granularity? Using techniques from Bayesian networks, information theory, and decision theory, we develop a formal theory according to which causal representations reflect a trade-off between compression and informativeness, where the optimal trade-off depends on the decision-theoretic value of information for a given agent in a given context. This theory predicts that, all else being equal, agents prefer causal models that are as compressed as possible. When compression is associated with information loss, however, all else is not equal, and our theory predicts that agents will favor compressed models only when the information they sacrifice is not informative with respect to the agent's anticipated decisions. We then show, across six studies reported here (N=2,364) and one study reported in the supplemental materials (N=182), that participants' preferences over causal models are in keeping with the predictions of our theory. Our theory offers a unification of different dimensions of causal evaluation identified within the philosophy of science (proportionality and stability), and contributes to a more general picture of human cognition according to which the capacity to create compressed (causal) representations plays a central role.

Keywords: Bayesian networks, compression, causal models, proportionality, stability, value of information.

Introduction

Scientists often aim to produce causal models of the world that balance informativeness with compression. That is, they aim to model data-generating processes in a way that captures as much information about those processes as possible, while omitting cumbersome or unnecessary details. For example, epidemiologists might produce a model of cancer rates in a population that treats smoking as a binary variable representing whether or not a person smokes cigarettes, but without specifying the average number of cigarettes the person smokes per day, and omitting additional background variables such as the person's blood type. Ordinary agents face an analogous challenge: in representing the social and physical world around us, each of us must determine which variables to include in our causal models, and at what level of granularity. For example, a causal model of a toddler's tantrums could include whether they napped or not as a binary variable, or a finer-grained specification of the number of minutes they napped; it could include the time of day, or omit this variable entirely. Any such choice of variables instantiates a particular trade-off between informativeness and compression. How do people navigate this choice in building causal models of the world?

In this paper, we begin from the premise that ordinary agents, like scientists, build causal models of the world, and that these causal models can be represented formally as *Bayesian networks* that support causal claims, causal explanations, predictions, and interventions (Gopnik and Tenenbaum, 2007; Griffiths et al., 2008; Pacer et al., 2013; Pearl, 2000; Spirtes et al., 2000). We then argue that these ordinary agents, like scientists, face a crucial problem: choosing which variables to include in their causal models. Using the Bayes net formalism and a decision-theoretic framework, we provide a formal theory of how to choose variables for a causal model so as to

achieve an optimal compression of the environment. We then corroborate our formal framework over the course of six experiments.

A Bayes Net Perspective on Variable Choice

In a Bayesian network, types of events are represented by *random variables*. These random variables are then related to each other by functions that represent the causal relationships between types of events. For instance, in our epidemiological example, a binary variable representing whether or not a person gets cancer will have its value determined by a function that takes as one of its arguments the value of a binary variable denoting whether or not someone smokes.

By definition, each random variable in a Bayesian network can also be represented as a function defined on a set representing all of the possible states of the network's target system. We typically assume that these functions are many-to-one (i.e., that they are surjective but not injective), such that random variables can be understood as compressions (often, massive compressions) of possibility space. For the same data-generating process, there are many different sets of variables that we can define on possibility space, and each choice of variables leads to a different causal model.[1] For example, one model could use a binary variable for a person's smoker status; another could include a variable representing the total number of cigarettes smoked in a person's life. In this case the former model is more compressed, but the latter is more informative. However, the more informative model is not always the superior model. In some cases, it may be no better-supported by the data than more compressed models, or the additional information that it encodes could be irrelevant in a given context. For instance, an epidemiologist will likely be

---

[1] Note that we do not assume here that there is a unique, most fine-grained representation of a given data-generating process. Rather, we take the fundamental nature of a data-generating process to be represented by a probability space, with respect to which the random variables in a causal model are measurable. See Appendix A for more details on how the variables in a causal model are defined with respect to a probability space.

uninterested in whether smokers are more likely to smoke with their right or left hand. Given these considerations, which model should be used in a given modelling context?

In philosophy of science, the question of how to make choices about which variable set to use when representing some system has been termed the "variable choice problem" (Woodward, 2016b). More precisely, the variable choice problem is the problem of determining which normative standards allow us to distinguish between appropriate choices of variables and cumbersome, unnatural choices of variables in contexts where empirical adequacy does not on its own vitiate in favor of one variable set or another. In this paper, we are especially concerned with versions of the variable choice problem that introduce a trade-off between compression and informativeness. As we discuss in the next section, this is a trade-off that arises across a number of areas in cognition and beyond.

The Case for Compression

The importance of compression for understanding one's environment is well-established within cognitive science. Rosch (1978) argues that classifying types of objects or events in one's environment requires one to balance the need for informative classification (i.e., maximal informativeness, which introduces pressure towards fine-grained categories) against a need for "cognitive economy" (i.e., less demanding representations, which introduces a pressure towards coarser representations) (see also Chater and Vitányi, 2003). This pressure towards cognitive economy in classification can be understood as a pressure towards compressed representations of one's environment, where this includes causal representations (see also Fauconnier and Turner, 2008; Murphy, 2004). Keil (2006) argues further that understanding how agents achieve optimal levels of compression in causal representations of their environment is crucial for understanding how those agents explain observed events. For example, if an agent knows that failing to discard

old food attracts pests, then that agent can exploit their knowledge of this causal relationship to explain the presence of pests when it occurs, and to intervene on their environment to avoid attracting pests in the future. This agent does not need to separately store a complex causal model of the relationship between different types of food (e.g., grains, vegetables, or meat) and the presence or absence of pests. Thus, compression allows for the recognition of high-level patterns of causal dependence; this accounts for the central role that compression plays in more general cognitive processes such as sense-making and understanding (see also Kirfel et al., 2021; Marzen and DeDeo, 2017; Pacer and Lombrozo, 2017; Wilkenfeld, 2019; Wojtowicz et al., 2021). Finally, Waldmann and Hagmayer (2006) show that people not only group objects together when they infer that they have similar causal effects, but also assume that objects have similar causal effects once they have been grouped together (see also Buchsbaum et al., 2015; Gopnik and Sobel, 2000).

In sum, there is a psychological and philosophical consensus that compression and causal explanation are closely related, and that our choice of variables for causal models often involves a trade-off between informativeness and compression. However, what is missing from this discussion is a precise quantification of the trade-off between informativeness and compression in Bayesian networks, as well as a systematic treatment of the value of the information contained in a Bayesian network for a given agent in a given context.

Outline of a Theoretical Framework and Empirical Hypotheses

Here, we provide a formal theory that quantifies the trade-off between compression and information loss that is often inherent in choices between variable sets in causal modeling. We then test this foundational framework empirically. Next, we show how this trade-off can additionally incorporate considerations about the value of information, reflecting the fact that not all information is equally valuable for a particular agent facing particular decisions. This allows

for more compressed causal models that, though they sacrifice informativeness relative to some more detailed model, are nevertheless optimal for a given agent in a given context, because the information lost in compression has no decision-theoretic value for the agent who uses the model to explain and intervene upon their environment. Having developed this more detailed formal framework, we also test it experimentally.

Importantly, on the picture of causal cognition that we use in this paper, we do not presuppose that agents generate a single, complete causal representation of their environment from raw data. Rather, we take it that agents approach their environment with a fragmented and revisable picture of the causal structure of the world. When asked to engage in explicit causal reasoning (e.g., when asked to evaluate the quality of a causal claim, or to generate a causal explanation or description based on data), agents refine this fragmented causal structure and make salient a specific, better-defined causal representation of their environment.

Based on our formal framework, and with this picture of flexible causal representation in the background, we obtain evidence consistent with the following three hypotheses:

H1: In general, people treat compression as a positive feature of a causal representation; all else being equal,[2] the more compressed a given representation is, the better.

H2. Compression can come at the cost of informativeness, and so, all else being equal, the optimal causal representation will achieve a balance of compression and informativeness.

H3: When people are asked to select a causal representation in the context of a particular decision problem, their tolerance for information loss in achieving a more compressed causal claim is moderated by the decision-theoretic value of the information that is lost.

---

[2] Throughout this paper, by "all else being equal," we mean holding fixed all features of the causal structure of a system, the interventional probability distributions over effects given causes, and the decision-theoretic context in which an agent finds themselves, apart from those explicitly stated to vary.

That is, when the information that is lost in moving to a more compressed causal claim is not decision-relevant, that information can be sacrificed without sacrificing the overall quality of a causal representation.

The remainder of the paper proceeds as follows. First, we review prior work (primarily from philosophy) on two questions about variable choice with implications for the compression of causal claims. These questions concern the granularity of selected variables (this is reflected in a claim's so-called "proportionality") and the choice of which variables to include (this is reflected in a claim's so-called "stability"). We then introduce a way to formalize both proportionality and stability in terms of the amount of information lost in the move from one causal model to another. This framework allows us to test our first two hypotheses (H1 and H2) empirically, which we do in Experiments 1-2. In Experiment 3, we rule out an alternative interpretation, showing that a causal contrast theory due to Lien and Cheng (2000) does not predict results that our formal theory is able to successfully accommodate.

Next, we consider how trade-offs between informativeness and compression are moderated by an agent's interests. In particular, information loss could be irrelevant to an agent in a given environment if the lost information would not affect their decisions. Addressing such cases requires extending our formal framework to incorporate the decision-theoretic value of information. After introducing the relevant theory and formalization, we again turn to human judgments to test our third hypothesis (H3) in Experiments 4-6. Specifically, we test the effect of the value of information on the trade-off between informativeness and compression in the evaluation (Experiment 4), generation (Experiment 5), and representation (Experiment 6) of causal claims.

Taken together, the paper makes the following novel contributions. On a theoretical level, our framework offers a novel, formal measure of the information lost in moving from one causal model to another, whether that loss is realized through changes in proportionality or stability. In so doing, our framework offers the first unified account of two dimensions of causal evaluation (proportionality and stability) that have previously received independent treatment. Our formalization also allows us to incorporate the decision-theoretic value of information, resulting in a new solution to the variable selection problem in building causal models. Empirically, we offer novel evidence that humans trade off informativeness and compression in evaluating causal claims, and moreover that the evaluation and production of causal claims is sensitive to the decision-theoretic value of information. We find no evidence that these trade-offs are handled differently when compression is achieved through changes in proportionality versus stability, lending empirical support to our formal unification. Finally, our findings contribute to a more general picture of human cognition as balancing tradeoffs between informativeness and compression in context-sensitive ways that take into account the value of information for a given agent in a given context.

Throughout our experiments, we occasionally consider causal relationships in which one cause can act as a moderator of the relationship between another cause and its effect. While our framework does not *only* handle these cases, the fact that it is able to handle them speaks to its advantages over other frameworks, such as the PowerPC theory of Cheng (1997), which generates predictions about human causal judgments only in cases where candidate causes are antecedently assumed to be independent of each other.

Dimensions for Comparing Causal Claims Across Compressions

On its own, the Bayesian network approach to causal representation treats causation as a binary relation; two variables are either causally related or they are not, and so the corresponding causal claim (e.g., "smoking causes cancer") is either appropriate or it is not. However, we take it that one can nevertheless make graded distinctions between causal claims along a large number of different dimensions, in keeping with experimental work on causal judgement (e.g., Cheng, 1997; Gerstenberg et al., 2021; Icard et al., 2017; Lombrozo, 2007, 2010; Morris et al., 2018; O'Neill et al., 2021, 2022; Quillien, 2020; Spellman, 1997). In our framework, we consider two dimensions along which causal models can vary, corresponding to the following two questions about variable choice in building a causal model: i) At what level of granularity should a variable be defined?, and ii) Which variables should be included in a causal model?

These two questions correspond to two graded dimensions that have been discussed in the philosophy literature on causation, especially by Woodward (2008, 2010, 2016, 2021a, 2021b), though by others as well (e.g., Blanchard, Vasilyeva, & Lombrozo, 2018; Bourrat, 2021; DiMarco, 2021; Franklin-Hall, 2016; Gebharter and Eronen, 2021; Harbecke, 2021; Hoffmann-Kolss, 2014; List and Menzies, 2009; Ross, 2015; Weslake, 2013). These dimensions are *proportionality* and *stability*. A causal claim's 'proportionality' depends on the extent to which it is informative about how possible changes to the cause would result in changes in its effect. Since coarse-graining or refining variables that stand in causal relationships to one another can change the proportionality of causal claims, the concept of proportionality offers a partial answer to the question, "at what level of granularity should a variable be represented?" By contrast, 'stability' refers to the degree to which a causal relationship is insensitive to changes in the values of unspecified background

variables, and thus offers a partial answer to the question, "which variables should be included in a causal model?" Below, we describe both proportionality and stability in greater detail.

Proportionality

Proportionality is described by Woodward as the degree to which a causal claim of the form 'C causes E,' where C and E are variables in a causal structure, is stated at the "level [of causal description] that is most informative about the conditions under which the effect will and will not occur'' (2021a, p. 389). For Woodward, the hierarchy of levels of description with which a causal relationship can be stated corresponds to a sequence of "vertically" related causal variables, where each causal variable in the sequence is a coarsening of the previous causal variables (2021a, p. 371). The standard example of a hierarchy of levels of description that differ with respect to their proportionality comes from Yablo (1992). Consider a pigeon who has been trained to peck at all and only red targets. In a causal model of a system containing this pigeon and various targets, one might have a variable C with the range of values {red target, non-red target}, and another variable E with the range of values {pigeon pecks, pigeon does not peck}. To be an accurate representation of the underlying data-generating structure, the model would have to be such that C is a cause of E. However, one could also generate a causal model in which C is replaced with a variable C' with the range of values {scarlet target, non-scarlet red target, non-red target}. Here, accuracy would also demand that we say that C' is a cause of E, since some changes in the value of C' lead to changes in the value of E; changing C' from either of its first two values to its third, and vice versa, leads to changes in the value of E.

According to Woodward's definition, the claim 'C causes E' has the same level of proportionality as the claim 'C' causes E.' This is because a function specifying how changes in C bring about changes in E would give an agent all the information that they need to appropriately

manipulate the effect, as would a function specifying how changes in C' bring about changes in E. However, 'C causes E' achieves a more compressed representation of the data-generating process than 'C' causes E.' This is because C is a *coarsening* of C'. That is, it defines a strictly more general equivalence class on possibility space: any scarlet target or non-scarlet red target is still a red target. So, to the extent that we aim to optimize proportionality in our causal representations, we are *licensed* to use compressed representations when those compressions do not result in a reduction in proportionality, as argued in Woodward (2021b).[3] If the pigeon had instead been trained to peck at scarlet targets (and not other red targets), then the claim 'C' causes E' would be more proportional than 'C causes E,' because the former would be more informative about the conditions under which the effect will and will not occur (i.e., that it will occur for red targets that are scarlet, but not for red targets that are not scarlet).

When testing the influence of proportionality in people's evaluations of causal claims, it makes sense to consider causal models in which compression is achieved by coarsening the range of a particular causal variable. Consider a causal claim $C' \rightarrow E$ that is embedded within a given causal model. If we replace C' with a coarsening C, but leave the rest of the model unchanged, then we compress the model by replacing a variable with its coarsening. We can then assess how much information about the likely value of E is conveyed by changes in C in this more compressed model, as compared to how much information about likely values of E is conveyed by changes in C' in the less compressed model. This tells us how proportional the claim $C \rightarrow E$ is, as compared

---

[3] In Yablo's original paper, as well as in earlier work on this topic by Woodward (2010), the definition of proportionality was such that the variable with the range of values {red target, non-red target} would be said to be a *more* proportional cause of the variable with range of values {pigeon pecks, pigeon does not peck} than the variable with range of values {scarlet target, non-scarlet red target, non-red target}. However, as Woodward (2021a) notes, the amount of information that the value of either of these causal variables conveys about the value of the effect variable is equal; specifying the value of either causal variable tells us what the value of the effect variable would be. For this reason, we follow the later Woodward in treating the two variables as equally proportional, with this licensing the choice of the more coarse-grained variable.

to the claim $C' \rightarrow E$. Thus, evaluations of the relative proportionality of causal claims involve comparisons of more and less compressed causal models of the same data-generating process. In what follows, this idea will be made mathematically precise.

To date, little empirical work has considered proportionality as a relevant dimension in people's evaluations of causal claims. Of most direct relevance, Lien and Cheng (2000) offer evidence that agents prefer proportional causal claims. Specifically, they show that people prefer to give causal explanations that "explain as much as possible with as few causal rules as possible," (p. 88). By "causal rules," they mean mappings from values of a causal variable to values of an effect variable. This preference for less informationally detailed causal relations tracks a preference for more coarse-grained explanations in cases where fine-graining only serves to complicate the description of the relation between cause and effect. Thus, they find that agents prefer just those kinds of coarsenings that are licensed by a preference for more proportional causal claims. However, their work does not address the question of whether and how judgments of more or less proportional causal claims instantiate a trade-off between compression and informativeness. In other relevant work, Bechlivanidis et al. (2017) report a preference for concreteness over abstraction in causal explanation (i.e., people favor explanations with finer-grained variable choices), even in cases where such concreteness does not add any information that would be relevant for predicting the effect in question. While this suggests that people may ignore compression in favor of greater detail under some conditions, their task involved evaluating causal explanations for token events, not type-level causal claims (see also Aronowitz and Lombrozo, 2020, for potentially relevant discussion). We take explaining why a particular event happened to be a cognitively distinct task from identifying patterns in the causal relations between types of

events, such that the norms governing the former might be different from those governing the latter.[4] Our focus here is on the latter cognitive task.

Stability

Stability is the extent to which a causal claim is sensitive to changes in unspecified background conditions (Woodward, 2010).[5] As an example of a highly stable causal relationship, consider "smoking causes lung cancer." Across a wide range of plausible changes to other features of the world, people who smoke are more likely to get lung cancer than those who do not, and this statistical association is due to a causal relationship between smoking and lung cancer. By contrast, consider Woodward's example of a binary variable C representing whether or not a person has a particular genetic mutation that typically results in dyslexia, and another binary variable E denoting whether or not that person eventually learns to read. Suppose that there is widespread failure to address and correct early reading failures in dyslexic children. Under these conditions, the causal claim 'C causes E' would hold; the presence or absence of the genetic mutation would lead to changes in the likelihood of a person learning to read. However, this claim is highly *un*stable. If we alter the background conditions so that dyslexia is treatable and is treated (as, in fact, it often is), then the causal effect of the genetic mutation on the probability of one's learning to read is greatly diminished.

---

[4] This is not to say that these tasks are entirely unrelated. For example, seeking an explanation for a particular event might lead someone to identify patterns in the causal relations between types of events, and beliefs about such causal relations surely constrain explanations for particular events. Our point here is simply that finding a preference for more detailed descriptions in some explanations for particular events does not imply a preference for finer-grained specifications of type-level causal relationships. To illustrate, the task of explaining why a particular student failed a test or why a particular tree stopped bearing fruit is a different cognitive activity from finding the right level of granularity for the variables used in stating theories about broad causal patterns like 'studying causes good performance on tests' or 'poisonous fungi cause trees to stop bearing fruit.' Our focus in this paper is on the latter sort of cognitive activity, rather than the former.

[5] By "unspecified" background conditions, we mean those conditions that are not explicitly part of the causal claim. For example, the stability of the claim "smoking causes lung cancer" depends on its persistence across changes in whether a person also lives in a high-air-pollution environment, but the stability of the claim "smoking and living in a high-air pollution environment causes lung cancer" does not, since in the latter case the amount of air pollution in the environment is specified in the causal claim.

When we compress a causal model by removing a set of variables, we can regard the removed variables as background conditions, and assess the stability of the causal relationships that remain after compression by observing how well they are preserved even as background conditions are removed. Whether a variable is designated as a background condition as opposed to a causal variable of interest is ultimately a distinction that we impose on the model, rather than one that is dictated by the nature of the system being modelled. That is, we can stipulate that a particular relationship $C \rightarrow E$ in a causal model is of interest, and designate other variables in the model as background conditions.[6] We can then assess the stability of the relationship $C \rightarrow E$ by compressing the model in which it is embedded so as to remove the background conditions, and then measuring the degree to which changes in the value of C still result in changes in the probability distribution over E. Thus, as in the case of proportionality, we can measure the stability of a causal claim by comparing how much information about the likely value of the effect variable is conveyed by changes in the causal variable when the claim is embedded in a more or less compressed causal model. This too will be made more mathematically precise in what follows.

Prior work offers some direct evidence that people favor causal claims that are more stable over those that are less stable. Specifically, Vasilyeva et al. (2018) provide evidence that people prefer more stable causal claims even when other dimensions of causal variation (such as causal strength) are carefully controlled. As with proportionality, however, this empirical work does not address the question of whether and how judgments of more or less stable causal claims instantiate a trade-off between compression and informativeness.

In sum, both proportionality and stability can be assessed by comparing how informative changes in a causal variable are with respect to likely values of an effect variable across more or

---

[6] See Watson and Silva (2022) for an example from the machine learning literature in which background and foreground causal variables are distinguished by stipulation, with fruitful results.

less compressed causal models in which the cause-effect relationship is embedded. This suggests that proportionality and stability are actually two species of the same genus. That is, both proportionality and stability are measures of how much information a causal relationship preserves across different types of compression. To test this hypothesis, in our experiments we manipulate whether participants are put in scenarios in which a compression of the implicit causal model involves coarsening a causal variable or eliding a background condition. Across multiple experiments, while we find evidence that varying the amount of information lost in compression affects peoples' judgments about the aptness of a causal representation, we find no evidence for an effect of varying whether compression involves coarsening a causal variable or eliminating a background condition. Thus, our unified analysis of proportionality and stability is in keeping with the findings of our experiments, but represents an important departure from prior work within philosophy, which has sometimes aimed to offer a formalization of stability and/or proportionality (e.g., Pocheville et al., 2017), but without offering a unifying framework for both.

Formalizing Information Loss Due to Compression

In this section, we formalize our notion of information loss due to compression, and how it can be used to evaluate both the proportionality and the stability of causal claims. (Our formalization here is given at a relatively high level of abstraction; see Appendix A for a more complete presentation in the language of graphical causal models. On the other hand, readers who prefer to move directly to an empirical test of hypotheses H1-H3 can skip to the section titled "Motivation for Experiments 1-2.") As a preliminary point, note that both a set of causal variables C and an effect variable E are random variables measurable with respect to the same probability space. Moreover, we assume that both random variables are situated within a **Bayesian network**. That is, there is an acyclic set of directed edges connecting the random variables, representing

causal relationships between them. Importantly, this graphical structure satisfies the **Markov condition** with respect to the probability distribution over the variables in the graph: all variables are independent of their non-descendants, conditional on their parents. This setting allows us to calculate the probability distribution over E given each possible intervention setting a set of variables C to each of its possible sets of values, in the style of Pearl (2000). We denote the probability that E takes a value e, given that C is set to a vector of values c via intervention, using the notation $p(e \,|\, do(c))$. We define the **causal mutual information** between C and E as follows:

$$CMI(C,E) = \sum_{c,e} q(c)\, p(e \,|\, do(c))\, log_2 \frac{p(e \,|\, do(c))}{p(e)}$$

where q is a probability distribution over possible interventions on C and c and e are possible values of C and E, respectively.[7] Note that this is just the standard definition of mutual information found in Shannon (1948), with interventional conditional probabilities substituted for conditional probabilities. As such, it measures the average amount of information that any possible intervention on a set of causal variables communicates about which of any possible values the effect variable will take. This makes it a measure not of the informativeness of any one causal claim, but rather a measure of the informativeness, with respect to effect variable E, of an entire causal model containing causal variables C.

We aim to define a lexical notion of compression that allows us to meaningfully state that one set of variables $\widehat{C}$ is **more compressed** than another set C. Intuitively, we want the relation between more and less compressed variable sets to imply that any distinction captured in the more compressed variable set $\widehat{C}$ is also captured in the less compressed variable set C, but that C can

---

[7] The idea of a probability distribution over possible interventions is non-standard, though it does have some precedent in the causal inference literature (Pearl, 1994). In particular, it is necessary to define the information capacity of a causal channel in a way that incorporates an interventional understanding of causation (Ay and Polani, 2008). Where q ranges over the values of multiple variables, we assume that *all* variables are intervened upon. That is, q does not allow for the possibility of intervening on only some variables.

contain distinctions that are not captured in $\widehat{C}$. To this end, we begin by noting that for any set of causal variables C, we can define a surjective but not injective **compression function** σ from the range of C into another set. This yields a set of variables $\widehat{C}$ whose set of possible values is the range of the function σ. Going forward, we will say that such a variable set $\widehat{C}$ is a **compression** of C. We stipulate that for any vector of values $\hat{c}$, the probability $q(\hat{c})$ is given by the equation $\sum_{c \in \sigma^{-1}(\hat{c})} q(c)$. Importantly, any probability distribution q over interventions and any other probability distribution p must satisfy the constraint that $p(e \,|\, do(\hat{c})) = \sum_{c \in \sigma^{-1}(\hat{c})} p(e\,|\,do(c)) \frac{q(c)}{q(\hat{c})}$.

In other words, the probability that $E=e$ given a coarse-grained intervention setting $\widehat{C}$ to $\hat{c}$ is given by the average, according to q, of each interventional conditional probability $p(e\,|\,do(c))$ for each $c \in \sigma^{-1}(\hat{c})$. Having made these stipulations, we arrive at a lexical, comparative measure of compression: one set of variables $\widehat{C}$ is more compressed than another set C if there is a surjective but not injective compression function σ from the range of C to the range of $\widehat{C}$.[8] In Appendix A, we define the compression relationship between variables more rigorously, but the definition here is sufficient for expositional purposes.

---

[8] On this lexical measure, it only makes sense to say that one set of variables is more/less compressed than another if the first is a compression of the second or vice versa. We deliberately do not provide a more general measure of compression as such a measure will invariably vitiate on questions not directly relevant to our interests here. For instance, a measure of how compressed the variable set C is that uses the entropy of C will be sensitive to the prior distribution over C, but it is unclear why the likelihood that different values of C occur should be relevant to how compressed C is. On the other hand, if we measure how compressed C is by simply counting the number of variable combinations, we lose the ability to say (for example) that a countably infinite discretization of the unit interval results in a more compressed representation. Ultimately, we are interested here in agents' preferences between nested representations of the same causal processes, and so we leave to one side a more general measurement of compression. We believe that comparisons between nested representations deserve special attention, since our choices between different nested representations ultimately determine whether or not any two events observed in our environment should be treated as instantiations of the same causal kind.

We are now in a position to define the amount of information about an effect variable that is lost in the move from a less compressed to a more compressed set of causal variables. We define this quantity as follows:

$$\mathscr{L}(C, \widehat{C}, E) = CMI(C, E) - CMI(\widehat{C}, E)$$

Thus, the amount of information about E that is lost in the move from the less compressed causal representation C to the more compressed representation $\widehat{C}$ is equal to the difference between the causal mutual information that C provides about E and the causal mutual information that $\widehat{C}$ provides about E. Following Shannon (1948), one should measure information loss in terms of the number of *bits* of information lost in the move from a less compressed representation to a more compressed one. In what follows, we will use this equation to derive qualitative predictions about when people will prefer more or less compressed causal representations of their environment, which we then confirm experimentally.

We focus on qualitative comparisons between participants and our model because our goal is to understand the conditions under which people prefer more compressed causal representations of their environment, and our measure provides a principled basis for tractably deriving predictions. Our aim is *not* to directly model the mechanisms by which people estimate compression, and so we leave open the possibility that other measures of compression can be used to derive similar predictions, and may offer more plausible accounts of human cognition at an algorithmic level. Having said this, in what follows we will note some respects in which salient rivals to our framework are less able to explain our data.

Measuring Proportionality Using Information Loss

Having formalized information loss due to compression, we can apply our framework to the compression achieved by replacing a causal variable C with a more compressed variable $\widehat{C}$ and

leaving all other variables unchanged. By comparing the amount of information that C communicates about some effect variable E to the amount of information that $\widehat{C}$ communicates about the same variable E, we can measure the amount of information that is lost in this compression. This corresponds to a comparison of the causal claims 'C causes E' and '$\widehat{C}$ causes E' with respect to their proportionality.

More precisely, let $\mathscr{C} = \left(C_1, ..., C_n \right)$ be a sequence of causal variables, with each C_i a compression of all variables $C_{j<i}$. We then say that, in the context of such a sequence, a variable C_i is **proportional** with respect to an effect variable E to the extent that $\mathscr{L}\left(\{C_j\}, \{C_i\}, E \right)$ is relatively small for all $j < i$, where what counts as "relatively" small is assessed with respect to the amount of information that is lost as each causal variable in the sequence $\mathscr{C} = \left(C_1, ..., C_n \right)$ is replaced with a more coarse-grained alternative. That is, proportional choices of causal variables are those that preserve information about the conditions under which an effect variable E will change, as compared to less compressed alternatives. Note that in this paper we only consider comparisons of proportionality between causal claims with different causal variables and a common effect variable, though one can in principle compare causal relationships that differ with respect to both cause and effect in terms of proportionality. We expect that our use of information loss to measure proportionality generalizes to such comparisons.[9]

Measuring Stability Using Information Loss

Our formalization of information loss due to compression can also be applied to compressions achieved through the omission of background variables. Recall from our earlier discussion that we can measure the stability of a causal relationship $C \rightarrow E$ embedded in a particular

---

[9] Specifically, one could measure the difference between the difference $CME(\, C, E) - CME(\, \widehat{C}, \widehat{E})$, where $\widehat{E}$ is a compression of E.

causal Bayes net by removing a set of variables $\boldsymbol{B}$ from that Bayes net and assessing how much information is lost in the move from the original Bayes net to the Bayes net that is created by removing the background variables. We stipulate that the variables in a set $\boldsymbol{B}$ are **background variables** with respect to a causal relationship $C \to E$ if and only if removing all variables in $\boldsymbol{B}$ and all edges going into or out of the variables in $\boldsymbol{B}$ creates a Bayesian network that still satisfies the Markov condition. The notion of causal stability can now be made precise, using our proposed measure of information loss. Specifically, we will say that the causal relationship between C and E is **stable** with respect to background condition $\boldsymbol{B}$ to the extent that the value of $\mathscr{L}(\{C, \boldsymbol{B}\}, \{C\}, E)$ is low. That is, the relationship $C \to E$ is stable with respect to $\boldsymbol{B}$ to the extent that the average amount of information about E that is communicated by interventions on both C and the variables in $\boldsymbol{B}$ is similar to the average amount of information about E that is communicated solely by interventions on C. Note that we assert only this monotonic relationship between $\mathscr{L}(\{C, \boldsymbol{B}\}, \{C\}, E)$ and stability, and make no claim about exactly how low $\mathscr{L}(\{C, \boldsymbol{B}\}, \{C\}, E)$ has to be for the relationship $C \to E$ to be sufficiently stable for a given person to eschew any mention of the background variable $\boldsymbol{B}$.

Relationship to Existing Approaches

Our framework is not the first to use information theory to quantify properties of causal relationships. Previous work in this vein includes specific attempts to measure proportionality and stability (Pocheville et al., 2017), as well as attempts to measure other properties of causal relationships, such as power, abstraction, strength, or specificity using formalism from information theory (Ay and Polani, 2008; Beckers and Halpern, 2019; Bourrat, 2021a, 2021b; P. E. Griffiths et al., 2015; Hoel, 2017; Korb et al., 2011). Moreover, our account of information loss can be understood as a version of *rate distortion theory* (Sims, 2016; Zaslavsky et al., 2018), which has

been applied to discrete categorization of stimuli and working memory, though not (as far as we know) to causal representation. On this interpretation, the compression function defines a distortion channel, or "information bottleneck" (Tishby et al., 2000), through which information is passed from the less compressed variable set C to the effect variable E, and the function $\mathscr{L}$ measures how much information is lost in the distorting process of compression. Despite these relationships to prior work, none of these approaches argue, as we do, that measurements of the proportionality and stability of a causal relationship can both be expressed in terms of information loss.

Motivation for Experiments 1-2

We have now introduced our theoretical framework for quantifying the amount of information lost in moving from a specific causal representation of a given data-generating process to a more compressed causal representation of that same process. We intend for this framework to serve both a normative and a descriptive role. On the normative side, we have argued above that two putatively positive features of causal claims, proportionality and stability, can be given a unifying account in terms of information loss in causal models. On the descriptive side, we hold that agents, all things being equal, trade off a preference for more compressed causal representations against a preference for causal representations that minimize relative information loss (this descriptive claim summarizes the hypotheses **H1** and **H2** above). These hypotheses can be tested empirically, and we report the results of two experiments that do so in the following two sections.

In order to investigate people's causal representations, we need a measurable response that reflects which variables they are representing and at what level of granularity. Asking people to generate or evaluate Bayes' nets would be a natural approach given our formalism, but doing so

would likely require training (see, for example, Bramley et al., 2015, 2017) and this approach would fail to reflect the way that people tend to explicitly represent causal relationships in everyday life. Instead, causal relationships are often expressed in the form of generic causal claims of the form "*C* causes *E*." Not coincidentally, we have made several such claims in the course of this paper (for example, "smoking causes lung cancer"). Such claims can be taken as evidence for causal representations on the assumption that they bear some systematic relationship to the variables an individual has represented in their internal causal model. Specifically, we take the generation or positive evaluation of a generic causal claim as evidence that the individual producing or endorsing that generic holds a causal model consistent with that claim. For example, suppose that an individual shown data about a fictional insect called the Bricofly evaluates the claim 'raising Bricofly larvae in a warm, humid tank causes them to develop blue wings' more positively than the claim 'raising Bricofly larvae in a warm tank causes them to develop blue wings.' We would take this to be evidence that the individual is representing the environment using a causal model of Bricofly development in which both different temperatures and different humidity levels of the tank are retained in their representation, as opposed to a model in which only the temperature of the tank is represented.

The assumption that the endorsement of generic causal claims reflects an individual's causal representation, which guides our methods in Experiments 1-4, is in keeping with a core tenet of the philosophy of science literature on causation: namely, that generic causal claims are derived from causal models. For instance, Papineau describes himself as "committed to reading 'C causes E' as saying that C is an ancestor of E in a system of generic causal equations," where one can understand a 'system of generic causal equations' as a variety of graphical causal model (Papineau, 2022, p. 20). Similarly, Woodward writes that "[causal] modeling techniques are at

least sometimes successful in reliably establishing generic causal claims" (Woodward, 2019, p. 765). Thus, the current paper's assumption of a close connection between the causal models used to represent a system and generic causal claims about a system is broadly in keeping with at least one branch of the philosophical mainstream.

Experiment 1

In Experiment 1, we test **H1**: that in general, people treat compression as a positive feature of a causal representation, such that all else being equal, the more compressed a given representation is, the better. To test this hypothesis, we presented participants with a description of the results of controlled experiments on a fictional variety of mushroom, fly, or rock, and asked them to rate how good it would be to include various claims in a summary of the described results. These claims included more and less compressed causal claims (e.g., the more compressed claim 'raising Bricofly [a fictional type of insect] larvae in a warm tank causes them to develop blue wings,' and the less compressed claim 'raising Bricofly larvae in a warm, humid tank causes them to develop blue wings'). We manipulated both the vignette used and whether the compression was achieved by coarse-graining a variable (thus manipulating proportionality) or removing a background variable (thus manipulating stability). We predicted that in the absence of information loss due to compression, participants would favor the more compressed representation, and that this would hold for manipulations of both proportionality and stability.

Experiment 1 was also designed to test **H2**: that compression can come at the cost of informativeness, such that the optimal causal representation will achieve a balance of compression and informativeness. To test this, we also manipulated the amount of information loss realized by the more compressed causal claim. For example, if raising Bricofly larvae in a warm, humid tank results in an 85% probability of their developing blue wings, and raising Bricofly larvae in a warm,

dry tank results in a 70% probability of their developing blue wings, then implicitly, assuming an equal probability of having a humid or dry tank, the probability of developing blue wings when Bricofly larvae are raised in a warm tank is 77.5%. Under these conditions, and assuming a 1% probability of developing blue wings for larvae raised in a either a cold, humid tank or a cold, dry tank, moving from a causal model that keeps track of the humidity or dryness of the tank to one that does not keep track of these properties results in a loss of information of .01, on the measure of information loss that we introduce above.[10] Thus, by manipulating the conditional probabilities in the data sets shown to participants, we were able to manipulate the amount of information loss inherent in endorsing a more compressed causal claim. We predicted that with greater information loss, we would see lower evaluations of the more compressed representation relative to the less compressed representation.

The data, stimuli, and pre-registrations for all experiments in this paper are available at https://osf.io/cnwk6/?view_only=f1c75535dc2c4a30a6f813a88baebd3a.

Participants

Participants were 450 adults recruited via Prolific. An additional 150 participants were excluded for failing comprehension checks or for rating poor causal claims non-negatively. The sample of participants was 49.6% female and 48.9% male, with an age range of 19-79 and a mean

---

[10] Here, we give explicit calculations of information loss for x=.7 and x=.85 to illustrate how these values are calculated. When x=.7, the probability p(Blue Wings) = $.25[.7 + .7 + .01 + .01] = .355$. The causal mutual information between the less compressed causal variable and the binary variable for whether a Bricofly develops blue wings is $.25\left[.7\log_2\frac{.7}{.355} + .7\log_2\frac{.7}{.355} + .01\log_2\frac{.01}{.355} + .01\log_2\frac{.01}{.355} + .3\log_2\frac{.3}{.645} + .3\log_2\frac{.3}{.645} + .99\log_2\frac{.99}{.645} + .99\log_2\frac{.99}{.645}\right] \approx .47.$ The causal mutual information between the more compressed causal variable and the binary Blue Wings variable is $.5\left[.7\log_2\frac{.7}{.355} + .01\log_2\frac{.01}{.355} + .3\log_2\frac{.3}{.645} + .99\log_2\frac{.99}{.645}\right] \approx .47$ so that the information lost due to compression is zero. When x=.85, p(Blue Wings) = $.25[.85 + .7 + .01 + .01] = .3925$ and so the causal mutual information between the less compressed causal variable and the binary variable for whether a Bricofly develops blue wings is $.25\left[.85\log_2\frac{.85}{.3925} + .7\log_2\frac{.7}{.3925} + .01\log_2\frac{.01}{.3925} + .01\log_2\frac{.01}{.3925} + .15\log_2\frac{.15}{.6075} + .3\log_2\frac{.3}{.6075} + .99\log_2\frac{.99}{.6075} + .99\log_2\frac{.99}{.6075}\right] \approx .55$, while the causal mutual information between the more compressed causal variable and the binary Blue Wings variable is $.5\left[.775\log_2\frac{.775}{.3925} + + .01\log_2\frac{.01}{.3925} + .225\log_2\frac{.225}{.6075} + .99\log_2\frac{.99}{.6075}\right] \approx .54$, yielding an information loss due to compression of approximately .01.

age of 40.[11] For all studies reported here, participation was restricted to users with a US-based IP address and a 95% rating based on at least 100 previous studies. All studies received IRB approval from the authors' University.

Materials and Procedures

Participants read a vignette in which they learned about a novel causal system, including the results of experiments involving that system. For example, in the insect vignette, participants were presented with one of the following reports of results of experiments on the fictional "Bricofly"[12]:

Report 1:

a) x% of all Bricofly larvae raised in a warm, humid tank developed blue wings;

b) 70% of all Bricofly larvae raised in a warm, dry tank developed blue wings;

c) 1% of all Bricofly larvae raised in a cold, humid tank developed blue wings;

d) 1% of all Bricofly larvae raised in a cold, dry tank developed blue wings.

Report 2:

a) x% of all Bricofly larvae raised in a warm tank and sprayed with water developed blue wings;

b) 70% of all Bricofly larvae raised in a warm tank and blown with dry air developed blue wings;

---

[11] Demographic data obtained from participants was not solicited as part of our study, but was provided by participants to Prolific upon joining the platform.

[12] The exact wording of this cover story for participants shown Report 1 was as follows. "Suppose that you are asked to briefly summarize the findings of a scientific report on the Bricofly, a newly discovered type of insect. Specifically, the report aims to investigate why some Bricofly develop blue wings. The report describes the outcomes of controlled experiments that involve raising Bricofly larvae in a tank that is kept at a temperature that is either warm or cold, and either humid or dry."

c) 1% of all Bricofly larvae raised in a cold tank and sprayed with water developed blue wings;

d) 1% of all Bricofly larvae raised in a cold tank and blown with dry air developed blue wings.

The value of x was varied between subjects and set at either 70, 85, or 98. These values correspond to information loss amounts of 0, .01, and .06 respectively in moving from a less compressed representation to a more compressed representation, assuming a uniform distribution over possible interventions on the causal variable(s) in the less compressed models in which the evaluated causal claims can be represented.

Table 1

Structure of Vignettes used in Experiments 1 and 2.

Vignette	Effect	Primary Cause	Secondary Cause	Background Condition
Drol (Mushroom)	Bumpy Stems	High/Low Mineral Soil	High/Low Sodium Soil	Watered with Salt/Fresh Water
Bricofly (Insect)	Blue Wings	Warm/Cold Tank	Humid/Dry Tank	Water Spray/Dry Air Blow
Chapagite (Rock)	Fissures	Warm/Cold Water	Salt/Fresh Water	Wrapped in Saline/Plain Cloth

After receiving a version of the findings described above, participants were then asked to rate, on a scale from -3 (very bad) to 3 (very good), "how good it would be to include each of the following statements in a summary of this report:"

[*Compressed*:] Raising Bricofly larvae in warm tank causes them to develop blue wings.

[*High*:] Raising Bricofly larvae [in a warm, humid tank/in a warm tank and spraying them with water] causes them to develop blue wings.

[*Low*:] Raising Bricofly larvae [in a warm, dry tank/in a warm tank and blowing them with air] causes them to develop blue wings.

The causal claims *High* and *Low* are so-named because the cause cited in *High* always confers the same or greater probability onto the effect than the causal claim *Low* (e.g., it is always the case that $p(\text{Blue Wings} \mid \text{Warm, Humid Tank}) \geq p(\text{Blue Wings} \mid \text{Warm, Dry Tank})$).

As an attention check, participants were also asked to rate the following causal claims:

[*Compressed (Bad):*] Raising Bricofly larvae in a cold tank causes them to develop blue wings.

[*High (Bad)*:] Raising Bricofly larvae [in a cold, humid tank/in a cold tank and spraying them with water] causes them to develop blue wings.

[*Low (Bad)*:] Raising Bricofly larvae [in a cold, dry tank/in a cold tank and blowing them with air] causes them to develop blue wings.

Since all of these claims cite a cause that *lowers* the probability of its effect, we take them to be infelicitous in the context of the scenario shown to participants. Thus, participants who gave ratings at or higher than the scale midpoint (i.e., a rating of 0-3) were excluded.
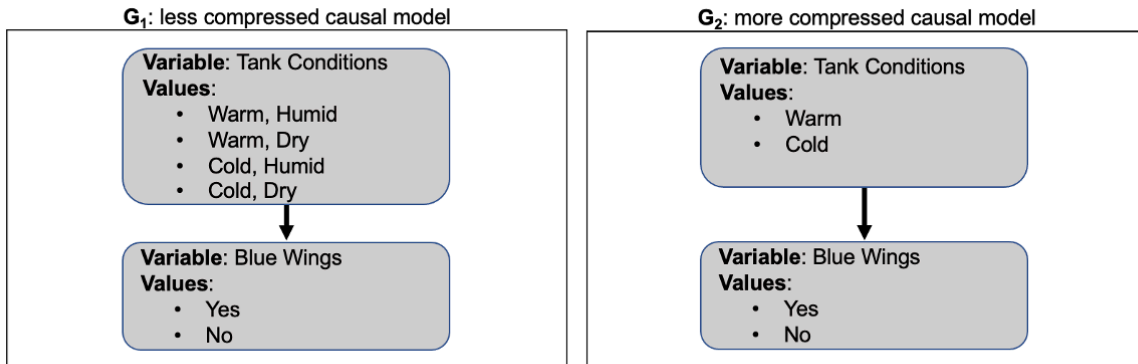
For participants shown Report 1, the claim *Compressed* is a compression achieved by coarsening a causal variable, thus varying proportionality. For participants shown Report 2, the claim *Compressed* is a compression achieved by eliding a background variable, thus varying stability.[13] Figure 1 shows the implicit causal model in which each of these claims is embedded.

---

[13] As described in the previous section, on our formal analysis, both the coarsening of a causal variable and the elision of a background condition can both be expressed as compressions of a partition over the same sample space, such that there is no difference between these two kinds of compression. We have fashioned our examples to match what is understood in the literature (e.g., Woodward, 2010) as a distinction between a refinement of the same variable and an elision of a background condition.

Figure 1

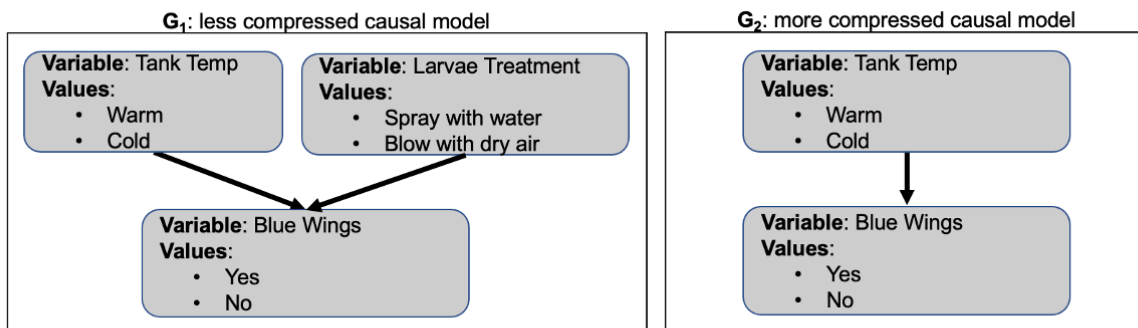*Graphs Showing the Causal Relationships between Variables in Experiment 1*

a) Compression achieved through coarsening a background variable (proportionality)

G_1: less compressed causal model

Variable: Tank Conditions
Values:
- Warm, Humid
- Warm, Dry
- Cold, Humid
- Cold, Dry

Variable: Blue Wings
Values:
- Yes
- No

Example claim: "Raising Bricofly larvae in a warm, humid tank causes them to develop blue wings."

G_2: more compressed causal model

Variable: Tank Conditions
Values:
- Warm
- Cold

Variable: Blue Wings
Values:
- Yes
- No

Example claim: "Raising Bricofly larvae in a warm tank causes them to develop blue wings."

**b) Compression achieved by removing a background variable (stability)**

G_1: less compressed causal model

Variable: Tank Temp
Values:
- Warm
- Cold

Variable: Larvae Treatment
Values:
- Spray with water
- Blow with dry air

Variable: Blue Wings
Values:
- Yes
- No

Example claim: "Raising Bricofly larvae in a warm tank and spraying them with water causes them to develop blue wings."

G_2: more compressed causal model

Variable: Tank Temp
Values:
- Warm
- Cold

Variable: Blue Wings
Values:
- Yes
- No

Example claim: "Raising Bricofly larvae in a warm tank causes them to develop blue wings."

Note. The top panel shows the compression used in the proportionality condition, and the bottom panel shows the compression used in the stability condition. In both panels, the right graph shows the more compressed causal model in which the claim *Compressed* is embedded, and the left graph shows the less compressed causal model in which the claims *High* and *Low* are embedded.

Results

Figure 2 presents the mean ratings for each of *Compressed*, *High*, and *Low* as a function of the amount of information loss incurred through compression. To test whether evaluation of less compressed causal claims relative to more compressed causal claims increased as a function of information loss due to compression, we computed (as pre-registered) two difference scores:

Figure 2

Evaluations of Causal Claims as a Function of Information Loss
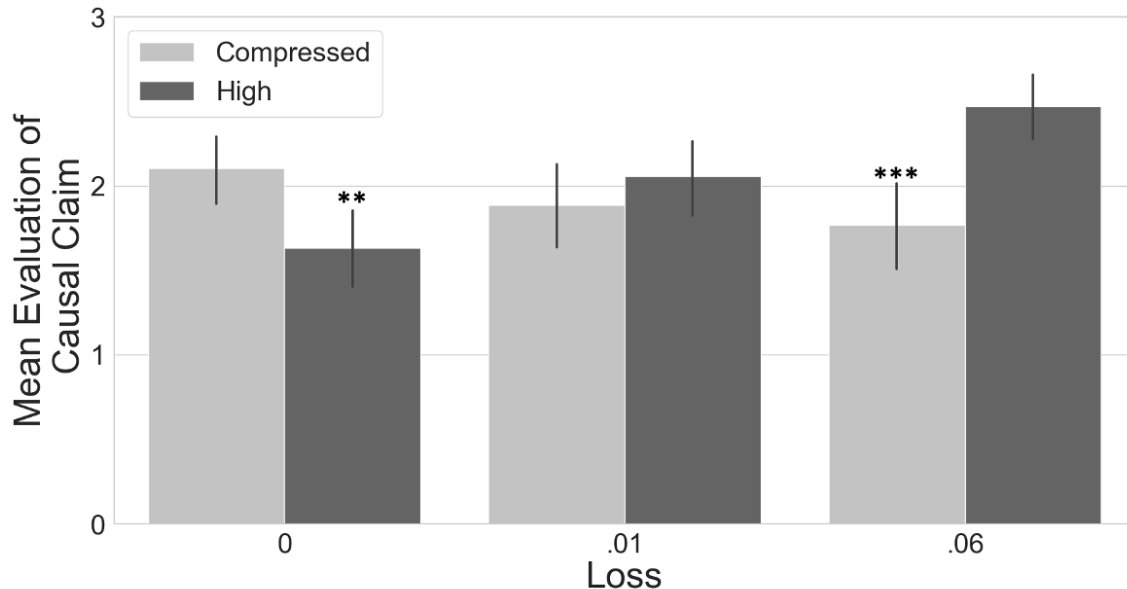
Note. Bar plots showing mean evaluations in Experiment 1 (with 95% confidence intervals) of causal claims under different loss conditions for all participants. Participants were asked to rate "how good it would be to include each of the following statements in a summary of this report," with ratings ranging from -3 (very bad) to 3 (very good). Double asterisks indicate significant within-participants differences at the .01 level in a mixed ANOVA, and triple asterisks indicate significant within-participants differences at the .001 level. Mixed ANOVA for each value of Loss found that at Loss=0, Compressed was rated more highly than High ($\eta^2 = .025$, $p = .002$). When Loss=.01, High was not rated significantly higher than Compressed ($\eta^2 = .005$, $p = .156$). When Loss=.06, High was rated significantly higher than Compressed ($\eta^2 = .058$, $p < .001$).

COMPRESSED-HIGH: The difference between the participant's evaluation of *Compressed* and their evaluation of *High* (e.g., the difference between the evaluation of 'Raising Bricofly larvae in a warm tank causes them to develop blue wings' and the evaluation of 'Raising Bricofly larvae in a warm, dry tank causes them to develop blue Wings'). This reflects the degree to which an agent sees a causal claim derived from a more compressed model as more fitting than the most compelling causal claim that can be derived from a less compressed model.

COMPRESSED-AVG(HIGH, LOW): The difference between the participant's

evaluation of *Compressed* and a uniform average of their evaluations of *High* and *Low*

(e.g., the difference between the evaluation of 'Raising Bricofly larvae in a warm tank

causes them to develop blue wings' and the average evaluation of 'Raising Bricofly larvae

in a warm, humid tank causes them to develop blue wings' and 'Raising Bricofly larvae in

a warm, dry tank causes them to develop blue wings').[14] This reflects the degree to which

an agent sees a causal claim derived from a more compressed model as more fitting, on

average, than either alternative claim that could be derived from a less compressed model.

In keeping with our hypothesis, we find that both of these difference scores are significantly

predicted primarily by the amount of information lost in compression. We regressed these

dependent variables against independent variables denoting the assigned vignette (Vignette),

whether the more compressed claim is achieved by coarsening a causal variable or removing a

background variable (Mode of Compression), and the amount of information loss inherent in

moving from the less-compressed to the more-compressed causal model in which causal claims

are embedded (Loss), as well as all possible interactions between the independent variables. The

regressions revealed that only Loss was a significant predictor of COMPRESSED-HIGH (

$\beta = -16.623, p < .001, R^2 = .083$), so that the more information that was lost in the move from a

more detailed to a more compressed causal model, the more participants preferred the claim *High*

to the claim *Compressed*. Loss was also the only significant predictor of COMPRESSED-

AVG(HIGH, LOW) ($\beta = -9.852, p < .001, R^2 = .041$). Notably, we found no evidence of a

significant interaction between Loss and Mode of Compression on these dependent variables

---

[14] Due to an error, the equation for COMPRESSED-AVG(HIGH, LOW) was pre-registered for both experiments as Evaluation of *Compressed* - .5(Evaluation of *High*-Evaluation of *Low*). However, the correct equation is Evaluation of *Compressed* - .5(Evaluation of *High*+Evaluation of *Low*).

(COMPRESSED-HIGH: $\beta = -.426$, $p = .875$, $R^2 = .083$; COMPRESSED-AVG(HIGH, LOW): $\beta = -2.106$, $p = .426$, $R^2 = .041$), nor did we find any significant interaction effects between Mode of Compression and any other independent variables.

For additional analyses based on each individual rating (*Compressed*, *High*, *Low*), see Figure 2 as well as Supplementary Materials. As a sanity check, we also analyzed the difference between participants' evaluation of *High* and their evaluation of *Low* (we label this difference 'HIGH-LOW'). As expected, only Loss was a significant predictor of HIGH-LOW ($\beta = 13.543$, $p < .001$, $R^2 = .130$), with larger values for HIGH-LOW as the probability of the effect given the cause in *High* increased (resulting in greater information loss).

In an exploratory analysis to better appreciate how any preference for compression (regardless of magnitude) varied as a function of loss, we calculated the percentage of participants who strictly preferred *Compressed* to *High*. This percentage was approximately 36% when Loss=0, 21% when Loss=.01, and 10% when Loss=.06. This shows that the percentage of participants strictly preferring a more compressed to a less compressed causal representation varies monotonically with the amount of information lost in compression.

Discussion

These results provide strong evidence in favor of the claim that participants' relative evaluations of more and less compressed causal claims are partially governed by the amount of information loss that is inherent in the more compressed causal claim. When there is no information loss, participants evaluate more compressed causal claims significantly more highly than less compressed causal claims (consistent with **H1**), suggesting that people award simplicity and penalize unnecessary complexity in their evaluation of causal claims. When information loss is moderate, there is no significant difference between participants' evaluations of more and less

compressed causal claims, suggesting that some participants prefer a compressed claim even when some information loss is inherent in compression (consistent with **H2**).

Importantly, we found that participants' pattern of evaluation of causal claims was similar across the condition in which compression was achieved by coarsening a causal variable and the condition in which compression was achieved by removing a background variable (see Supplementary Materials for additional figures by Mode of Compression). In keeping with our analysis, this suggests that the amount of information that is lost due to compression is related to evaluations of the quality of causal claims in the same way across both of these ways of compressing a causal representation of one's environment. This, in turn, is in keeping with our unified analysis of judgements of the proportionality and stability of causal claims in terms of information loss.

Note that we assume throughout our analysis of the results a semantics of causal claims in keeping with that advocated in Woodward (2003, p. 203). Woodward puts forward a permissive semantics of causal explanation according to which a (set of) causal variable(s) C taking some (set of) value(s) c explains an effect variable E taking its value e when there is *some* intervention setting C to the value(s) $c' \neq c$ such that this intervention leads to E taking some value $e' \neq e$. Though Woodward's causal semantics is stated deterministically, one can naturally render it probabilistic by requiring instead that there exists some intervention setting C to the value(s) $c' \neq c$ that confers higher probability on E taking some value $e' \neq e$ than existed when $C=c$. On that approach, Compressed (Good), High (Good) and Low (Good) are all true, and Compressed (Bad), High (Bad) and Low (Bad) are all false. We take participants' responses to follow this semantics, as participants tended to judge all three of the "Good" causal claims above the scale midpoint, and all three of the "Bad" causal claims below the scale midpoint. Thus, we claim that the interpretation

of the data generated by our participants is one in which differences in the aptness of asserting each of the three "Good" causal claims are tracking differences in aptness among claims that are already taken as true. While this interpretation is most consistent with our theoretical framework, we recognize that it is under-motivated by our current results, and leave its full defense as an empirical theory of causal semantics to future work.

Experiment 2

In Experiment 1, participants evaluated the three key causal claims (*Compressed*, *High*, and *Low*) on the same screen. This could have introduced unintended task demands. For instance, participants may have felt that endorsing *Compressed* was redundant with the endorsement of both *High* and *Low*, or that endorsing *Compressed* (when the option to select more fine-grained options was available) implied the causal irrelevance of the unspecified factor. To ensure that the results of Experiment 1 were robust to such considerations, we replicated the study with the amendment that participants were shown the same data twice, and first asked to evaluate *Compressed*, and only asked to evaluate *High* and *Low* after their evaluation of *Compressed* was completed.

Participants

Participants were 483 adults recruited via Prolific. An additional 117 participants were excluded for failing comprehension checks or rating poor causal claims non-negatively. The sample of participants was 50.1% female and 48.6% male, with an age range of 19-81 and a mean age of 38.

Materials and Procedures

The procedure was identical to that used in Experiment 1, with three exceptions. First, as described above, participants completed their evaluation of *Compressed* before being asked to evaluate *High* and *Low*. Second, sentence (b) in both descriptions used in the first experiment was

amended to replace 70% with 55%. Analogous replacements were made for the other two vignettes. Third, the value of *x* in sentences (a) and (b) was varied between participants and was set at either 55, 85, or 98, leading to information loss amounts of 0, .04, and .11 respectively, under the assumption of a uniform distribution over possible interventions on causes in the less-compressed representation of the data-generating process. Thus, Experiment 2 replicates Experiment 1 for a different range of loss values.

Results

Figure 3

Evaluations of Causal Claims as a Function of Information Loss
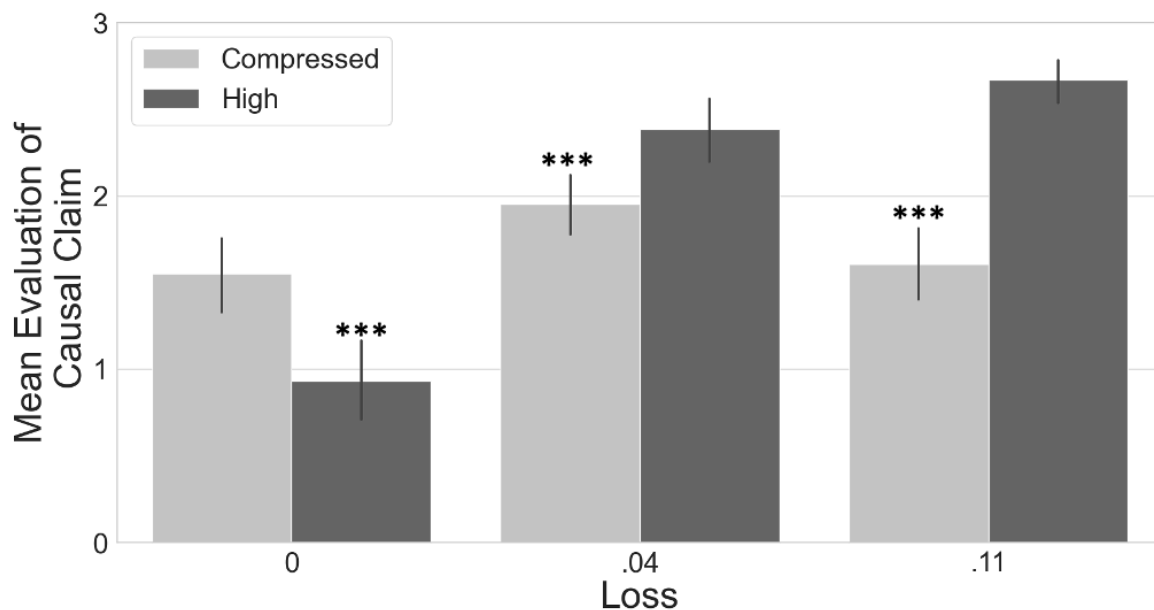
Note. Bar plots showing mean evaluations in Experiment 2 (with 95% confidence intervals) of causal claims under different loss conditions for all participants. Triple asterisks indicate significant within-participants differences at the .001 level. Mixed ANOVA for each value of Loss found that at Loss=0, Compressed was rated more highly than High ($\eta^2 = .044$, $p < .001$). At Loss=.04, High was rated more highly than Compressed ($\eta^2 = .036$, $p < .001$). At Loss=.11, High was rated more highly than Compressed ($\eta^2 = .179$, $p < .001$).

Figure 3 presents the mean ratings for each of *Compressed*, *High*, and *Low* as a function of Loss. We performed the same regressions as in Experiment 1. Once again, we find that the

amount of information lost in compression was a significant predictor of all three dependent variables (COMPRESSED-HIGH: $\beta = -14.543$, $p < .001$, $R^2 = .213$; COMPRESSED-AVG(HIGH, LOW): $\beta = -6.391$, $p < .001$, $R^2 = .064$; HIGH-LOW: $\beta = 16.303$, $p < .001$, $R^2 = .296$). Thus, as the amount of information lost due to compression increased, *Compressed* was again evaluated more negatively in comparison with more detailed causal claims. In addition, we again saw larger values for HIGH-LOW as the probability of the effect given the cause in *High* increased (resulting in greater information loss). For additional analyses based on each individual rating (*Compressed*, *High*, *Low*), see Figure 3 and Supplementary Materials.

In a further replication of Experiment 1, the manner in which compressed causal claims were generated (i.e., either by coarsening a causal variable or removing a background variable) was not a significant predictor of any of the three dependent variables measured, nor did it interact with Loss (Regression statistics for the interaction between Mode of Compression and Loss: COMPRESSED-HIGH: $\beta = -1.803$, $p = .171$, $R^2 = .213$; COMPRESSED-AVG(HIGH, LOW): $\beta = -1.172$, $p = .327$, $R^2 = .064$; HIGH-LOW: $\beta = 1.263$, $p = .281$, $R^2 = .296$; see Supplementary Materials for additional figures broken down by Mode of Compression). This again suggests that measures of information loss in causal model compression provide a unifying account of the value of both proportional and stable causal claims.

In an exploratory analysis, we measured the percentage of participants who strictly preferred *Compressed* to *High* across all three loss levels. This percentage was approximately 39% when Loss=0, 10% when Loss=.04, and 2% when Loss=.11.

Discussion

The results of Experiment 2 replicate the positive results of Experiment 1 for a different range of loss levels, and under conditions such that *Compressed* was evaluated separately from

High and *Low*. This renders the aforementioned concerns about task demands less plausible. Nevertheless, our results in Experiments 1 and 2 remain subject to two salient concerns. First, Experiments 1-2 varied proportionality by adding or omitting qualifiers to a variable (e.g., warm *humid* tank versus warm tank). One might worry that tank temperature and tank humidity are better treated as separate variables in this case. More canonical manipulations of proportionality involve a *continuum* that can be coarsened into discrete ranges (e.g., a scale with ten values that is coarsened into two ranges of values). Thus, the results of Experiments 1-2 leave open the possibility that these more canonical manipulations of proportionality would show divergence between effects of information loss on proportionality and stability. Second, Experiments 1 and 2 were designed to provide positive support for H1 and H2, but were not designed to differentiate our account from another alternative hypothesis: that evaluations of more and less compressed causal claims do not reflect information loss, as our account suggests, but instead differences in *causal contrast* (consistent with Lien and Cheng, 2000). Experiment 3 was designed to addresses both of these concerns.

Experiment 3

Lien and Cheng (2000) develop an account of how people differentiate between genuine and spurious causes, and in so doing present results in keeping with the claim that when agents choose between candidate causal explanations of a given event, they choose the one that maximizes *causal contrast*, which is given by the following equation:[15]

$$\text{Cont}(e; c) = p(e \mid do(c)) - p(e \mid do(\neg c)).$$

---

[15] We have re-written Lien and Cheng's contrast measure in terms of Pearl's do-calculus. In the original formulation, causal contrast is written as $\text{Cont}(e; c) = p(e \mid c) - p(e \mid \neg c)$. Although not cited in Lien and Cheng's paper, the causal contrast equation is very similar to Allan's ΔP_C measure (Allan, 1980).

Applying Lien and Cheng's formula to the values from Experiments 1 and 2 reveals that as the probability of the effect for *High* increases, the difference in contrast between the compressed causal claim and the high causal claim decreases (see Table 2 for calculations demonstrating this trend). So, it seems that our results in Experiment 1-2 might just as well be explained by the hypothesis that participants are basing their judgments on the difference in contrast between *Compressed* and *High* as they are by our hypothesis that participants are balancing compression against information loss. To distinguish between these two hypotheses, Experiment 3 used a similar paradigm to Experiments 1-2, but where participants were shown data sets for which information loss and causal contrast generated different qualitative predictions. This design allows us to test which of these two quantities is a more plausible candidate for the cue that participants are using to evaluate causal claims. Experiment 3 also differed from Experiments 1-2 in manipulating proportionality through coarsenings of a continuous quantity.

Participants

Participants were 458 adults recruited via Prolific. An additional 185 participants were excluded for failing comprehension checks or rating poor causal claims non-negatively. The sample of participants was 49.9% female and 50.1% male, with an age range of 18-79 and a mean age of 37.

Table 2

Information loss and causal contrast for three different scenarios shown in Experiments 1 and 2, in the proportionality conditions.

Tank Condition	% of Bricofly Developing Blue Wings	% of Bricofly Developing Blue Wings	% of Bricofly Developing Blue Wings
Warm, Humid Tank	70%	85%	98%
Warm, Dry Tank	70%	70%	70%
Cold, Humid Tank	1%	1%	1%
Cold, Dry Tank	1%	1%	1%
Loss	0	.01	.06
Contr(Blue Wings; Warm Tank)	.7 - .01 = .69	.5[.85 + .7] - .01 = .765	.5[.98 + .7] - .01 = .83
Contr(Blue Wings; Warm, Humid Tank)	.7 – (1/3)[.7+.01+.01] = .46	.85 – (1/3)[.7+.01+.01] = .61	.98 – (1/3)[.7+.01+.01] = .74
Contr(Blue Wings; Warm Tank) - Contr(Blue Wings; Warm, Humid Tank)	.69 - .46 = .23	.765 - .61 = .155	.83 - .74 = .09

Materials and Procedures

Participants read a vignette in which they learned about a novel causal system, including the results of experiments involving that system. As in Experiments 1 and 2, the fictional experiments involved either insects, mushrooms, or rocks. For example, in the insect vignette, participants assigned to the proportionality condition were presented with one of the data scenarios shown in Table 3, and asked to evaluate the following three causal claims on a scale from -3 to 3:

[*Compressed*:] Raising Bricofly larvae in a moderate-temperature tank causes them to develop blue wings.

[*High*:] Raising Bricofly larvae in a moderately warm tank causes them to develop blue wings.

[*Low:*] Raising Bricofly larvae in a moderately cold tank causes them to develop blue wings.

The causal claims evaluated in the stability case were identical to those evaluated in Experiments 1 and 2, only with "moderate-temperature tank" replacing "warm tank" in the insect vignette, and a similar substitution made in other vignettes. Table 3 also shows the values of both information loss and the difference in contrast between *Compressed* and *High* for all three data sets. As can be seen from the table, Scenarios 2 and 3 both differ from Scenario 1 by the same amount with respect to the difference in contrast between *Compressed* and *High*, but only Scenario 2 differs from Scenario 1 with respect to information loss. Thus, if we believe that information loss and not causal contrast is affecting participants' evaluations of causal claims, then we would predict that participants will treat Scenarios 1 and 3 similarly, but treat Scenario 2 differently from both Scenarios 1 and 3.

Table 3

Information loss and causal contrast for three different scenarios shown in Experiment 3.

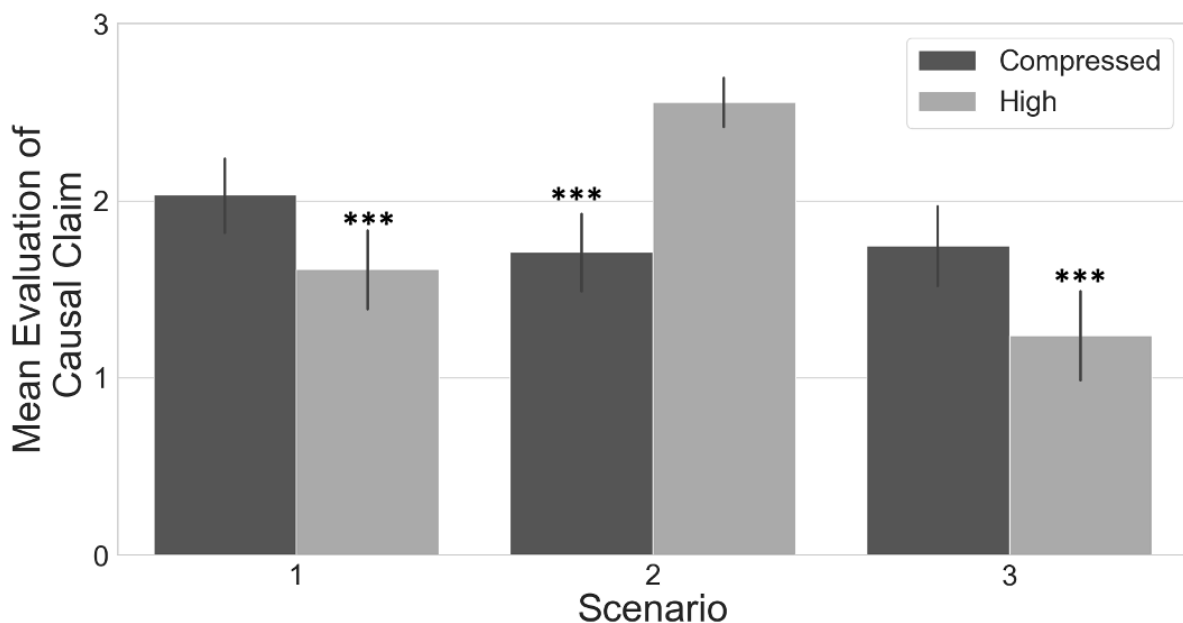| Scenario 1		Scenario 2		Scenario 3	
Tank Condition	**% of Bricofly Developing Blue Wings**	**Tank Condition**	**% of Bricofly Developing Blue Wings**	**Tank Condition**	**% of Bricofly Developing Blue Wings**
Extremely Cold Tank (0-24 degrees)	1%	Extremely Cold Tank (0-24 degrees)	1%	Extremely Cold Tank (0-24 degrees)	43%
Moderately Cold Tank (25-49 degrees)	70%	Moderately Cold Tank (25-49 degrees)	70%	Moderately Cold Tank (25-49 degrees)	70%
Moderately Warm Tank (50-74 degrees)	70%	Moderately Warm Tank (50-74 degrees)	98%	Moderately Warm Tank (50-74 degrees)	70%
Extremely Warm Tank (55-99 degrees)	1%	Extremely Warm Tank (55-99 degrees)	1%	Extremely Warm Tank (55-99 degrees)	43%
Loss	0	**Loss**	.06[16]	**Loss**	0
Contr(Comp)-Contr(High)	.23	**Contr(Comp)-Contr(High)**	.09	**Contr(Comp)-Contr(High)**	.09

---

[16] In the pre-registration for Experiment 3, this figure is mistakenly given as '.31.'

Results

Figure 4 shows the results of Experiment 3 for all three scenarios across both modes of compression. As we were primarily concerned with differential evaluations of *Compressed* and *High* across different scenarios, we ran mixed ANOVA for the within-participants difference between *Compressed* and *High* in each scenario. We found that in Scenarios 1 and 3, *Compressed* was strictly preferred to *High* (Scenario 1: $\eta^2 = .019, p = .001$; Scenario 3: $\eta^2 = .032, p < .001$). This is consistent with **H1** (since participants favored compression when information loss was zero). Moreover, the preference for *Compressed* over *High* did not differ across Scenarios 1 and 3 ($\beta = .041$, $p = .637$), which is consistent with the predictions of information loss, but not those of causal contrast (see Table 3).

Figure 4

Mean evaluations of causal claims for three scenarios in Experiment 3.

Note. Triple asterisks indicate significant within-participants differences at the .001 level, while double asterisks indicate significant within-participants differences at the .01 level.

Unlike Scenarios 1 and 3, in Scenario 2, *High* was strictly preferred to *Compressed* ($\eta^2 = .120, p < .001$). This is consistent with the hypothesis that compression trades off with information loss, such that less compressed causal claims may be favored when information loss is not negligible. Moreover, the difference in ratings between *Compressed* and *High* differed across Scenarios 2 and 3 ($\beta = -22.276$, $p < .001$, $R^2 = .232$); this is consistent with the predictions of information loss, but not with those of causal contrast (see Table 3). Notably, we do see a significant effect of the proportionality/stability condition on the difference in evaluations between *Compressed* and *High* in Scenario 1 ($\eta^2 = .031, p = .006$), but not for Scenarios 2 and 3, suggesting that the finding for Scenario 1 is likely spurious. A full report of all pre-registered analyses for Experiment 3 is provided in the Supplemental Materials.

Discussion

Although Lien and Cheng (2000)'s causal contrast theory was developed as an account of how people differentiate between genuine and spurious causes, rather than how people determine a level of compression at which to represent the causal structure of their environment, causal contrast nevertheless offers a natural alternative to our own account of information loss in explaining why people might favor more or less compressed causal claims. Experiment 3 was designed to provide a direct test of the predictions of information loss versus those of causal contrast in explaining judgments like those elicited in Experiments 1-3. The results provide clear support for information loss: differences in information loss (holding differences in causal contrast fixed) predicted different patterns in ratings, while differences in causal contrast (holding information loss fixed) did not. Moreover, because our proportionality cases in Experiment 3 involve coarsenings of a continuous quantity (e.g., temperature), the results bolster our claim that

the information loss framework provides a unified account of evaluations of both proportionality and stability.

Having established our first key hypotheses (**H1** & **H2**) – that people treat both informativeness and compression as positive features of causal claims, to be traded off against one another – we turn to our third key hypothesis (**H3**) – that agents will tolerate information loss for the sake of compression whenever the lost information is not decision-theoretically valuable to that agent in a given context.

Decision Theory and The Value of Lost Information

The results of Experiments 1-3 demonstrate that the strength of participants' preference for more compressed causal claims over less compressed claims is predicted by the amount of information loss achieved by moving from a model in which the less compressed claim is embedded to a model in which the more compressed claim is embedded. However, these results also show that when the total amount of information loss is low, the preference for more detailed causal claims over less detailed ones is not significant. Additionally, they show that even as the total amount of information loss increases, mean evaluations of compressed causal claims remain positive. This suggests that, all things considered, people assign value to *lossy compressions* of a causal model, even when they have the opportunity to endorse a lossless compression of the same underlying data, where a "lossless" compression is understood as one that preserves all four variable values.

What explains the value that people assign to compressed causal claims that elide information about the underlying dynamics of the processes that they represent? As stated in the introduction, our hypothesis is that agents' evaluations of causal claims that compress their target systems with information loss are driven at least in part by agents' judgments as to the *decision-*

theoretic value of the information that is lost in compression. That is, when agents judge that the information that is lost in the move to a more compressed causal model is *not relevant* to their choice between a set of feasible actions, they evaluate the compressed causal claim more positively than they would if the lost information *were* relevant to their choice between a set of actions.

To illustrate, consider an agent who is tasked with stocking headache medicines at a pharmacy. These medicines can have one of two ingredients (Reptol or Psylo), each of which comes in one of two types (Type-1 or Type-2). The agent wants to stock all and only those headache medicines that relieve headaches in *y%* of patients, and has access to the data in Table 4. Suppose that the agent is in the proportionality condition. If *y*=50, such that they would stock any medicine that reduces severity of headaches in at least 50% of patients, then whether a Reptol-based or Psylo-based medicine is of Type-1 or Type-2 is irrelevant to their decision; they will stock Reptol-based medicines and not Psylo-based medicines. Thus, the compression from 'Type-1 Reptol causes reduced severity of headaches' to 'Reptol causes reduced severity of headaches' incurs no loss of decision-theoretically valuable information for this agent; it elides only decision-irrelevant information about the type of Reptol, and we would expect the agent to be indifferent between the two claims if asked to evaluate their aptness in describing the efficacy of headache medicines more generally, or to show a preference for the more compressed claim. By contrast, if *y*=80, such that the agent would stock any medicine that reduces severity of headaches in at least 80% of patients, and if *x*>80, then only Type-1 Reptol should be stocked, according to the agent's own preferences. Thus, the compression from 'Type-1 Reptol causes reduced severity of headaches' to 'Reptol causes reduced severity of headaches' *does* elide decision-relevant information about the type of Reptol. Under this condition, we would therefore expect an agent to

evaluate the less compressed causal claim more positively than the more compressed causal claim.

We use this motivating example as part of our materials in Experiment 4.

Table 4

Data shown to participants in Experiment 4.

Medication	Facts about Medication (Proportionality Condition)	Facts about Medication (Stability Condition)	% of Patients with Reduced Severity of Headaches
A	Active ingredient Type-1 Reptol	Active ingredient Reptol and taken with food	x% (x > 80)
B	Active ingredient Type-2 Reptol	Active ingredient Reptol and taken without food	70%
C	Active ingredient Type-1 Psylo	Active ingredient Psylo and taken with food	1%
D	Active ingredient Type-2 Psylo	Active ingredient Psylo and taken without food	1%

Mathematically, we can define the value of the information lost when moving from a less compressed causal representation to a more compressed representation as follows. We begin with a set of variables that can be partitioned into a singleton set containing an **action variable** A, a singleton set containing an **effect variable** E, and a subset of **observable variables** $\boldsymbol{O}$. Let $u:R_A \times R_E \to \mathbb{R}$ be a real-valued utility function defined on the range of the action variable A and the effect variable E. Each value $u(a,e)$ of this function represents the utility, to some agent, of setting A to a particular value a (i.e., performing a particular action) when E takes a particular value e. The expected utility of an action a is defined as follows:

$$\mathbb{E}(u \mid a) = \sum_e p(e \mid do(a))\, u(a,e)$$

The expected utility of an action a, given an observation that the variables in $\boldsymbol{O}$ take the set of values $\boldsymbol{o}$, is given by the equation:

$$\mathbb{E}(u \mid a; \boldsymbol{o}) = \sum_e p(e \mid do(a), \boldsymbol{o})\, u(a,e)$$

where $p(e \mid do(a), o) = \dfrac{p(e, o \mid do(a))}{p(o \mid do(a))}$. With these two equations in hand, we define the

decision-theoretic value of the information contained in the observable variables O as follows:

$$VOI_u(O) = \sum_o p(o) \max_a \mathbb{E}(u \mid a; o) - \max_a \mathbb{E}(u \mid a)$$

That is, the value of the information contained in a set of observable variables is the average

difference between the maximum utility an agent can expect when they have observed the value

of all observable variables and the maximum utility that agent can expect when they have not made

any observations. This is the standard decision-theoretic notion of value of information as defined

by Blackwell (1953) and Good (1960).

Equipped with this definition, consider a case in which we move from a less compressed

set of observable variables O (such as whether a medicine contains Type-1 Reptol, Type-2 Reptol,

Type-1 Psylo, or Type-2 Psylo) to a more compressed set of observable variables $\widehat{O}$ (such as

whether a medicine contains Reptol or Psylo) for a fixed action variable A and an effect variable

E. We can now calculate the decision-theoretic **value of lost information** (VOLI) with respect to

this move, for an agent with utility function $u : R_A \times R_E \to \mathbb{R}$:

$$VOLI_u(O, \widehat{O}) = VOI_u(O) - VOI_u(\widehat{O})$$

This quantity tells us how costly the compression from O to $\widehat{O}$ is for an agent whose preferences

over values of A and E are given by the utility function u.

There is an important connection between our measure of the value of lost information and

our measure of overall information loss. Specifically, for a natural formalization of an agent whose

sole goal is to guess the correct distribution over an effect variable, the value of lost information

in a compression is equal to the total amount of lost information involved in compression. To see

this, consider an agent for whom the range of their action variable A consists of all possible

probability distributions over an effect variable E, with each distribution in A representing the act

of making a prediction about the probability of each value of E obtaining. The agent's utility

function is such that for any action a and value e of E, $u(a,e) = \log_2 \frac{a(e)}{p(e)}$. This utility function

rewards the agent for assigning high probability to E taking the value e when this actually occurs,

with greater reward when said occurrence was unlikely according to the marginal distribution over

E. Now consider a set of causal variables C with compression $\widehat{C}$. For such an agent,

$VOLI_u(C, \widehat{C}) = \mathscr{L}(C, \widehat{C}, E)$.[17] So, for the special case of an agent who only aims to make correct

predictions about E, the value of lost information just is the total amount of information lost.

In this light, participants' responses in Experiments 1-3 can be interpreted as reflecting a

trade-off between: i) satisfying the decision-theoretic goal of making correct predictions about E,

and ii) maintaining a compressed causal representation of the system under study. However, in the

following experiments we deliberately place agents in decision scenarios in which their goals

involve more than solely guessing the correct distribution over the effect variable, to examine the

more general role of the value of lost information in agents' evaluations of more and less

compressed causal claims. Thus, what emerges from Experiments 1-3 is a general account

according to which people evaluating more and less compressed causal representations engage in

a two-way tradeoff between compression and maintaining decision-theoretically valuable

information, but where, in at least some cases, peoples' decision-theoretic goals include making

accurate predictions (and therefore show sensitivity to information loss).

---

[17] To see this, consider the quantity $\sum_c q(c) \, max_{a \in A} \sum_e p(e|do(c)) \, u(a,e)$. The quantity

$\sum_e p(e|do(c)) \, u(a,e)$ is maximized when $a(e) = p(e|do(c))$ for all e. Thus,

$\sum_c q(c) \, max_{a \in A} \sum_e p(e|do(c)) \, u(a,e) = \sum_{c,e} q(c) \, p(e|do(c)) \log_2 \frac{p(e|do(c))}{p(e)} = CME(C,E)$. The

quantity $max_a \sum_e p(e) \, u(a,e)$ is maximized at zero for $a(e) = p(e)$. Thus,

$VOI_u(C) = CME(C,E) - 0 = CME(C,E)$. Repeating for $\widehat{C}$ yields $VOI_u(\widehat{C}) = CME(\widehat{C}, E)$. Thus,

$VOLI_u(C, \widehat{C}) = CME(C,E) - CME(\widehat{C}, E) = \mathscr{L}(C, \widehat{C}, E)$.

Motivation for Experiment 4

Experiment 4 tests **H3**: when an agent is placed in a decision context, that agent's evaluations of compressed causal claims will be sensitive to whether or not the value of the information lost in compression is zero or strictly positive. Specifically, we expect agents to evaluate compressed causal claims more positively when the value of the information lost in the relevant compression is zero.

One implication of these predictions is that agents' causal representations of their environments are primarily guided by their prudential *values*. That is, agents build causal models of their environments so as to achieve a representation that: i) allows for expected-utility-maximizing interventions on their environment, and ii) encodes observations that facilitate the choice of expected-utility-maximizing actions. In this respect, our hypothesis and the results supporting it are in keeping with work on value-guided task construal by Ho et al. (2022), as well as theoretical work by Brodu (2011), Kinney (2019), and Kinney and Watson (2020), arguing that prudential factors such as an agent's interest in realizing certain values of an effect variable and the value of the information provided by a causal variable determine the overall quality of compressed causal claims. However, Ho et al.'s experimental paradigms did not test the extent to which explicitly *causal* representations are value-guided, and so our experiments are also the first to test these ideas as descriptive claims about human evaluations of causal claims.

Experiment 4

In Experiment 4, we designed vignettes with a similar structure to those used in Experiments 1-3, but which also allowed us to manipulate whether the VOLI realized in the move from a less compressed to a more compressed causal model of the data presented in the vignette was zero or strictly positive. We hypothesized that when VOLI is zero, participants would tolerate

information loss, and therefore rate causal claims embedded in a more compressed representation at least as highly as those embedded in a less compressed representation. By contrast, when VOLI is strictly positive, we hypothesized that participants would be less tolerant of information loss, since the more compressed representation fails to include information that is prudentially valuable to them, and they would therefore prefer those claims embedded in a less compressed representation. As in our earlier studies, we also hypothesized that these patterns of evaluation would hold regardless of whether compression was achieved by coarsening a causal variable or eliding a background condition.

Participants

Participants were 372 adults recruited via Prolific. An additional 408 participants were excluded for failing comprehension checks, and an additional 26 participants were excluded due to experimenter error.[18,19] The sample of participants was 49.2% female and 48.5% male, with an age range of 19-93 and a mean age of 42.

Materials and Procedures

We presented participants with the results of fictional experiments involving headache medications or training programs for new employees at a company. Participants were told that headache medications/training programs had to be shown to be effective in a certain percentage of people in order to be recommended for stocking on a pharmacy's shelves/being required for new employees. Participants were then asked to answer questions about the decisions they would make

---

[18] A bug in our code meant that these 26 participants did not receive accurate feedback on the answers that they gave to multiple choice questions that were used to exclude participants from analysis. This resulted in slightly fewer data points being collected than were pre-registered. We exclude the data produced by these participants from our analysis below for the sake of accuracy, but note that all reported significant results still hold if the data from these participants is included.

[19] Given the large number of exclusions, we repeated key analyses including all participants. Patterns of findings were similar overall; see Supplementary Materials for details.

about different headache medicines or training programs, and then asked a second set of questions about their pattern of decision making with respect to the first set of questions.

To illustrate, participants shown the vignette about headache medicines were shown the data in Table 4. The value of x was varied between participants, as was whether the first or second set of facts about each medication was shown. Participants shown these data were then told that their manager had asked them to stock any medicine that was shown to reduce the severity of headaches in at least y% of patients, where $1 < y < x$. They were then asked to answer the following yes-or-no questions:

A1. Would you stock Medication A?

A2. Would you stock Medication B?

A3. Would you stock Medication C?

A4. Would you stock Medication D?

The correct answer to A1 is always 'Yes', and the correct answer to A3 and A4 is always 'No.' Whether the correct answer to A2 is 'Yes' or 'No' depends on whether y is less than or greater than 70. This was manipulated between participants, by setting $y=50$ for some and $y=80$ for others. Participants who answered these questions incorrectly were given a second chance to answer. If they answered incorrectly again, their data were excluded from analysis. Participants were then shown a summary of the correct answers to the first set of questions, and then asked to say whether the following statements, or analogous statements for those assigned to different conditions, were true or false:

B1. Whenever a medication contains Reptol, you would stock it, regardless of whether it is Type-1 or Type-2.

B2. Whenever a medication contains Reptol, you would stock it if it is Type-1 but not if it is Type-2.

B3. Whenever a medication contains Psylo, you would *not* stock it, regardless of whether it is Type-1 or Type-2.

B4. Whenever a medication contains Psylo, you would stock it if it is Type-1 but not if it is Type-2.

The statement B1 is true when $y \leq 70$ and false when $y > 70$, B2 is true when $y > 70$ and false when $y \leq 70$, B3 is always true, and B4 is always false. Participants who answered these questions incorrectly were given a second chance to answer. If they answered incorrectly again, their data were excluded from analysis.

Participants were then asked to evaluate, on a scale from -3 (very bad) to 3 (very good), how good it would be to include the following causal claims in a summary of the data prepared for a colleague:

[*Compressed*:] Reptol causes reductions in the severity of headaches.

[*High*:] [Type-1 Reptol/Reptol taken with food] causes reductions in the severity of headaches.

[*Low*:] [Type-2 Reptol/Reptol taken with food] causes reductions in the severity of headaches.

[*Compressed (Bad)*:] Psylo causes reductions in the severity of headaches.

[*High (Bad)*:] [Type-1 Psylo/Psylo taken with food] causes reductions in the severity of headaches.

[*Low (Bad)*:] [Type-2 Psylo/Psylo taken with food] causes reductions in the severity of headaches.

See Figure 5 for a schematic representation of the compression inherent in endorsing the causal claim *Compressed* in the conditions in which compression involves coarsening a causal variable. As in Experiments 1-3, data from participants who assigned non-negative evaluations to the bad causal claims were excluded from analysis.

Between participants, we manipulated the vignette used (see Table 5 for a comparison of the pharmaceutical and employee training vignettes); the amount of information loss realized by the more compressed causal claim (by setting the value of x to either 85 or 98, resulting in information loss amounts of .01 and .06, respectively); whether compression was achieved by coarse-graining a variable, thus manipulating proportionality, or eliding a background variable, thus manipulating stability (this was done in the pharmaceutical vignette by showing participants either the first or second set of facts about each medication, respectively); and whether the decision-theoretic value of the information lost in the move from a less compressed to a more compressed causal model was zero or strictly positive (by setting the value of y to either 50 or 80, respectively).

Table 5

Structure of vignettes used in Experiment 4.

Vignette	Effect	Primary Cause	Secondary Cause	Background Condition
Pharma	Reduced Severity of headaches	Reptol/Psylo	Type-1/Type-2	Taken with/without food
Training	Becoming a successful employee	Case Studies/Simulations	Task-focused/Problem-focused	Held on weekends/weekdays

Figure 5

Graphs showing the Causal Relationships between Variables in Experiment 4

Compression achieved through coarsening a background variable (proportionality)

G_1: less compressed causal model

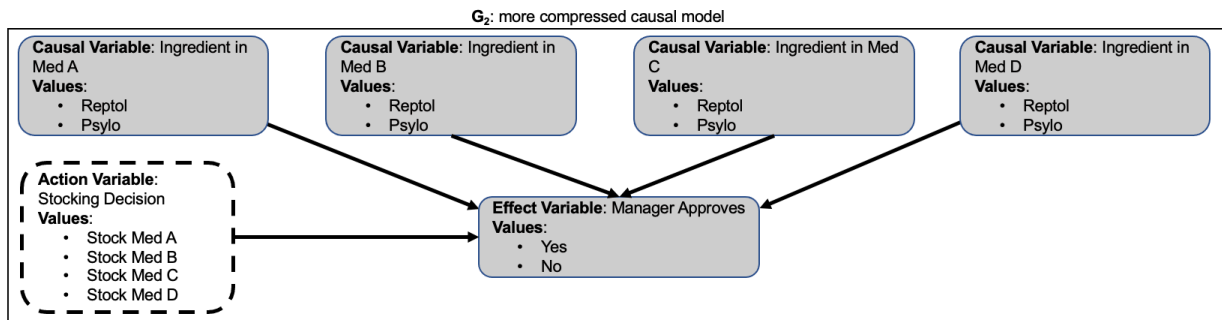| **Causal Variable**: Ingredient in Med A **Values:** • Type-1 Reptol • Type-2 Reptol • Type-1 Psylo • Type-2 Psylo | **Causal Variable**: Ingredient in Med B **Values:** • Type-1 Reptol • Type-2 Reptol • Type-1 Psylo • Type-2 Psylo | **Causal Variable**: Ingredient in Med C **Values:** • Type-1 Reptol • Type-2 Reptol • Type-1 Psylo • Type-2 Psylo | **Causal Variable**: Ingredient in Med D **Values:** • Type-1 Reptol • Type-2 Reptol • Type-1 Psylo • Type-2 Psylo |

Action Variable: Stocking Decision **Values:** • Stock Med A • Stock Med B • Stock Med C • Stock Med D

Effect Variable: Manager Approves **Values:** • Yes • No

Example claim: "Type-1 Reptol causes reduced severity of headaches."

G_2: more compressed causal model

| **Causal Variable**: Ingredient in Med A **Values:** • Reptol • Psylo | **Causal Variable**: Ingredient in Med B **Values:** • Reptol • Psylo | **Causal Variable**: Ingredient in Med C **Values:** • Reptol • Psylo | **Causal Variable**: Ingredient in Med D **Values:** • Reptol • Psylo |

Action Variable: Stocking Decision **Values:** • Stock Med A • Stock Med B • Stock Med C • Stock Med D

Effect Variable: Manager Approves **Values:** • Yes • No

Example claim: "Reptol causes reduced severity of headaches."

Note: The top panel shows the uncompressed graph used in the proportionality condition (in which the claims *High* and *Low* are embedded), and the bottom panel shows the compressed graph used in the same condition (in which the claim *Compressed* is embedded). Whether the agent's manager approves of their choice of medication to stock depends on their choice of medication and the active ingredient in that medication, as indicated in the graph. The agent can observe the active ingredients in each medication, which is in turn informative about the likelihood that their manager approves their choice of medication.

The final manipulation described above has no analog in Experiments 1-3, and so we explain it here in more detail. In the pharmaceutical case, when *y=50*, participants should recommend medications containing Reptol regardless of whether that Reptol is Type-1 or Type-2, and regardless of whether the medication must be taken with or without food. Under these conditions, the information lost in the compression is of no use to the agent's decision-making. Thus, VOLI in this case is zero. However, when *y=80*, only medications containing Type-1 Reptol/Reptol taken

with food should be stocked on the shelves, and so the VOLI for the compression is strictly positive, on the assumption that agents assign strictly higher utility to complying with their manager's directions than to failing to comply. By manipulating whether VOLI was zero or strictly positive between participants, we were able to measure the extent to which the decision-theoretic *value* of the amount of information lost, controlling for the amount of information loss itself, influenced participants' evaluations.

Results

Figure 6 shows mean evaluations of each causal claim as a function of information loss under the different VOLI conditions. These results are in keeping with our prediction that participants would prefer more compressed causal representations in cases where VOLI was zero than in cases where VOLI was positive, and that in the positive-VOLI case, any preference for compression would decrease with information loss. We first performed the same regressions as Experiments 1 and 2, with an additional binary independent variable added to represent whether the VOLI due to compression in the vignette shown to a participant was zero or strictly positive (once again, we also regressed our dependent variables on all possible interactions between all four of our independent variables). We found that COMPRESSED-HIGH, the difference between evaluations of *Compressed* and *High*, was significantly predicted by the following: i) whether VOLI was zero or strictly positive ($\beta = -.429$, $p = .001$, $R^2 = .242$), such that the difference between evaluations of *Compressed* and *High* was negligible when VOLI was zero, but much larger and negative (reflecting higher ratings for *High* than *Compressed*) when VOLI was strictly positive; ii) the amount of information loss due to compression ($\beta = -7.655$, $p = .010$, $R^2 = .242$), such that the extent to which ratings for *High* dominated ratings for *Compressed* increased as information loss increased; and iii) the interaction between whether VOLI was zero or strictly positive and the

information loss due to compression ($\beta = -6.573$, $p = .027$, $R^2 = .242$), so that when VOLI was strictly positive, increases in information loss due to compression resulted in lower values of COMPRESSED-HIGH.

Figure 6

Evaluations of Causal Claims by VOLI and Information Loss
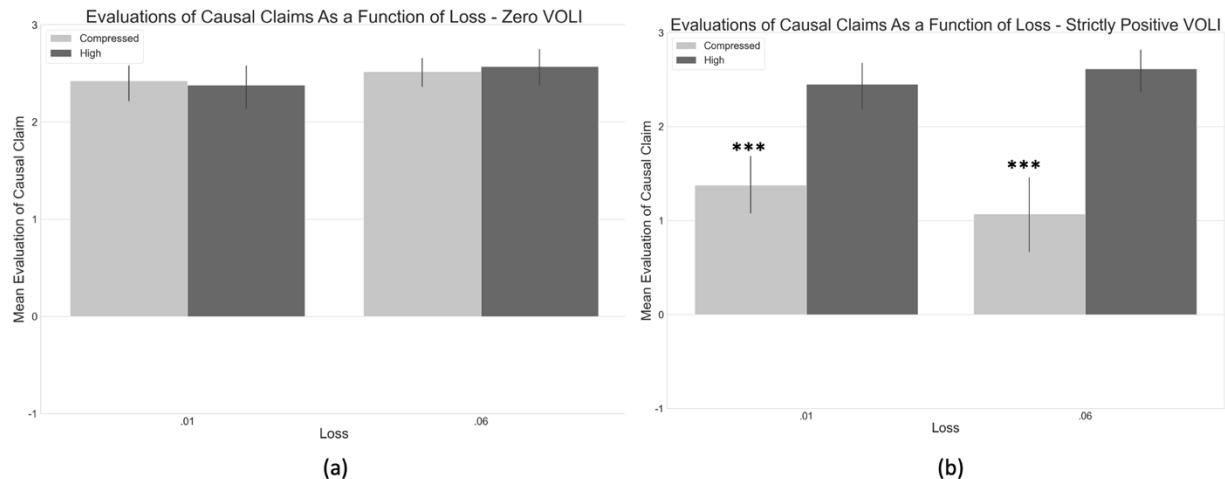
Note: Bar plots showing mean evaluations in Experiment 4 (with 95% confidence intervals) of causal claims as a function of loss for participants assigned to the zero-VOLI condition (a) and the positive-VOLI condition (b). Triple asterisks indicate significant within-participants differences at the .001 level in a mixed ANOVA. Mixed ANOVA found that when the VOLI associated with compression is zero, there was no significant difference between Compressed and High ($\eta^2 = .001$, $p = .847$). There was a significant interaction effect between: i) the between-subjects difference in whether compression was achieved by coarsening a causal variable or removing a background variable, and ii) the within-subjects difference between evaluations of Compressed and High ($\eta^2 = .011$, $p = .017$). When VOLI is strictly positive, there was a significant difference between Compressed and High ($\eta^2 = .175$, $p < .001$). There was also a significant effect of the interaction between the amount of information lost due to compression and the vignette shown to participants ($\eta^2 = .017$, $p = .047$), and a significant interaction effect between the between-subjects difference in information loss due to compression and the within-subjects difference between evaluations of Compressed and High ($\eta^2 = .015$, $p = .005$).

We followed up this interaction with independent tests of each VOLI condition. This revealed that when we restricted our analysis to just those cases in which VOLI is strictly positive, the sole significant predictor of COMPRESSED-HIGH is the amount of information lost in compression ($\beta = -11.809$, $p = .015$, $R^2 = .069$); by contrast, in those conditions for which VOLI

was zero, none of the independent variables were significant predictors of COMPRESSED-HIGH. Finally, the original regression also revealed an interaction between: i) whether VOLI was zero or strictly positive, ii) the amount of information loss due to compression, and iii) the vignette shown to a participant ($\beta = -6.847, p = .021,\ R^2 = .242$). When VOLI was zero, participants who were shown the Pharma vignette tended to assign lower values to COMPRESSED-HIGH than those who were shown the training vignette. When VOLI was strictly positive, this relationship was reversed; participants who were shown the Pharma vignette tended to assign higher values to COMPRESSED-HIGH than those shown the Training vignette. As in Experiments 1-2, we found that the interaction between information loss due to compression, VOLI, and whether compression was achieved by coarsening a causal variable (proportionality) or removing a background variable (stability) was not a significant predictor of any of our dependent variables (COMPRESSED-HIGH: $\beta = -.200,\ p = .946,\ R^2 = .242$).

See the Supplemental Materials for the full pre-registered findings of Experiment 4.

Discussion

Experiment 4 found support for **H3**: When participants were asked to select a causal representation in the context of a particular decision problem, their tolerance for information loss was moderated by the decision-theoretic value of the information that was lost. Specifically, when the value of information lost in moving to a more compressed representation was zero, they rated the compressed claim (*Compressed*) and the less compressed claim associated with the highest probability (*High*) similarly, regardless of how much information loss was associated with the more compressed claim. But when the value of information lost through compression was strictly positive, participants favored the less compressed claim associated with the highest probability (*High*) over the more compressed claim (*Compressed*), and additionally showed sensitivity to the

amount of information lost, with the compressed claim more strongly disfavored when it was associated with greater information loss. This suggests that agents do not merely consider whether lost information has strictly positive value or zero value when evaluating claims embedded in more or less compressed causal models, but instead continuously trade off an all-things-considered preference for compression against the decision theoretic value of the information lost in compression, where there is an agential goal both to have enough information to accomplish a task and to be as predictively accurate as possible.

Our finding that manipulating the decision context in which an agent finds themselves changes whether they prefer a more or less compressed representation of the same data helps to address a concern about our earlier experiments. Namely, in Experiment 1, one might think that some participants prefer a more compressed representation of the data in the Loss=.01 condition because they judge that a 15% difference in the probability of blue wings across tank humidity conditions is not statistically significant. If this were the case, then one should see the same pattern of preferences in this condition regardless of whether the information lost in compression is decision-theoretically valuable. That we do not see this suggests that participants do tend to think that the observed difference in probabilities is statistically meaningful, and that this statistical difference matters in cases where it can be exploited to make better decisions.

Regressions on the individual evaluations of each causal claim show a contrast between the effect of different amounts of information loss in Experiments 1 and 2 and the effect of different values for VOLI in Experiment 4. In Experiments 1 and 2, changes in the amount of information loss involved in compression mostly led to higher evaluations for *High*, without significant changes to evaluations of *Compressed*. By contrast, manipulations of VOLI in Experiment 4 largely led to changes in evaluations of *Compressed*, without significant changes in the value of

High. This suggests that when more information is lost in compression, agents avoid that information loss by ensuring that they include detailed causal claims in their descriptions of a data-generating process. However, when the lost information is not decision-theoretically valuable to them, they do not downgrade their evaluations of more detailed causal claims, but instead upgrade their evaluations of causal claims embedded in a more compressed causal model.

Experiment 5

Experiments 1-4 offer support for hypotheses H1-H3. However, in all four of these studies participants were explicitly asked to say how good or bad it would be to include a given claim in a summary of a data-generating process to be presented to a colleague. This raises the possibility that our results are not explained by a general trade-off between compression and information loss in causal representation, but by the pragmatics of linguistic communication. The most plausible candidate to fill this role is Grice's maxim of quantity (Grice, 1975), according to which a speaker should strive to be as informative as possible without providing any unnecessary information. In fact, we regard such maxims as highly amenable to our framework: much of what we have claimed can be seen as a formalization of Gricean ideas, which we return to in the General Discussion. Nonetheless, our own claims are more general insofar as they pertain to *causal representation* generally, not only interpersonal, linguistic communication. For this reason, in a follow up to Experiment 4, we asked participants to describe what they themselves had learned, rather than selecting the most apt summary to present to another person. We sought to test whether participants would, as predicted by our account, produce more compressed representations when VOLI was zero and the amount of information loss inherent in compression is low. Experiment 5 also goes beyond prior experiments in requiring participants to generate their own representation, rather than select between provided options.

Participants

Participants were 400 adults recruited via Prolific. An additional 20 participants were excluded for failing comprehension checks.[20]

Materials and Procedures

As in Experiment 4, we presented participants with the results of fictional experiments involving headache medications or training programs for new employees (see Table 5), making only slight changes to the structure of the training vignette in some conditions. Participants were also told that their manager had instructed them to only stock a medicine/recommend a certain program if its effectiveness was above a threshold y. This enabled us to perform all of the same between-participants manipulations of the scenarios as in Experiment 4. Participants were then asked questions A1-A4 from Experiment 4. Participants who answered these questions incorrectly were given a second chance to answer and were excluded if they answered incorrectly a second time. Participants were then asked: "please summarize in a few sentences what you have learned about the efficacy of [active ingredients in headache medicines/different kinds of training programs]." Below this prompt they were presented with a text field that required them to input at least fifty characters.

Text responses were coded by two independent researchers, according to a pre-registered rubric. The four codes that coders could assign, along with a typical example of the kind of response that would receive each code, is shown in Table 6. (For the full coding rubric, see the pre-registration.) Note that it was possible for a response to receive none of these codes (e.g., if a

---

[20] Unlike in Experiment 4, we did not ask participants to state whether statements B1-B4 were true or false, with participants who answered incorrectly excluded. This decision is motivated by the finding, in Experiment 4, that this exclusion criterion led to a high number of exclusions without changing the overall pattern of results.

participant responded by saying 'I did not learn anything,' then their response did not receive any code). It was also possible for a response to receive multiple codes.

The coders were in agreement with respect to the appropriate codes to assign to 69.75% of all responses. Disagreements were resolved through discussion. As pre-registered, we report here estimates of Cronbach's measure of reliability (i.e., Cronbach's alpha), which were: .589 (95% CI: [.500, .663]) for judgements as to whether a response warranted Code 1, .796 (95% CI: [.751, .832]) for judgements as to whether a response warranted Code 2, .775 (95% CI: [.726, .815)) for judgments as to whether a response warranted Code 3, and .840 (95% CI: [.805, .868]) for judgments as to whether a response warranted Code 4.

Table 6

Codes assigned to open-ended responses in Experiment 5, with typical instances of each code.

Code	Gloss	Typical Instance
Compressed-Implicit	Captures a more compressed representation, where collapsing across sub-types (e.g., Type 1 and Type 2 Reptol) is implicit.	"Reptol causes reductions in the severity of headaches."
Compressed-Explicit	Captures a more compressed representation, where collapsing across sub-types (e.g., Type 1 and Type 2 Reptol) is explicit.	"Reptol of either type is effective in reducing the severity of headaches" or "Reptol is effective at reducing the severity of headaches whether it is taken with or without food"
DIFF	Captures a less compressed representation by differentiating between sub-types (e.g., Type 1 and Type 2 Reptol).	"Type-1 Reptol is more effective than Type-2 Reptol" or "Reptol taken with food is more effective than Reptol taken without food."
High	Captures a less compressed representation by focusing on the sub-type that maximizes the probability of the effect the between sub-types (e.g., Type 1 Reptol).	"Type-1 Reptol causes reductions in the severity of headaches" or "Reptol causes reductions in the severity of headaches when taken with food."

Results

Figure 7

Proportion of Responses Receiving Each Code by VOLI Condition
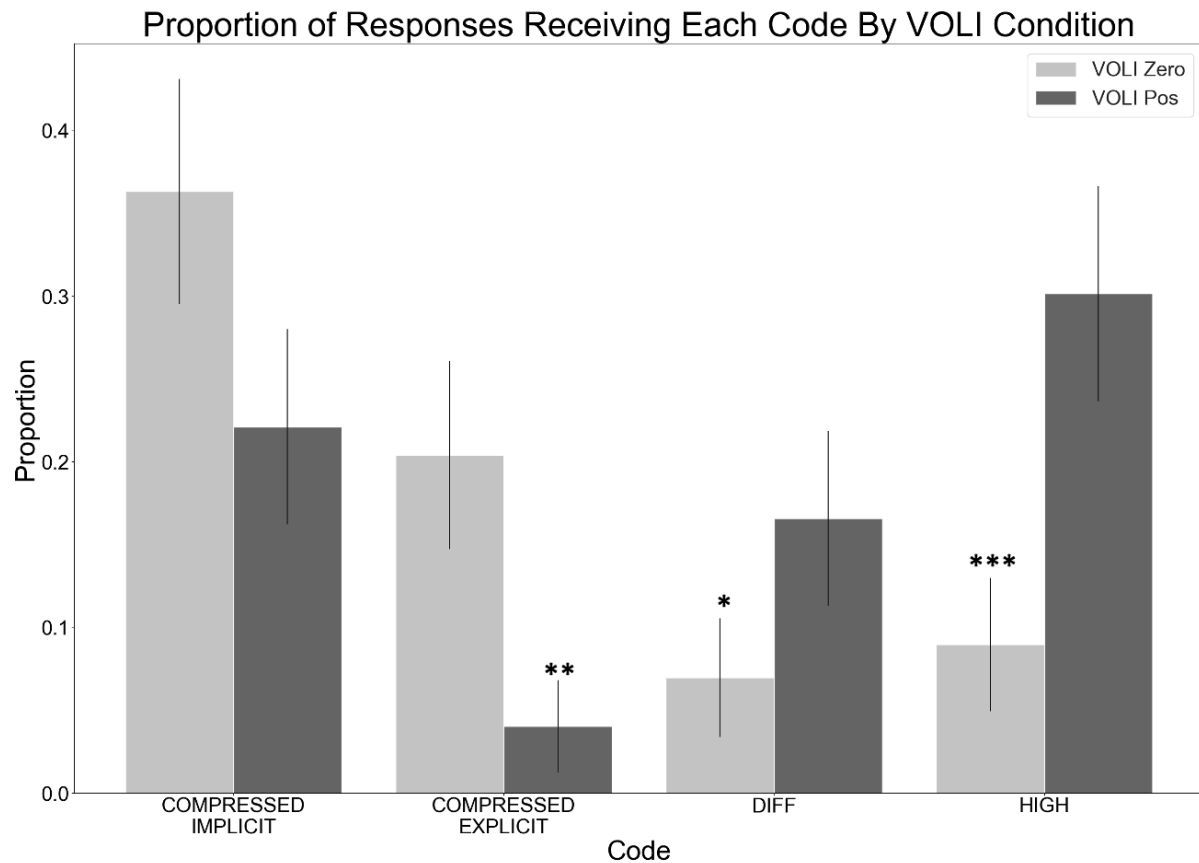
Note. Proportion of participant responses receiving each code in each VOLI condition, with 95% confidence intervals. A single asterisk represents significance at the 5% level, and double asterisks represent significance at the 1% level.

Figure 7 shows the proportion of responses receiving each of the four codes described above under the condition that VOLI is zero or that VOLI is strictly positive. We constructed four binary variables representing whether or not a response was assigned each of the four codes, and performed logistic regressions for each variable against independent variables representing: i) whether VOLI was zero or strictly positive, ii) the amount of information loss inherent in compression, iii) the vignette shown to participants, iv) whether compression was achieved by

coarsening a causal variable or removing a background variable (proportionality vs. stability), and

v) all interaction terms between all independent variables.

The binary variable representing whether the Value of Lost Information is zero or strictly

positive was the only significant predictor of: i) whether a response received the code

COMPRESSED-EXPLICIT ($\beta = -.909$, $p = .014$),[21] ii) whether a response received the code

DIFF ($\beta = .735$, $p = .023$),[22] and iii) whether a response received the code HIGH ($\beta = .735$,

$p = .006$). There were no other significant effects (see Table 7). Overall, participants generated

more compressed responses when VOLI was zero (vs. strictly positive), and more DIFF and HIGH

responses when VOLI was strictly positive (vs. zero).

Table 7

Logistic regression coefficients and p-values for the independent variables representing the mode of compression in the vignette shown to a participant (proportionality versus stability) and information loss inherent in compression, for each of the four binary coding variables.

	COMPRESSED-IMPLICIT		COMPRESSED-EXPLICIT		DIFF		HIGH	
Variable	β	p	β	p	β	p	β	p
Mode of Compression	-.257	.210	.359	.268	-.227	.502	-.354	.186
Information Loss	.217	.964	-.876	.921	-7.827	.334	-6.417	.317

---

[21] The three-way interaction between the three independent variables representing the vignette shown to participants, the amount of information loss inherent in compression, and the decision-theoretic value of lost information was left out of this regression, since including it leads to perfect separation and a failure of logistic regression.
[22] The four-way interaction between all independent variables was left out of this regression, since including it leads to perfect separation and a failure of logistic regression.

Table 7 shows logistic regression coefficients and *p*-values for the logistic regressions for each of the four coding variables for both the independent variable representing whether compression is achieved by coarsening a causal variable or eliding a background condition and the variable measuring the amount of information loss inherent in compression. In all cases, we see no significant relationships.

Discussion

The results of Experiment 5 replicate the finding from Experiment 4 that manipulating the decision-theoretic value of the information that is lost when moving from a more detailed to a more compressed causal representation of some data-generating process changes agents' propensity to generate more or less compressed representations of that process. When the decision-theoretic value of the information lost in compression is zero, participants tend to give more compressed representations of the data-generating process that do not make fine-grained distinctions between the efficacy of different possible interventions on that system, *even when those distinctions are explicitly present in the data they are asked to analyze*. By contrast, when the decision-theoretic value of the information lost in compression is positive, participants tend to give more fine-grained accounts of what they learned about a data-generating process, holding fixed all other factors about the data that they are shown. However, unlike in Experiment 4, we do not find a significant effect of the *amount* of information lost on participants' responses when the value of the information lost is zero. As in previous experiments, we did not see a significant effect of how compression was achieved (i.e., proportionality versus stability) on the types of causal representations that participants produced in an open-ended setting. This suggests once again that judgments of both proportionality and stability can be understood in terms of the value of the

information lost in a compression of one's causal model of the environment, in relation to a particular decision context.

Note further that because participants did not have access to the data that they were summarizing when writing their open-ended responses as to what they learned about the causal system described in the stimuli, one can understand the granularity of these open-ended responses as indicative of how different causal features of the data were encoded in participants' memories, and then reported to reflect what they had *learned* (versus what they would share with a colleague). Thus, the results of this experiment suggest that participants were more likely to encode more compressed causal summaries of data in their memory when doing so would not result in the loss of valuable information. This provides more direct evidence for our claim that our model explains not just the granularity of people's linguistic representation of causal structure in the context of interpersonal communication, but also their mental representation of the same structure. That said, Experiment 5 still involved some element of communication insofar as participants were providing responses for experimenters. Among other things, Experiment 6 takes a step farther away from representations in the service of interpersonal communication, demonstrating effects of information loss on compression in other forms of representation.

Experiment 6

Experiments 1-3 found that participants favored more compressed representations when the amount of information lost in compression was zero. This suggests that, all else being equal, participants assign some cost to storing a less compressed representation. Experiments 4-5 found that participants favored more compressed representations when the value of information lost in compression was zero. This suggests that the more valuable the information lost, the more willing participants should be to incur the representational costs of storing a less compressed

representation. Schematically, this relationship between the cost of compression, on the one hand, and the value of lost information, on the other, can be written as follows:

Propensity to Form Compressed Representation ∝ Cost of Non-Compression – VOLI

It follows from these observations that it should be possible to manipulate the propensity to form compressed representations not only by manipulating VOLI (as we did in Experiments 4-5), but also by manipulating cost. In Experiment 6, we do precisely this, and find that participants are more inclined to store a compressed representation of a data set when: i) compression is less costly, and ii) the amount (and therefore, the value) of the information lost in compression is lower. In contrast with Experiments 1-4 (and potentially Experiment 5), our key dependent variable does *not* involve communication on a participant's part. Rather, it directly measures the level of compression at which an agent chooses to store information about a data set.

Participants

Participants were 202 adults recruited via Prolific. An additional 29 participants were excluded for failing comprehension checks. The sample of participants was 50% female and 50% male, with an age range of 18-76 and a mean age of 36.

Materials and Procedures

Participants were told that they would be playing a role-playing game (RPG) in which their character began the game with 1000 health points and 1000 gold coins, and where the goal of the game was to end with as many health points and gold coins as possible. They were also told that over the course of the game, their character might sustain damage, reducing their total health points. To regain health points, they would be able to use gold coins to purchase remedies. Each remedy has a numerical strength, from 0-100, where the remedy's strength denotes the number of health points that can be regained by taking the remedy. To illustrate, if a participant's character

has 800 health points after sustaining damage, and takes a remedy with a strength of 50, then the character will have 850 health points after taking the remedy.

Participants were shown a table with a list of remedies that they would potentially be able to buy, along with the strength of each remedy. The contents of each table depended on whether the participant had been randomly assigned to a "low loss" or "high loss" condition, and are shown in Table 8. The "low loss" condition is so-named because if this table were to be compressed into one that only compared the strength of mushrooms and flowers, less information about the differential strengths of remedies would be lost than if the same compression were performed in the "high loss" condition.

Table 8

Data shown to Participants in Experiment 6.

Low Loss		High Loss	
Remedy	**Strength**	**Remedy**	**Strength**
Green Mushroom	82	Green Mushroom	99
Yellow Mushroom	81	Yellow Mushroom	81
Purple Mushroom	80	Purple Mushroom	63
Yellow Flower	19	Yellow Flower	35
Red Flower	17	Red Flower	1

Participants were told that, later in the game, they would not have access to the information in the table they were shown, even though that information would be relevant. However, they would be able to use gold coins to purchase one of two note cards summarizing the data in the table. They would have access to the information on their purchased note card throughout the course of the game. Those two note cards, along with their costs, were as follows:

More Compressed: "Mushrooms are stronger than flowers." Cost: 100 Gold Coins.

Less Compressed: "Green mushrooms are strongest, followed by yellow mushrooms, then purple mushrooms, then yellow flowers, then red flowers." Cost: [100/120] Gold Coins.

The cost of the less compressed note card was manipulated between participants, and set to either 100 or 120 gold coins, to reflect an additional cost of a less-compressed data representation of either 0 or 20. We predicted that, due to the increased loss of valuable information in the high loss category, participants would choose the less compressed causal representation more often in the high loss condition than in the low loss condition. We predicted further that this difference would be most pronounced when there is a positive cost to storing a less-compressed data representation; when storing less-compressed data comes with an explicit cost, agents will be more likely to incur that cost when doing so enables them to store a greater amount of valuable information.

In the remainder of the experiment, we told participants that their character had incurred damage, and then asked them to state how many gold coins they were willing to pay for a yellow mushroom to counteract that damage. They had access to the information on their purchased note card when making this decision. They were then asked to rate how helpful they found their note card, and then asked if they could recall the strength of each remedy from the table they were shown. These subsequent dependent variables were collected for exploratory purposes, but were not part of our core analysis, which concerns participants' choice of note card.

Results

Figure 8 shows the proportion of participants selecting the less-compressed causal representation as a function of both the additional cost of the less compressed representation and the amount of valuable information lost in compression. As is clear from the figure, when there is no additional cost to storing a less compressed causal representation, a large majority of participants elect to do so. However, once there is a substantive cost to storing a less compressed causal representation of data, we see significant variation as to which participants are willing to

select a less-compressed representation, with participants much more likely to do so when the

amount of valuable information lost in compression is low.

Figure 8

Proportion of participants in Experiment 6 choosing a less compressed data representation, by level of information loss and the additional cost of a less compressed representation.
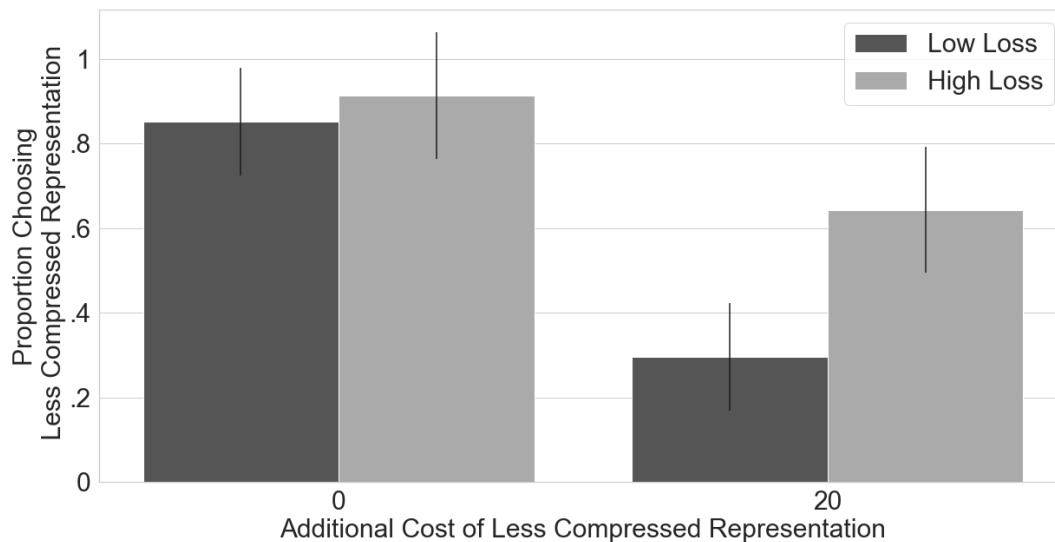
Note. Bars show 95% confidence intervals for proportions.

We performed a binary logistic regression for a dependent variable representing whether

or not a participant chose the less compressed data representation, regressing that variable against:

i) the additional cost of the less compressed representation (Cost), and ii) a binary variable

representing whether a participant was assigned to the low and high loss condition (Loss). The low

loss condition was coded as -1 and the high loss condition was coded as 1. We found that both

variables were significant predictors of whether or not a participant chose the less compressed

representation (Cost: $\beta = -.116$, $p < .001$; Loss: $\beta = .602$, $p = .001$); participants with an additional

cost of storing a less compressed representation were, all else being equal, less likely to do so,

while participants in the high loss condition were, all else being equal, more likely to store a less

compressed representation of the data, in keeping with our predictions. (In the Supplemental

Materials, we report the results of another version of this experiment that replicates the effect of manipulating the cost of a less-compressed causal representation on participants' propensity to store such a representation, but does not find a significant effect of the overall value of the information lost in compression on participants' choice as to the level of compression at which they represent data.)

Discussion

The results of this experiment provide strong evidence that both the cost of storing a less compressed representation and the amount of valuable information lost in compression affect people's propensity to store more or less compressed representations of their environment. Unlike Experiments 1-4 (and perhaps Experiment 5), the results of Experiment 6 cannot be explained in terms of participants adhering to communicative norms, as their selection of a note card summarizing the data for use later in the game was not an act of interpersonal communication, but was instead a choice about how to store information at a particular level of compression. One can view the choice of note card as a choice on the participant's part about how to represent the information in the data that they were shown within their extended mind (Clark & Chalmers, 1998) for subsequent use in navigating the environment of the role-playing game or as a "pragmatic action" that the agent performs to change their own computational state (Kirsh & Maglio, 1994). Thus, the results of Experiment 6 speak in favor of our argument that agents choose causal representations by weighing a preference for compression against a desire to retain decision-theoretically relevant information.

The results of Experiment 6 also speak to a potential alternative explanation of the findings from Experiments 4-5. The alternative explanation is that our findings do not reflect our posited trade-off between compression and the value of information, but instead reflect the effects of

adding a new feature (e.g., whether a medicine should be stocked) to a classification task. Specifically, Waldmann and Hagmayer (2006) show that participants presented with identical data about the causal capacities of exemplars nonetheless make different attributions of causal capacity depending on the categories they bring to the task and use to classify the exemplars. If the mechanisms identified by Waldmann and Hagmayer are applied to Experiment 4, for example, then the four ingredients in the pharmaceutical example could be described as possessing the features shown in Table 9, where our manipulation of decision-theoretic threshold (50% vs. 80%) plays the role of an additional feature (stocked in the pharmacy or not) that shapes the categories participants use to classify exemplars. The two types of Reptol would then be grouped together when and because they both share the feature of being stocked at the pharmacy, rather than because of any effect of the change in value of the information lost in compression brought about by manipulating the decision-theoretic threshold. However, a similar analysis *cannot* be used to explain the results of Experiment 6, since the cost of a less-compressed representation is not a feature of any of the remedies that participants compress over (e.g., changing the cost of the more detailed note card from 100 gold coins to 120 gold coins does not change any features of a yellow mushroom).[23]

Table 9:

Putative features of the four causes in the pharmaceutical vignette of Experiment 4.

Ingredient	p(e|c) (i.e., Effectiveness)	Stocked at Pharmacy?
Type-1 Reptol	.85/.98	Yes
Type-2 Reptol	.7	Yes/No
Type-1 Psylo	.01	No
Type-2 Psylo	.01	No

---

[23] It is also worth noting that while the mechanism identified in Waldmann and Hagmayer (2006) could explain why participants were influenced by the threshold manipulation in Experiments 4-5, it is not clear how it could explain the interaction with information loss that was also observed in Experiment 4.

General Discussion

As established in the introduction, representing a causal system involves a trade-off between informativeness and compression. How do agents manage this trade-off? In this paper we have put forward a theoretical framework, using the formal apparatus of Bayesian networks and information theory, that quantifies how much information is lost in the move from a less compressed causal representation of an agent's environment to a more compressed representation of the same environment. We propose that agents trade off an all-things-considered preference for compression against a desire to avoid losing information due to compression. In so doing, we are also able to offer a unified account of the proportionality and stability of causal claims.

Experiments 1 and 2 support this part of our proposal. They show that when no information is lost in compression, people prefer causal claims embedded in more compressed models to those embedded in less compressed ones. When information loss due to compression is modest, we do not see a significant difference in peoples' evaluations of claims embedded in more and less compressed models. By contrast, when information loss due to compression is considerable, we see a strong preference for causal claims that are embedded in less compressed, more detailed causal models. Whether compression was achieved by coarsening a causal variable (proportionality) or removing a background condition (stability) did not make a significant difference with respect to participants' evaluations of causal claims, suggesting that our theoretical framework provides a unifying account of these important dimensions along which causal claims can be compared. Experiment 3 further corroborated this part of our proposal, while also showing that our account is able to explain results that are less well-explained in terms of causal contrast.

We further elaborated our framework by considering how decision-theoretic factors may influence agents' preferences over causal claims. Specifically, we introduced a framework for

quantifying the value of the information lost in the move from a less compressed causal representation to a more compressed representation of the same environment, for an agent in a particular decision context. This allowed us to state precisely whether the information lost in a given compression is or is not valuable to a particular agent, in keeping with the broader theory that an agent's construal of the causal structure of their environment is fundamentally informed by the prudential values of that agent. We tested this aspect of our framework in Experiments 4-5, which found that when participants evaluate causal claims, they engage in a trade-off between a preference for compression and a preference for *valuable* information, where the value of information is determined by the decision-theoretic context in which a particular data-generating process is presented to participants. Importantly, this preference for valuable information is continuous; in Experiment 4, we found that the greater the amount of valuable information lost in a compression, the more negatively participants evaluate a compressed causal claim. In keeping with our hypothesis, discussed in the introduction, that human agents are capable of selecting from among different possible causal models of their environment to fit a specific context, the results of Experiment 4 suggest that this selection process is strongly influenced by the decision-theoretic structure of a given context in which a mental causal model is deployed.

Finally, we noted that, while we took our dependent variables in Experiments 1-4 to measure the extent to which participants were representing their environment at a given level of compression, they may be alternatively interpreted as reflecting solely communicative norms. To bolster the case for a representational interpretation of our findings, Experiment 5 involved open-ended reports of what participants had *learned*, and we additionally conducted Experiment 6, which measured the effect of both the cost of a less compressed representation and the value of the information lost in compression on a variable that is much more plausibly interpreted as

measuring participants' representational preferences than their communicative ones. This experiment found a significant effect of both the cost of a less compressed representation and the value of the information lost on participants' propensity to represent data at a certain level of compression.

The overall takeaway from our six experiments is as follows. When agents represent the causal structure of their environment, they have an extremely wide latitude with respect to how compressed that representation should be. Ultimately, the level of compression that an agent chooses for such a representation is determined by a trade-off wherein agents seek to minimize the loss of valuable information while maximizing compression. We note that in Experiment 5, we do not see a significant effect of the *amount* of information lost in compression on participants' preferences as to the level of compression at which they represent data. Thus, based on the data collected here, we take it to be something of an open question whether, when a particular decision context is specified, minimizing the loss of information involves a continuous preference for less information loss when the information loss is valuable (as observed in Experiment 4), or a categorical preference for representations in which the value of lost information is zero and against representations in which the value of lost information is positive (as observed in Experiment 5). We speculate that information loss is likely to play *some* role even when the value of information is zero, given both the results of Experiments 4 and the fact that agents may have uncertainty regarding their own prudential values, or anticipate that their values might change. What *is* clear is that the tradeoff between informativeness and compression holds independently of whether compression is achieved by coarsening a causal variable or eliding a background condition. This suggests a common framework for understanding these two salient varieties of compression.

Finally, it is worth noting that while our discussion here pertains to explicitly *causal* representations of data-generating processes, it may be extended further to include non-causal predictive models that might be used in agential representation of the environment and cognition. As long as these models use random variables and involve predicting a particular outcome of interest, much of the formal apparatus and experimental paradigms developed here will remain applicable. Thus, we hope that our results here have provided a general framework for thinking through the relationship between information loss, compression, and prudential values as the chief drivers of the processes whereby humans and other agents construct formal representations of their environment, for use in both navigating and intervening upon the world in which they live.

Relationship to Existing Model Selection Techniques

In statistics, the tradeoff between model simplicity and informativeness is most commonly understood through the lens of various "information criteria" used for model selection (Akaike, 1998; Schwarz, 1978). Indeed, there are similarities between our discussion of causal model selection and the much larger statistical literature on model selection via information criteria, in which the fundamental tradeoff is between selecting a model that captures all the data and a model that contains a small number of parameters. Nevertheless, there are also important differences. First, when we coarse-grain a causal variable in a model and leave the rest of the model unchanged, we do not necessarily change the number of statistical parameters in the model (e.g., the probability distribution over a continuously-valued variable may have the same number of parameters as a distribution over a discretization of that variable), though such coarsening may change the parameters that maximize the likelihood of the data. Nevertheless, in our framework a model with a more coarse-grained variable is considered to be more compressed, all else being equal, than the same model with a more fine-grained version of the same variable. In addition, we trade off

compression or simplicity against the amount of information shared between causes and their effects, rather than the likelihood of a model given some set of observed data. Our measure is meant to be applied in cases where more and less compressed models of some system are all well-supported by data, and yet there is nevertheless a trade-off between the simplicity of a model and the informativeness of cause-effect relationships within that model. That said, there are likely conditions under which our framework and other statistical techniques for model selection will make similar recommendations. We leave it to future work to examine these conditions in depth.

Limits of the Current Study

While we take our experimental results to confirm the theoretical framework presented, we acknowledge that they have some important limitations. As noted at the outset, we focus on qualitative predictions derived from principled measures of compression and information loss, but we do not claim that our specific measures correspond to the algorithms by which humans compute these quantities. This leaves it to future work to develop and test quantitative process models. Additionally, our experiments were conducted solely on U.S.-based participants recruited through on-line platforms, and so we are limited in the extent to which we can generalize to all agents faced with the task of creating causal models of the world. In particular, it is a largely open question whether individual scientists or groups of scientists would instantiate the same trade-offs observed here. In addition, our results are constrained to relatively simple causal scenarios that can be presented and understood in a matter of minutes, such that our results are of less significance in understanding deliberative, detailed causal understanding. This is a notable limitation if – as seems likely – variable choice interacts with information search in a dynamic process of inquiry, whereby current representations guide interventions, which in turn generate the data that revise representations. Investigating this interactive process is an important direction for future research.

Although this issue is addressed to a large degree by Experiments 5-6, one could still worry about the extent to which our predictions differ from what one might predict simply by assuming that listeners take speakers to follow Grice's *maxim of quantity*, which states that speakers should utter all and only the information relevant in a given context (Grice, 1975). In particular, if we assume that our manipulation of value-of-information thresholds affects what speakers take to be relevant, then speakers should utter less detailed claims (i.e., more coarse-grained claim) when the details they omit are irrelevant, and more detailed claims (i.e., more fine-grained claims) otherwise. This is consistent with what we observe in Experiment 4. However, the findings from Experiments 1, 2, and 4 highlight important ways in which our framework goes beyond – though is not inconsistent with – this basic Gricean idea. First, all three of these experiments demonstrated a graded effect of information loss on the granularity of causal claims, with more fine-grained representations becoming increasingly frequent as information loss increases. One sees a similar effect of the value of information in Experiment 4: restricting to just those cases where the value of information is positive, one sees a preference for more detailed, fine-grained causal claims as the amount of valuable lost information increases. Thus, information loss and the value of information offer a continuous quantification of the less precise and typically binary term "relevance," and our findings lend it support. Finally, Experiment 6 finds predicted effects on a form of representation (a personal memory aid) that does not have an interpersonal, communicative function. This is consistent with our focus on effects of compression and the loss of valuable information on causal representation quite generally, rather than interpersonal linguistic communication in particular.

Directions for Future Work

The current results suggest several intriguing avenues for future work that would generalize our findings to other areas of psychology. One such avenue emerges when we consider that the human capacity for compressed representation of causal structure begins very early in life. Future work in developmental psychology could show whether the process of selecting compressed representations of formal structure is fundamentally goal-oriented in very young humans. If this is the case, it would lend further support to the theory that our early-emerging and core commitments regarding the causal structure of data-generating processes are shaped by our pragmatic goals as agents.

Another intriguing line for future research concerns the psychology of social categorization. When we group people into categories such as race or gender, we situate them within a socially-constructed causal nexus based on group stereotypes. Our choices of social categories used to classify people are extremely ethically fraught, such that understanding why we group people in the particular ways that we do is a central goal of social psychology. If we are correct in thinking that the classification schema used in causal reasoning generally are downstream of our prudential goals as agents, then work in social psychology might establish that the same holds for social classification. This could yield a new analysis of the ethics of stereotyping, according to which the moral valence of a particular classification schema for individuals is tied to the moral valence of the prudential goals that led to that schema (in fact, see Kinney & Lombrozo, 2024, for recent evidence to this effect).

On the formal and mathematical side, as acknowledged above, our measure of information loss is an application of rate distortion theories developed in other areas of cognitive science. In rate distortion theories, one often finds a setting of parameters such that agents' preferences over

more and less distorted or compressed information channels follow a "rate distortion curve" showing the optimal level of distortion (Zaslavsky et al., 2018). A potentially fruitful formal project would involve spelling out, in full formal detail, how our measure of information loss can be re-stated as a rate distortion curve, with the decision-theoretic value of information being used to set key parameters that determine the shape of the curve. Such a formal study would amount to a significant unification of the literature in rate distortion theory and the literature on causal variable choice.

Finally, we note that the vignettes used in all our experiments only include cases where the causes in question generate their effects; we do not ask participants about cases where the causal relationship of interest is one where the occurrence of some cause *prevents* the occurrence of some effect. We believe that such cases can be accommodated by our formal model: if the values of an effect variable E each represent the non-occurrence of some effect, rather than the occurrence thereof, then one could calculate information loss between such an effect variable and different causal variables without any change in our mathematical approach. However, it is possible that participants would respond differently to versions of the experiments presented in the current paper where the causal relationships described are preventative rather than generative. Running these kinds of experiments would be an intriguing avenue for future work, especially since any observed differences between the predictions of our model and the results of such experiments would require at least some revision of the current theory.

These potential directions for future research, alongside our analysis of the results of the current studies, speak to the fruitfulness of our theoretical approach that combines information theory and decision theory to offer a unified analysis of causal structure selection in cognition. We believe that future applications of this approach may lead to a more comprehensive understanding

of the relationship between goal-setting, planning, and representation of the environment in humans, non-human animals, and artificially intelligent agents.

Conclusion

We have proposed a theoretical framework for measuring the amount of information lost in the move from a less-compressed to a more-compressed causal model of an environment. This framework allows us to give a unified account of the proportionality and the stability of causal claims. This framework can additionally quantify and incorporate the decision-theoretic *value* of the information that is lost in compression. Over the course of six experiments, we demonstrated the empirical adequacy of this framework in predicting people's evaluation and generation of causal claims. This suggests that, as per our hypothesis, human representations of the causal structure of the environment do trade-off valuable information against compression in a context-dependent way.

References

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected papers of hirotugu akaike* (pp. 199–213). Springer.

Allan, L. G. (1980). A note on measurement of contingency between two binary variables in judgment tasks. *Bulletin of the Psychonomic Society*, *15*(3), 147–149. https://doi.org/10.3758/BF03334492

Aronowitz, S., & Lombrozo, T. (2020). Experiential Explanation. *Topics in Cognitive Science*, *12*(4), 1321–1336. https://doi.org/10.1111/tops.12445

Ay, N., & Polani, D. (2008). Information flows in causal networks. *Advances in Complex Systems*, *11*(01), 17–41.

Bechlivanidis, C., Lagnado, D. A., Zemla, J. C., & Sloman, S. (2017). Concreteness and

abstraction in everyday explanation. *Psychonomic Bulletin & Review*, *24*(5), 1451–1464.

https://doi.org/10.3758/s13423-017-1299-3

Beckers, S., & Halpern, J. Y. (2019). Abstracting causal models. *Proceedings of the Aaai

Conference on Artificial Intelligence*, *33*(01), 2678–2685.

Blackwell, D. (1953). Equivalent comparisons of experiments. *The Annals of Mathematical

Statistics*, 265–272.

Blanchard, T., Vasilyeva, N., & Lombrozo, T. (2018). Stability, breadth and

guidance. *Philosophical Studies*, *175*, 2263-2283.

Bourrat, P. (2021a). Heritability, causal influence and locality. *Synthese*, *198*(7), 6689–6715.

Bourrat, P. (2021b). Measuring Causal Invariance Formally. *Entropy*, *23*(6), 690.

Bramley, N. R., Dayan, P., Griffiths, T. L., & Lagnado, D. A. (2017). Formalizing Neurath's

ship: Approximate algorithms for online causal learning. *Psychological Review*, *124*(3),

301.

Bramley, N. R., Lagnado, D. A., & Speekenbrink, M. (2015). Conservative forgetful scholars:

How people learn causal structure through sequences of interventions. *Journal of

Experimental Psychology: Learning, Memory, and Cognition*, *41*(3), 708.

Brodu, N. (2011). Reconstruction of epsilon-machines in predictive frameworks and decisional

states. *Advances in Complex Systems*, *14*(05), 761–794.

Buchsbaum, D., Griffiths, T. L., Plunkett, D., Gopnik, A., & Baldwin, D. (2015). Inferring action

structure and causal relationships in continuous sequences of human action. *Cognitive

Psychology*, *76*, 30–77.

Chater, N., & Vitányi, P. (2003). Simplicity: A unifying principle in cognitive science? *Trends in Cognitive Sciences*, *7*(1), 19–22.

Cheng, P. W. (1997). From covariation to causation: A causal power theory. *Psychological Review*, *104*(2), 367.

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, *58*(1), 7–19.

DiMarco, M. (2021). Wishful Intelligibility, Black Boxes, and Epidemiological Explanation. *Philosophy of Science*, *88*(5), 824–834.

Fauconnier, G., & Turner, M. (2008). *The way we think: Conceptual blending and the mind's hidden complexities*. Basic books.

Franklin-Hall, L. R. (2016). High-level explanation and the interventionist's 'variables problem.' *The British Journal for the Philosophy of Science*, *67*(2), 553–577.

Gebharter, A., & Eronen, M. I. (2021). Quantifying proportionality and the limits of higher-level causation and explanation. *The British Journal for the Philosophy of Science*, 714818. https://doi.org/10.1086/714818

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2021). A counterfactual simulation model of causal judgments for physical events. *Psychological Review*, *128*(5), 936.

Good, I. J. (1960). The paradox of confirmation. *The British Journal for the Philosophy of Science*, *11*(42), 145–149.

Gopnik, A., & Sobel, D. M. (2000). Detecting blickets: How young children use information about novel causal powers in categorization and induction. *Child Development*, *71*(5), 1205–1222.

Gopnik, A., & Tenenbaum, J. B. (2007). Bayesian networks, Bayesian learning and cognitive

    development. In *Developmental science* (Vol. 10, Issue 3, pp. 281–287). Citeseer.

Grice, H. P. (1975). Logic and conversation. In *Speech acts* (pp. 41–58). Brill.

Griffiths, P. E., Pocheville, A., Calcott, B., Stotz, K., Kim, H., & Knight, R. (2015). Measuring

    causal specificity. *Philosophy of Science*, *82*(4), 529–555.

Griffiths, T. L., Kemp, C., & Tenenbaum, J. B. (2008). *Bayesian models of cognition*.

Harbecke, J. (2021). Causal Proportionality as an Ontic and Epistemic Concept. *Erkenntnis*, 1–

    23.

Ho, M. K., Abel, D., Correa, C. G., Littman, M. L., Cohen, J. D., & Griffiths, T. L. (2022).

    People construct simplified mental representations to plan. *Nature*, *606*(7912), Article

    7912. https://doi.org/10.1038/s41586-022-04743-9

Hoel, E. P. (2017). When the map is better than the territory. *Entropy*, *19*(5), 188.

Hoffmann-Kolss, V. (2014). Interventionism and higher-level causation. *International Studies in

    the Philosophy of Science*, *28*(1), 49–64.

Icard, T. F., Kominsky, J. F., & Knobe, J. (2017). Normality and actual causal strength.

    Cognition, *161*, 80–93.

Keil, F. C. (2006). Explanation and understanding. *Annu. Rev. Psychol.*, *57*, 227–254.

Kinney, D. (2019). On the explanatory depth and pragmatic value of coarse-grained,

    probabilistic, causal explanations. *Philosophy of Science*, *86*(1), 145–167.

Kinney, D., & Watson, D. (2020). Causal feature learning for utility-maximizing agents.

    International Conference on Probabilistic Graphical Models, 257–268.

Kinney, David & Lombrozo, Tania (2024). Tell Me Your (Cognitive) Budget and I'll Tell You

    What You Value. Cognition.

Kirfel, L., Icard, T., & Gerstenberg, T. (2021). Inference from explanation. *Journal of Experimental Psychology: General*.

Kirsh, D., & Maglio, P. (1994). On distinguishing epistemic from pragmatic action. *Cognitive Science*, *18*(4), 513–549.

Korb, K. B., Nyberg, E., & Hope, L. R. (2011). A new causal power theory. In *Causality in the Sciences* (pp. 628–652). Oxford University Press.

Lien, Y., & Cheng, P. W. (2000). Distinguishing genuine from spurious causes: A coherence hypothesis. *Cognitive Psychology*, *40*(2), 87–137.

List, C., & Menzies, P. (2009). Nonreductive physicalism and the limits of the exclusion principle. *The Journal of Philosophy*, *106*(9), 475–502.

Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive Psychology*, *55*(3), 232–257.

Lombrozo, T. (2010). Causal–explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology*, *61*(4), 303–332.

Marzen, S. E., & DeDeo, S. (2017). The evolution of lossy compression. *Journal of The Royal Society Interface*, *14*(130), 20170166.

Morris, A., Phillips, J. S., Icard, T., Knobe, J., Gerstenberg, T., & Cushman, F. (2018). *Causal judgments approximate the effectiveness of future interventions*.

Murphy, G. (2004). *The big book of concepts*. MIT press.

O'Neill, K., Henne, P., Bello, P., Pearson, J., & De Brigard, F. (2021). *Degrading causation*.

O'Neill, K., Henne, P., Bello, P., Pearson, J., & De Brigard, F. (2022). Confidence and gradation in causal judgment. *Cognition*, *223*, 105036.

Pacer, M., & Lombrozo, T. (2017). Ockham's razor cuts to the root: Simplicity in causal

explanation. *Journal of Experimental Psychology: General*, *146*(12), 1761.

Pacer, M., Williams, J., Chen, X., Lombrozo, T., & Griffiths, T. (2013). Evaluating

computational models of explanation using human judgments. *arXiv preprint

arXiv:1309.6855*.

Papineau, D. (2022). The statistical nature of causation. *The Monist*, *105*(2), 247–275.

Pearl, J. (1994). A probabilistic calculus of actions. In *Uncertainty Proceedings 1994* (pp. 454–

462). Elsevier.

Pearl, J. (2000). *Causality: Models, Reasoning and Inference*. Cambridge University Press.

Pocheville, A., Griffiths, P. E., & Stotz, K. (2017). Comparing causes—An information-theoretic

approach to specificity, proportionality and stability. *Proceedings of the 15th Congress of

Logic, Methodology and Philosophy of Science*, 93–102.

Quillien, T. (2020). When do we think that X caused Y? *Cognition*, *205*, 104410.

Rosch, E. (1978). *Principles of categorization*.

Ross, L. N. (2015). *Causal control: A rationale for causal selection*.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 461–464.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical

Journal*, *27*(3), 379–423.

Sims, C. R. (2016). Rate–distortion theory and human perception. *Cognition*, *152*, 181–198.

Spellman, B. A. (1997). Crediting causality. *Journal of Experimental Psychology: General*,

126(4), 323.

Spirtes, P., Glymour, C., & Scheines, R. (2000). Causation, Prediction, and Search. *MIT Press

Books*, *1*.

Tishby, N., Pereira, F. C., & Bialek, W. (2000). The information bottleneck method. *arXiv Preprint Physics/0004057.*

Vasilyeva, N., Blanchard, T., & Lombrozo, T. (2018). Stable causal relationships are better causal relationships. *Cognitive Science*, *42*(4), 1265–1296.

Waldmann, M. R., & Hagmayer, Y. (2006). Categories and causality: The neglected direction. *Cognitive Psychology*, *53*(1), 27–58.

Watson, D., & Silva, R. (2022). *Causal discovery under a confounder blanket* (arXiv:2205.05715). arXiv. http://arxiv.org/abs/2205.05715

Weslake, B. (2013). Proportionality, contrast and explanation. *Australasian Journal of Philosophy*, *91*(4), 785–797.

Wilkenfeld, D. A. (2019). Understanding as compression. *Philosophical Studies*, *176*(10), 2807–2831.

Wojtowicz, Z., Chater, N., & Loewenstein, G. (2021). *The motivational processes of sense-making*.

Woodward, J. (2008). Mental causation and neural mechanisms. *Being Reduced: New Essays on Reduction, Explanation, and Causation*, 218–262.

Woodward, J. (2010). Causation in biology: Stability, specificity, and the choice of levels of explanation. *Biology & Philosophy*, *25*(3), 287–318.

Woodward, J. (2016a). Causal cognition: Physical connections, proportionality, and the role of normative theory. *Of Psychology: Causality and Psychological Subject*, 105.

Woodward, J. (2016b). The problem of variable choice. *Synthese*, *193*(4), 1047–1072.

Woodward, J. (2019). On Wolfgang Spohn's Laws of Belief. *Philosophy of Science*, *86*(4), 759–772.

Woodward, J. (2021a). *Causation with a human face: Normative theory and descriptive psychology*. Oxford University Press.

Woodward, J. (2021b). Explanatory autonomy: The role of proportionality, stability, and conditional irrelevance. *Synthese*, *198*(1), 237–265.

Yablo, S. (1992). Mental causation. *The Philosophical Review*, *101*(2), 245–280.

Zaslavsky, N., Kemp, C., Regier, T., & Tishby, N. (2018). Efficient compression in color naming and its evolution. *Proceedings of the National Academy of Sciences*, *115*(31), 7937–7942.

Appendix A

In this appendix, we provide our full formal framework for measuring the amount of information that is lost in the move from one Bayesian network representing the causal structure of some system to a second Bayesian network, where this second Bayesian network amounts to a more compressed representation of a given target system than the first. This allows us to state precisely our claim that both the proportionality and stability of a causal claim can be defined in terms of information loss between causal models. It also provides the theoretical framework within which our first three experimental paradigms are situated.

Variables, Coarsenings, and Causal Bayesian Networks

We begin with a **probability space**, which is a triple (Ω, Σ, p) in which Ω is a **sample space** of primitive possibilities (i.e., a set of possible worlds), Σ is an **algebra** on Ω (i.e., a set of subsets of Ω that is closed under union, complement, and intersection), and p is a **probability distribution** on Σ that satisfies the Kolmogorov axioms. A **random variable** $X: \Omega \rightarrow R_X$ is function from the sample space into some set R_X (i.e., the **range** of the random variable). As stated in the introduction, in this paper we assume that random variables are surjective but not injective

functions on the sample space, meaning that multiple possible worlds are often mapped to the same value of a random variable. This clarifies one sense in which representations that use random variables are compressions of their targets; they clump together many possible observations under a single label. A random variable is said to be **measurable** with respect to a probability space (Ω, Σ, p) if and only if for any $x \in R_X$, $X^{-1}(x) \in \Sigma$. This allows us to assign a probability to the event that the variable X takes the value x, using the equation $p(X = x) = p(X^{-1}(x))$.

For any random variable X that is measurable with respect to a probability space (Ω, Σ, p) let $\sim_X$ be an equivalence relation defined on Ω such that $\omega \sim_X \omega'$ if and only if $X(\omega) = X(\omega')$. A random variable $\hat{X}$ is a **coarsening** of X if and only if, for any $\omega, \omega' \in \Omega$: i) if $\omega \sim_X \omega'$ then $\omega \sim_{\hat{X}} \omega'$, and ii) there exists an $\omega, \omega' \in \Omega$ such that $\omega \sim_{\hat{X}} \omega'$ and $\omega \not\sim_X \omega'$. If $\hat{X}$ is a coarsening of X, then X is a **refinement** of $\hat{X}$. The definition of coarsening captures the intuitive idea that coarser-grained random variables define a more general compression of the possibility space on which they are defined than their more fine-grained counterparts. That is, all possibilities treated as equivalent by a random variable X are also treated as equivalent by its coarsening $\hat{X}$, but some possibilities treated as equivalent by $\hat{X}$ are not treated as equivalent by X.

Moving to the use of Bayesian networks to represent causal structure, let $\mathscr{V}_{\mathscr{P}}$ be a set of random variables that are each measurable with respect to a probability space $\mathscr{P} = (\Omega, \Sigma, p)$. Let $\mathscr{E}_{\mathscr{V}_{\mathscr{P}}}$ be an acyclic set of ordered pairs, or edges, relating the variables in $\mathscr{V}_{\mathscr{P}}$. These are represented pictorially as arrows in the causal graph depicting relations of direct causation from one variable to another. A **causal Bayes net** $\mathscr{G}_{\mathscr{P}}$ is a pair $(\mathscr{V}_{\mathscr{P}}, \mathscr{E}_{\mathscr{V}_{\mathscr{P}}})$ that satisfies the following conditions:

1. According to the probability distribution p in the probability space $\mathscr{P}$ with respect to which the Bayes net is defined, all variables are independent of their non-descendants, conditional on their parents (**Markov condition**).

2. There is no set of edges $\mathcal{E}^{*}_{\mathscr{V}_{\mathscr{P}}} \subset \mathcal{E}_{\mathscr{V}_{\mathscr{P}}}$ such that $(\mathscr{V}_{\mathscr{P}}, \mathcal{E}^{*}_{\mathscr{V}_{\mathscr{P}}})$ satisfies the Markov condition

   (**Minimality condition**).

3. No element of $\mathscr{V}_{\mathscr{P}}$ is a coarsening of or identical to any other element of $\mathscr{V}_{\mathscr{P}}$ (**Co-possibility condition**).

The Markov and Minimality conditions formalize the idea that the value of each variable in a causal Bayes net is determined by all and only its parents (i.e., its direct causes), plus an exogenous source of error not accounted for in the Bayes net and not correlated with the error in any other variables. The Co-possibility condition ensures that all functional relationships between variables in a causal Bayes net are indeed causal, rather than logical, in nature.

An important feature of a causal Bayes net is that it allows us to calculate the probability distribution over the variables in the Bayes net given one or more hypothetical interventions setting the value(s) of variables in the Bayes net, in keeping with the "do-calculus" of (Pearl, 2000). For any given causal Bayes net $\mathscr{G}_{\mathscr{P}}$, we can calculate the probability distribution over any variable V in the set $\mathscr{V}_{\mathscr{P}}$, given an intervention setting some set of variables X to some set of values x, using the following formula:

$$p_{\mathscr{G}_{\mathscr{P}}}(v \mid do(x)) = \begin{cases} p\left(v \mid par_{\mathscr{G}_{\mathscr{P}}}(V)\right) if\ V \notin X \\ 1\, if\, V \in X\ and\ v\ is\ consistent\ with\ x \\ 0\ otherwise \end{cases}$$

where $par_{\mathscr{G}_{\mathscr{P}}}(V)$ denotes the values taken by the parents of V in $\mathscr{G}_{\mathscr{P}}$. The idea here is that when the variable that the distribution is defined over is not intervened upon, the distribution is determined solely by the value taken by the parents of that variable. In practice, these values are not always known, but may be known if the variable(s) intervened upon include parents of the variable over which the distribution is defined. Where parents are not known, they are marginalized over. This allows us to derive the probability distribution that would be defined over any variable in the causal

Bayes net, if any other variable in the same causal Bayes net were set to some value via an exogenous, "surgical" intervention on the data-generating system.

For any given causal Bayes net, we can define an equivalence relation $\sim_{\mathscr{V}_{\mathscr{P}}}$ on the sample space Ω, such that $\omega \sim_{\mathscr{V}_{\mathscr{P}}} \omega'$ if and only if for all $V \in \mathscr{V}_{\mathscr{P}}$, $V(\omega) = V(\omega')$. A causal Bayes net $\mathscr{G}_{\mathscr{P}} = \left(\mathscr{V}_{\mathscr{P}}, \mathscr{G}_{\mathscr{V}_{\mathscr{P}}}\right)$ is a **more compressed** representation of a given target process than an alternative causal Bayes net $\mathscr{G'}_{\mathscr{P}} = \left(\mathscr{V'}_{\mathscr{P}}, \mathscr{G}_{\mathscr{V'}_{\mathscr{P}}}\right)$ if and only if: i) for any $\omega, \omega' \in \Omega$, if $\omega \sim_{\mathscr{V'}_{\mathscr{P}}} \omega'$, then $\omega \sim_{\mathscr{V}_{\mathscr{P}}} \omega'$ and ii) there exists an $\omega, \omega' \in \Omega$, such that $\omega \sim_{\mathscr{V}_{\mathscr{P}}} \omega'$ but $\omega \not\sim_{\mathscr{V'}_{\mathscr{P}}} \omega'$. Thus, in an analogy to the coarsening-of relation between variables, a more compressed Bayes net defines a more general equivalence relation over the sample space than a less compressed Bayes net defined over the same sample space and representing the same underlying dynamics.

Measuring Information Loss

We are now in a position to introduce a formal measure of the amount of information that is lost about an effect of interest E, with respect to some sets of causal variables of interest **C** and **C'**, when we move from one Bayes net $\mathscr{G}_{\mathscr{P}}$ to a more compressed Bayes net $\mathscr{G}'_{\mathscr{P}}$ representing the same data-generating process. Specifically, we define an information loss function $\mathcal{L}\left(\mathscr{G}_{\mathscr{P}}, \mathscr{G}'_{\mathscr{P}}, \boldsymbol{C}, \boldsymbol{C'}, E, q\right)$ as follows:

$$\mathcal{L}\left(\mathscr{G}_{\mathscr{P}}, \mathscr{G}'_{\mathscr{P}}, \boldsymbol{C}, \boldsymbol{C'}, E, q\right) = \sum_c q(c) \sum_e p_{\mathscr{G}_{\mathscr{P}}}(e \mid do(c)) \log_2 \frac{p_{\mathscr{G}_{\mathscr{P}}}(e \mid do(c))}{p(e)} - \sum_{c'} q(\boldsymbol{c'}) \sum_e p_{\mathscr{G}'_{\mathscr{P}}}(e \mid do(\boldsymbol{c'})) \log_2 \frac{p_{\mathscr{G}'_{\mathscr{P}}}(e \mid do(\boldsymbol{c'}))}{p(e)}$$

The probabilities $q(\boldsymbol{c})$ and $q(\boldsymbol{c'})$ are respectively interpreted as the probability of an intervention setting the variables in **C** to the vector of values $\boldsymbol{c}$ and the probability of setting the variables in **C'** to the vector of values $\boldsymbol{c'}$.

Measuring Proportionality Using Information Loss

Recall that one way of moving from a causal Bayes net $\mathscr{G}_{\mathscr{P}}$ to a more compressed Bayes net $\mathscr{G}'_{\mathscr{P}}$ is by replacing a variable C in the graph $\mathscr{G}_{\mathscr{P}}$ with its coarsening $\widehat{C}$, and leaving all other variables unchanged. By measuring the amount of information that is lost in the move from $\mathscr{G}_{\mathscr{P}}$ to $\mathscr{G}'_{\mathscr{P}}$, we can compare the amount of information that C communicates about some effect variable E to the amount of information that $\widehat{C}$ communicates about the same variable E, thereby comparing the causal claims 'C causes E' and '$\widehat{C}$ causes E' with respect to their proportionality.

More precisely, let $G_{\mathscr{P}} = \left(\mathscr{G}^1_{\mathscr{P}},...,\mathscr{G}^n_{\mathscr{P}}\right)$ be a sequence of causal Bayes nets, such that the only difference between two causal Bayes nets $\mathscr{G}^i_{\mathscr{P}}$ and $\mathscr{G}^{i+1}_{\mathscr{P}}$ is the replacement of a single variable with a coarsening thereof.[24] This yields a sequence of variables $C = \left(C_1,...,C_n\right)$, with each C_i a variable in the causal Bayes net $\mathscr{G}^i_{\mathscr{P}}$ and a coarsening of all variables $C_{j<i}$. We then say that, in the context of a such a sequence, a variable C_i is **proportional** with respect to an effect variable E to the extent that $\mathcal{L}\left(\mathscr{G}^j_{\mathscr{P}},\mathscr{G}^i_{\mathscr{P}},\{C_j\},\{C_i\},E,q\right)$ is relatively small for all $j < i$. That is, proportional choices of causal variables are those that preserve information about the conditions under which an effect variable E will change, as compared to more fine-grained alternatives.

Measuring Stability Using Information Loss

Recall from our earlier discussion that we can measure the stability of a causal relationship $C \rightarrow E$ embedded in a particular causal Bayes net by removing a set of variables $\boldsymbol{B}$ from that Bayes net and assessing how much information is lost in the move from the original Bayes net to the Bayes net that is created by removing the background variables. This claim can now be made

---

[24] Formally, for any $i < n$ there is a bijection $f:\mathscr{V}^i_{\mathscr{P}} \rightarrow \mathscr{V}^{i+1}_{\mathscr{P}}$ such that $f(C_i) = C_{i+1}$ where C_{i+1} is a coarsening of C_i, and for all $V_i \in \mathscr{V}^i_{\mathscr{P}} \setminus \{C_i\}$, $f(V_i) = V_i$. There is also a bijection $g:\mathscr{E}^i \rightarrow \mathscr{E}^{i+1}$ such that, for any $g((W,Z)) = (W_g,Z_g)$: i) if $W = C_i$ then $W_g = C_{i+1}$, ii) if $Z = C_i$ then $Z_g = C_{i+1}$, iii) if $W \neq C_i$ then $W_g = W$ and iv) if $Z \neq C_i$ then $Z_g = Z$.

precise, using our proposed measure of information loss. Let $\mathscr{G}_{\mathscr{P}} = \left(\mathscr{V}_{\mathscr{P}}, \mathscr{E}_{\mathscr{V}_{\mathscr{P}}} \right)$ be a causal Bayes

net containing a cause C, an effect E, and a set of background variables $\boldsymbol{B}$. Let $\mathscr{G}_{\mathscr{P}}^{-\boldsymbol{B}} = \left(\mathscr{V}_{\mathscr{P}}^{-\boldsymbol{B}}, \mathscr{E}_{\mathscr{V}_{\mathscr{P}}^{-\boldsymbol{B}}} \right)$

be a causal Bayes net with the same structure as $\mathscr{G}_{\mathscr{P}}$ but with all variables in $\boldsymbol{B}$ and all edges going

into or out of variables in $\boldsymbol{B}$ removed.[25] The causal relationship between C and E is **stable** with

respect to background condition $\boldsymbol{B}$ to the extent that the value of $\mathcal{L}\left(\mathscr{G}_{\mathscr{P}}, \mathscr{G}_{\mathscr{P}}^{-\boldsymbol{B}}, \{C\} \cup \boldsymbol{B}, \{C\}, E, q \right)$ is

low. That is, the relationship $C \to E$ is stable with respect to $\boldsymbol{B}$ to the extent that the average amount

of information about E that is communicated by interventions on both C and the variables in $\boldsymbol{B}$ is

similar to the average amount of information about E that is communicated solely by interventions

on C.

Simulations

      To demonstrate the relationship between various parameter values and the overall amount

of information lost in coarse-graining, consider a causal variable C with the range of values $\{c_1,$

$c_2, c_3\}$. Suppose that E is a binary effect variable with values $\{e_1, e_2\}$ of which C is a cause. Fixing

$p(e_1|do(c_1))$ at some value x, let $p(e_1|do(c_2))$ and $p(e_1|do(c_3))$ range between zero and one.

Consider the case in which we compress the variable C by mapping c_1 and c_2 to the value $\hat{c}_1$ of a

more compressed variable $\hat{C}$, and mapping c_3 to the value $\hat{c}_2$. The amount of information lost in

this compression, for all possible values of $p(e_1|do(c_2))$ and $p(e_1|do(c_3))$, is shown in Figure 9

for x=.5 and x=1, under the assumption that $q(c_1) = q(c_2) = q(c_3) = 1/3$. As can be seen from

the figure, when x=.5, information loss is largest for extreme values of $p(e_1|do(c_2))$, so that there

is a greater difference in the effect of c_1 and the value c_2 that it is conjoined with in the

compression. By contrast, when x=1, information loss is greatest for low values of $p(e_1|do(c_2))$.

---

[25] That is, $\mathscr{V}_{\mathscr{P}}^{-\boldsymbol{B}} = \mathscr{V}_{\mathscr{P}} \smallsetminus \boldsymbol{B}$ and $\mathscr{E}_{\mathscr{V}_{\mathscr{P}}^{-\boldsymbol{B}}} = \{X, Y : X \in \boldsymbol{B} \vee Y \in \boldsymbol{B}\}$.

Figure 9

Results of simulations showing the amount of causal information lost in compression under various parametric conditions.

(a)   (b)
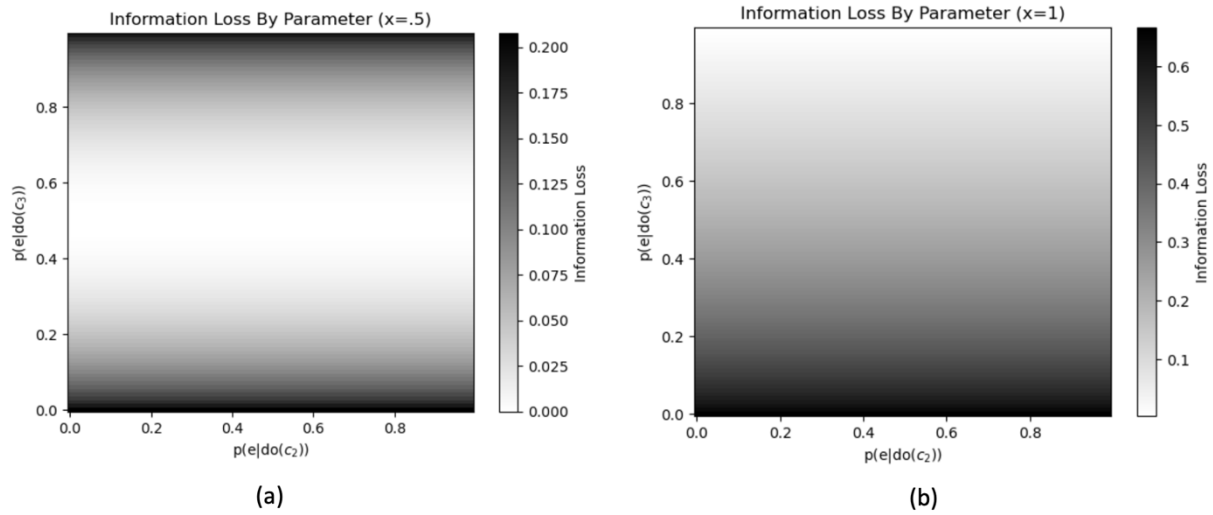
Finally, Figure 10 shows the amount of information that is lost in the same compression when we fix $p(e_1|do(c_1)) = .1$, $p(e_1|do(c_2)) = .5$, and $p(e_1|do(c_3)) = .9$, but vary the distribution q over possible interventions on C. As should be clear from this figure, the more that the probability density of q is concentrated on the causal values c_1 and c_2 that are put together during the process of compression, the greater the overall amount of causal information that is lost in compression.

Figure 10

Results of simulations showing the amount of causal information lost in compression under various

settings of the distribution q.