Evaluations of Causal Claims Reflect a Trade-Off Between Informativeness and Compression

David Kinney (david.kinney@princeton.edu)
Department of Psychology, Princeton University
Princeton, NJ 08540 USA

Tania Lombrozo (lombrozo@princeton.edu)
Department of Psychology, Princeton University
Princeton, NJ 08540 USA

Abstract

The same causal system can be accurately described in many ways. What governs the evaluation of these choices? We propose a novel, formal account of causal evaluation according to which evaluations of causal claims reflect the joint demands of maximal informativeness and maximal compression. Across two experiments, we show that evaluations of more and less compressed causal claims are sensitive to the amount of information lost by choosing the more compressed causal claim over a less compressed one, regardless of whether the compression is realized by coarsening a single variable or by eliding a background condition. This offers a unified account of two dimensions along which causal claims are evaluated (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: causation; compression; proportionality; stability.

Introduction

Not all causal claims are created equal. Consider a fictional mushroom, the Drol, that tends to develop bumpy stems when planted in high-mineral soil. The claim 'planting Drol in high-mineral soil causes them to have bumpy stems' is an example of a type-level causal claim. Yet it is only one of many ways to describe the same causal system. Suppose that high-mineral soil can also be either high or low in sodium, with this distinction making no difference to the likelihood of a Drol developing bumpy stems. Under these conditions, the claim 'planting Drol in high-mineral, high-sodium soil causes them to have bumpy stems,' is still true, but in the terminology of Woodward (2008, 2010, 2018a, 2018b, 2021), the original claim is more appropriate since it expresses a more *proportional* relationship between cause and effect. That is, by describing the cause-effect relationship in a manner that approximates a one-to-one function, it encodes a more informative relationship between cause and effect, which renders it more useful for causal reasoning (Lien & Cheng, 2000).

While 'planting Drol in high-mineral soil causes them to have bumpy stems' is more informative than less proportional alternatives, it may not be maximally informative. In particular, it may omit factors that moderate the causal relationship between the mineral content of soil and a Drol having bumpy stems. Consider a scenario in which Drol that are planted in high-mineral soil *and watered with salt water* are much more likely to develop bumpy stems than Drol planted in high-mineral soil and watered with fresh water. In the terminology of Woodward (2010, 2018b, 2021), under these conditions the claim 'planting Drol in high-mineral soil and watering them with salt water causes them to have bumpy stems' would be more *stable*, or robust with respect to variation in unspecified background conditions, than the alternative that omits any specification of the kind of water used when planting Drol. Data suggest that human causal reasoning is sensitive to stability, with more stable causal claims evaluated more favorably (Vasilyeva, Blanchard, & Lombrozo, 2018).

Why are some causal descriptions judged better than others? And more specifically, why might proportionality and stability guide our evaluations of causal claims? Here we propose a novel, unified account of these phenomena. The core idea is that causal claims can be understood as balancing demands for maximal informativeness, on the one hand, and maximal *compression*, on the other. We formalize this idea with mathematical definitions of proportionality and stability according to which both properties of a causal claim can be measured by assessing the degree to which that claim achieves a loss-minimal compression of a more fine-grained description of the same data. Like other trade-offs in human cognition, such as that between cognitive economy and informativeness in classification (Rosch, 1999), the result is a 'basic level' of causal description - one that most efficiently meets our informational needs in a given situation.

Our approach has two important implications. First, and consistent with prior work on evaluations of causal claims, our formalizations can capture the graded nature of causal judgment (Cheng, 1997; Spellman, 1997; Lombrozo, 2007, 2010; Icard, Kominsky, & Knobe, 2017; Morris et al., 2018; Quillien, 2020; Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2021; O'Neill, Henne, Bello, Pearson, & De Brigard, 2021). Within the interventionist, Bayesian network-based approach to causal inference, explanation, and description made prominent by Pearl (2000) and Spirtes, Glymour and Scheines (2000), we are licensed to draw causal conclusions of the form 'X causes Y,' where X and Y are types of events, just in case the Bayesian network representing the data-generating process is such that there is a directed path from a random variable representing types of events X to a random variable representing types of events Y. We hold that, in addition, evaluations of causal descriptions have a graded structure, and that this graded structure can be captured by our analyses of proportionality and stability.

Second, our approach is consistent with the thesis that pro-

portionality and stability are two instances of the same general property of a causal claim, namely the degree to which the claim minimizes information loss due to compression. In this way our proposal departs from prior treatments of proportionality and stability, and offers a unifying framework that makes additional predictions about causal judgments in everyday cognition and in scientific practice. This allows us to subsume our understanding of evaluations of causal claims under a broader cognitive framework in which compression plays a central role (Keil, 2006; Pacer & Lombrozo, 2017; Wilkenfeld, 2019; Kirfel, Icard, & Gerstenberg, 2021).

In what follows, we offer background on the causal Bayes net formalism and its extension to proportionality and stability, with the introduction of information loss. We also review prior empirical work on proportionality and stability in human causal reasoning. We then present results from two experiments in which participants are asked to evaluate various causal descriptions of the same underlying system, where these causal claims contain different levels of detail, and imply different levels of information loss as a result of compression. In keeping with our hypothesis, we find that participants evaluate less detailed causal claims similarly to more detailed causal claims when replacing the more detailed claim with the less detailed one results in less information loss.

Background

Coarsening and Causal Bayes Nets A **probability space** is a triple (Ω, Σ, p), where Ω is a **sample space**, Σ is an **algebra** on Ω (i.e., a set of subsets of Ω closed under union, intersection, and complement), and $p : \Sigma \to [0,1]$ is a **probability distribution** on Σ. A **random variable** is a function $X : \Omega \to R_X$, where the **range** R_X of X is any set. A random variable is **measurable** with respect to a probability space (Ω, Σ, p) iff, for any $x \in R_X$, $X^{-1}(x) \in \Sigma$.

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 $\omega \sim_X \omega'$ iff $X(\omega) = X(\omega')$. A second random variable $\tilde{X}$ that is measurable with respect to (Ω, Σ, p) is a **coarsening** of X iff: i) for any $\omega, \omega' \in \Omega$, if $\omega \sim_X \omega'$ then $\omega \sim_{\tilde{X}} \omega'$, and ii) there exists a $\omega, \omega' \in \Omega$ such that $\omega \sim_{\tilde{X}} \omega'$ but $\omega \not\sim_X \omega'$. If $\tilde{X}$ is a coarsening of X, then X is a **refinement** of $\tilde{X}$. This definition captures the intuitive idea that $\tilde{X}$ is a coarsening of X iff the partition of possibility space achieved by $\tilde{X}$ is such that any possibilities treated as equivalent by X are also equivalent according to the coarsening $\tilde{X}$, but that some possibilities treated as equivalent by $\tilde{X}$ are treated as distinct by the more fine-grained X.

Let $\mathcal{V}_\mathcal{P}$ be a set of random variables that are all measurable with respect to the same probability space $\mathcal{P} = (\Omega, \Sigma, p)$. Let $\mathcal{E}$ be an acyclic set of ordered pairs, or edges, relating the variables in $\mathcal{E}$. The set of edges $\mathcal{E}$ allows us to define parent and descendant relations between variables in the obvious way. A **causal Bayes net** is a pair $\mathcal{G}_\mathcal{P} = (\mathcal{V}_\mathcal{P}, \mathcal{E})$ such that: i) according to the probability distribution p, all elements of $\mathcal{V}_\mathcal{P}$ are independent of their non-descendants, conditional on their parents (**Markov Condition**), ii) there is no set of edges $\mathcal{E}^* \subset \mathcal{E}$ such that $(\mathcal{V}_\mathcal{P}, \mathcal{E})$ satisfies the Markov condition according to the probability distribution p in the probability space with respect to which all elements of $\mathcal{V}_\mathcal{P}$ are measurable (**Minimality Condition**), and iii) no variable in $\mathcal{V}_\mathcal{P}$ is a coarsening of or identical to any other variable in $\mathcal{V}_\mathcal{P}$ (**Co-possibility Condition**). These conditions ensure that the graphical structure induced by $\mathcal{E}$ represents all causal dependencies between the variables in $\mathcal{V}_\mathcal{P}$ without any excess edges, and that any dependencies between variables are causal, rather than logical, in nature. A variable X causes Y according to $\mathcal{G}_\mathcal{P}$ just in case Y is a descendant of X.

For any given causal Bayes net $\mathcal{G}_\mathcal{P}$, we can calculate the probability distribution over any variable V in the set $\mathcal{V}_\mathcal{P}$, given an intervention setting some set of variables $\mathbf{X}$ to some set of values $\mathbf{x}$, using the following formula:

$$p_{\mathcal{G}_\mathcal{P}}(v|do(\mathbf{x})) = \begin{cases} p(v|\,\mathrm{par}_{\mathcal{G}_\mathcal{P}}(V)) & \text{if } V \notin \mathbf{V} \\ 1 & \text{if } V \in \mathbf{X} \text{ and } v \in \mathbf{x} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where $\mathrm{par}_{\mathcal{G}_\mathcal{P}}(V)$ denotes the values taken by the parents of $V_\mathcal{P}$ in $\mathcal{G}_\mathcal{P}$. 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.

Measuring Information Loss Let $\mathcal{G}_\mathcal{P}$ be a causal Bayes net, and let $\mathcal{G}'_\mathcal{P}$ be a graph generated by replacing the set of variables $\mathbf{X}$ with the set of variables $\mathbf{X}'$, and the set of edges $\mathcal{E}$ with the set of of edges $\mathcal{E}'$. All variables in both Bayes nets are measurable with respect to the same probability space $\mathcal{P}$. Thus, they are taken to represent the same underlying data-generating process. The **information loss** due to the change from $\mathcal{G}_\mathcal{P}$ to $\mathcal{G}'_\mathcal{P}$, with respect to an effect variable Y and the change in variables from $\mathbf{X}$ to $\mathbf{X}'$, is given by the equation

$$\begin{aligned} &\mathcal{L}(\mathcal{G}_\mathcal{P}, \mathcal{G}'_\mathcal{P}, \mathbf{X}, \mathbf{X}', Y, q) \\ &\quad = \sum_{\mathbf{x}} q(do(\mathbf{x})) \sum_{y} p(y) \log_2 \frac{p(y)}{p_{\mathcal{G}_\mathcal{P}}(y|do(\mathbf{x}))} \\ &\qquad - \sum_{\mathbf{x}'} q(do(\mathbf{x}')) \sum_{y} p(y) \log_2 \frac{p(y)}{p_{\mathcal{G}'_\mathcal{P}}(y|do(\mathbf{x}'))} \end{aligned} \quad (2)$$

where q is a probability distribution over possible interventions on $\mathbf{X}$ and $\mathbf{X}'$. In information-theoretic language, information loss is the difference between the average **Kullback-Leibler** divergence between the marginal distribution over Y and the distribution over Y given an intervention on $\mathcal{G}_\mathcal{P}$ setting $\mathbf{X}$ to $\mathbf{x}$, and the average Kullback-Leibler divergence between the marginal distribution over Y and the distribution over Y given an intervention on $\mathcal{G}'_\mathcal{P}$ setting $\mathbf{X}'$ to $\mathbf{x}'$. Where information loss is negative, information is gained rather than lost in the move from $\mathcal{G}_\mathcal{P}$ to $\mathcal{G}'_\mathcal{P}$.

Proportionality According to Woodward, a causal relationship is proportional to the extent that it is stated at the

"level [of causal description] that is most informative about the conditions under which the effect will and will not occur" (2021, 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 (2021, p. 371). This can be made precise in terms of information loss. Let $\mathbf{G}_{\mathcal{P}} = (\mathbf{G}^1_{\mathcal{P}}, \ldots, \mathbf{G}^n_{\mathcal{P}})$ be a series of causal Bayes nets, such that the only difference between two causal Bayes nets $\mathbf{G}^i_{\mathcal{P}}$ and $\mathbf{G}^{i+1}_{\mathcal{P}}$ is the replacement of a single variable with a coarsening thereof.[1] This yields a sequence of variables $\mathbf{C} = (C_1, \ldots, C_n)$, with each C_i a variable in the causal Bayes net $\mathbf{G}^i_{\mathcal{P}}$ 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 Y to the extent that $\mathcal{L}(\mathbf{G}^j_{\mathcal{P}}, \mathbf{G}^i_{\mathcal{P}}, \{C_j\}, \{C_i\}, Y, q)$ 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 Y will change, as compared to more fine-grained 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 results generalize to such comparisons.

Stability Recall that the stability of a causal relationship is its robustness to changes in background conditions. This implies that if a causal Bayes net $\mathcal{G}_{\mathcal{P}}$ contains a stable causal relationship between a cause X and effect Y, then one can eliminate variables representing background conditions from $\mathcal{G}_{\mathcal{P}}$ without losing any information contained in the relationship between interventions on X and changes in Y. This admits of straightforward formalization in terms of information loss. Let $\mathcal{G}_{\mathcal{P}} = (\mathcal{V}_{\mathcal{P}}, \mathcal{E})$ be a causal Bayes net containing a cause X, an effect Y, and a set of background variables $\mathbf{B}$. Let $\mathcal{G}^{-\mathbf{B}}_{\mathcal{P}} = (\mathcal{V}^{-\mathbf{B}}_{\mathcal{P}}, \mathcal{E}^{-\mathbf{B}})$ be a causal Bayes net such that $\mathcal{V}^{-\mathbf{B}}_{\mathcal{P}} = \mathcal{V}_{\mathcal{P}} \setminus \mathbf{B}$ and $\mathcal{E}^{-\mathbf{B}} = \mathcal{E} \setminus \{(W,Z) : W \in \mathbf{B} \vee Z \in \mathbf{B}\}$. That is, $\mathcal{G}^{-\mathbf{B}}_{\mathcal{P}}$ is just $\mathcal{G}_{\mathcal{P}}$ with all variables in $\mathbf{B}$ removed. The causal relationship between X and Y is **stable** with respect to background condition $\mathbf{B}$ to the extent that the value of $\mathcal{L}(\mathcal{G}_{\mathcal{P}}, \mathcal{G}^{-\mathbf{B}}_{\mathcal{P}}, \{X\} \cup \mathbf{B}, \{X\}, Y, q)$ is low.

Summary The preceeding formalizations of the proportionality and stability of causal relationships show how the task of measuring both properties can be subsumed under a more general measure of information loss. In what follows, we present experiments designed to test whether participants' evaluations of the quality of a compressed causal claim are predicted by the amount of information that is lost by choosing that claim over a less **compressed** alternative, where a causal claim is compressed to the extent that it elides either fine-grained details about the cause of some effect or the background conditions moderating the relationship between cause and effect.

Previous Work From a theoretical perspective, the work that is closest to our framework consists of previous attempts to quantify properties of causal relationships in Bayesian networks using tools from information theory. These include specific attempts to measure proportionality and stability (Pocheville, Griffiths, & Stotz, 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 & Polani, 2008; Korb, Nyberg, & Hope, 2011; Griffiths et al., 2015; Hoel, 2017; Beckers & Halpern, 2019; Bourrat, 2021). However, 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.

On the experimental side, Lien and Cheng (2000) present evidence effectively showing that humans prefer to justify their inferences using more proportional causal claims, although they do not use the term 'proportional'. By contrast, Bechlivanidis, Lagnado, Zemla, and Sloman (2017) find that participants prefer causal explanations with more detail to those with less detail, even when the less detailed explanations are just as proportional. However, their experiments ask participants to evaluate explanations of specific events rather than type-level causal claims. Vasilyeva, Blanchard, and Lombrozo (2018) show that participants are more willing to endorse causal and explanatory claims with high stability, even when other factors are held fixed. However, to our knowledge, there is no work aiming to empirically investigate whether there is a unifying explanation of the preference for both proportional and stable causal claims.

Experiments

Experiment 1

In Experiment 1, we hypothesized that when more information that is lost when a less compressed causal claim is replaced with a more compressed causal claim, the more compressed claim will be evaluated less positively by participants relative to the less compressed claim. To test this, 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. We manipulated the vignette used, the amount of information loss realized by the more compressed causal claim, and whether the compression was achieved by coarse-graining a variable (thus manipulating proportionality) or eliding a background variable (thus manipulating stability).

[1] Formally, for any $i < n$ there is a bijection $f : \mathcal{V}^i_{\mathcal{P}} \to \mathcal{V}^{i+1}_{\mathcal{P}}$ such that $f(C_i) = C_{i+1}$, where C_{i+1} is a coarsening of C_i, and for all $V_i \in \mathcal{V}^i_{\mathcal{P}} \setminus \{C_i\}$, $f(V_i) = V_i$. There is also a bijection $g : \mathcal{E}^i \to \mathcal{E}^{i+1}$ 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 $Z_g = Z$.

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

Table 1: Structure of vignettes used in both experiments.

Participants Participants were 450 adults recruited via Prolific. 150 additional participants were excluded for failing comprehension checks or for rating poor causal claims non-negatively. For both studies, participation was restricted to users with a US-based IP address and a 95% rating based on at least 100 previous studies. Both studies were pre-registered, and IRB approval was obtained from Princeton University. Data, stimuli, and pre-registrations are available at https://osf.io/prmu6/?view_only=90aee64c5b0943b0a1afbabebcc268e6.

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 mushroom vignette, participants were presented with one of the following descriptions of results of experiments on the "Drol" mushroom:

> D-1: a) x% of all Drol planted in high-mineral, high-sodium soil have bumpy stems; b) 70% of all Drol planted in high-mineral, low-sodium soil have bumpy stems; c) 1% of all Drol planted in low-mineral, high-sodium soil have bumpy stems; d) 1% of all Drol planted in low-mineral, low-sodium soil have bumpy stems.
>
> D-2: a) x% of all Drol planted in high-mineral soil and watered with salty water have bumpy stems; b) 70% of all Drol planted in high-mineral soil and watered with fresh water have bumpy stems; c) 1% of all Drol planted in low-mineral soil and watered with salty water have bumpy stems; d) 1% of all Drol planted in low-mineral soil and watered with fresh water have bumpy stems.

The value of x was varied between subjects and set at either 70, 85, or 98. 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 the findings of the descriptions given above:

- *Compressed:* Planting Drol in high-mineral soil causes them to have bumpy stems.
- *High:* Planting Drol in [high-mineral, high-sodium soil/high mineral soil and watering them with salty water] causes them to have bumpy stems.
- *Low:* Planting Drol in [high-mineral, low-sodium soil/high mineral soil and watering them with fresh water] causes them to have bumpy stems.

For participants shown Description 1, the claim *Compressed* is a compression achieved by coarsening a causal variable. For participants shown Description 2, the claim *Compressed* is a compression achieved by eliding a background variable.[2] The values of x correspond to information loss amounts for *Compressed* of 0, .03, and .31 respectively, assuming a uniform distribution over possible interventions. Vignettes involving flies and rocks followed an identical structure (see Tab. 1). Participants were randomly assigned to one of eighteen possible conditions, which differed with respect to which of the three vignettes they were shown, whether they were asked to evaluate a compressed claim achieved by coarsening a causal variable (proportionalilty) or eliding a background variable (stability), and the amount of information loss inherent in compression. Finally, participants were also asked to evaluate three poor causal claims, constructed by substituting the value of the primary causal factor (e.g., changing high-mineral to low-mineral). These were included to help anchor the scale and verify participant understanding; we do not discuss them further here in the interest of space.

Results 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:

- V-A. The difference between the participant's evaluation of *Compressed* and their evaluation of *High* (e.g., the difference between the evaluation of 'Planting Drol in high-mineral soil causes them to have bumpy stems' and the evaluation of 'Planting Drol in high-mineral, high-sodium soil causes them to have bumpy stems').
- V-B. 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 'Planting Drol in high-mineral soil causes them to have bumpy stems' and the average evaluation of 'Planting Drol in high-mineral, high-sodium soil causes them to have bumpy stems' and 'Planting Drol in high-mineral, low-sodium soil causes them to have bumpy stems').[3]

[2] 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.

[3] Due to an error, the equation for V-B 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*).

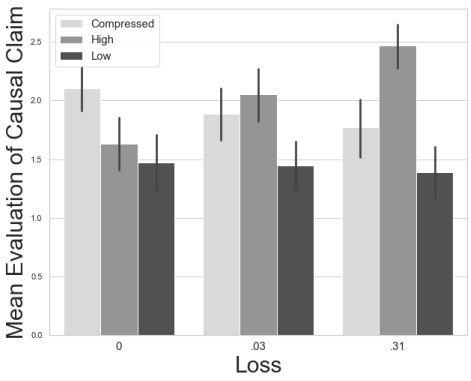

Figure 1: Mean evaluations of claims in Experiment 1, with bars showing 95% CIs. 'Loss' corresponds to information loss due to compression inherent in choosing *Compressed* over *High* and *Low*.

The score V-B measures a participant's preference for a more compressed claim over *either* of the more detailed claims.

We regressed these dependent variables against independent variables denoting the assigned vignette (Vignette), whether the more compressed claim manipulated proportionality or stability (Condition), and the amount of information loss (Loss), as well as all possible interactions. The regressions revealed that only Loss was a significant predictor of V-A ($\beta = -2.67$, $p = .004$). However, Loss was not a significant predictor of V-B ($\beta = -1.23$, $p = .169$). We did find a significant relationship between Vignette and V-B ($\beta = -.48$, $p = .025$), though we refrain here from interpreting this result. Notably, we found no evidence of a significant effect of Condition on these dependent variables (V-A: $\beta = -.20$, $p = .284$; V-B: $\beta = -.16$, $p = .379$), nor did we find any significant interaction effects between Condition and any other independent variables.

As a sanity check, we also analyzed the difference between the participant's evaluation of *High* and their evaluation of *Low* (V-C). As expected, only Loss was a significant predictor of V-C, with the value of V-C increasing as the probability of the effect given the description of the cause in *High* increases with Loss ($\beta = 2.89$, $p < .001$).

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 36% when Loss=0, 21% when Loss=.03, and 10% when Loss=.31. Mixed ANOVA for each value of Loss found that at Loss=0, *Compressed* was rated more highly than both *High* ($\eta^2 = .014$, $p = .018$) and *Low* ($\eta^2 = .022$, $p = .003$). When Loss=.03, *High* was not rated significantly higher than *Compressed* ($\eta^2 = .007$, $p = .096$), but was rated higher than *Low* ($\eta^2 = .035$, $p < .001$). When Loss=.31, *High* was rated significantly higher than both *Compressed* ($\eta^2 = .064$, $p < .001$) and *Low* ($\eta^2 = .145$, $p < .001$).

Discussion These results provide strong evidence in favor of the claim that participants' relative evaluations of more and less compressed causal claims is partially governed by the amount of information loss that is inherent in the more compressed causal claim. As can be seen in Fig. 1, which plots participants' absolute evaluations of each causal claim at each loss level, the lack of a significant effect of Loss on V-B is due to the fact that increases in the difference between evaluations of *High* and *Comp* are accompanied by an increase in the difference between *Comp* and *Low*. When there is no information loss, participants evaluate more compressed causal claims significantly more highly than less compressed causal claims, 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. That no evidence was found for any effect of Condition supports the thesis that the proportionality and stability of a causal claim are both measured by information loss.

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 asked first to evaluate *Compressed* and second to independently evaluate *High* and *Low*.

Participants 483 adults were recruited via Prolific. 117 additional participants were excluded for failing comprehension checks or rating poor causal claims non-negatively.

Materials and Procedures The procedure was identical to that used in Experiment 1 with three exceptions. First, as described above, participants were asked to evaluate *Compressed* as part of a separate task than their evaluation of *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 (a) and (b) was varied between subjects and set at either 55, 85, or 98, leading to information loss amounts of 0, .07, and .41 respectively. Thus, we replicated Experiment 1 for a different range of loss values.

Results We performed the same regressions as in Experiment 1. Loss was a significant predictor of all three dependent variables (V-A: $\beta = -2.97$, $p < .001$, V-B: $\beta = -1.68$, $p = .002$, V-C: $\beta = 2.57$, $p < .001$). Fig. 2 shows the relationship between Loss and participants' absolute evalua-

tions of *Compressed*, *High*, and *Low*. Only the evaluation of *High* is shown to be significantly linearly predicted by Loss ($\beta = 2.81$, $p < .001$). This suggests that the relationship between Loss and the three dependent variables is driven primarily by an increased evaluation of *High* that is not accompanied by an increased evaluation of *Compressed* or *Low*.

In a further replication of Experiment 1, Condition was not a significant predictor of any of the three dependent variables measured (V-A: $\beta = .228$, $p = .124$; V-B: $\beta = .218$, $p = .101$; V-C: $\beta = -.021$, $p = .875$). We did observe that the interaction between Condition and Loss was a significant predictor of V-A ($\beta = -1.32$, $p = .032$) and that the interaction between Vignette and Loss was a significant predictor of V-C ($\beta = -1.73$, $p = .009$). We refrain from interpreting these results here as they only hold for some dependent variables.

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=.07, and 2% when Loss=.41. Mixed ANOVA for each value of Loss found that at Loss=0, *Compressed* was rated more highly than both *High* ($\eta^2 = .038$, $p < .001$) and *Low* ($\eta^2 = .049$, $p < .001$). At Loss=.07, *High* was rated more highly than *Compressed* ($\eta^2 = .029$, $p < .001$) and *Low* ($\eta^2 = .188$, $p < .001$). At Loss=.41, *High* was rated more highly than *Compressed* ($\eta^2 = .151$, $p < .001$) and *Low* ($\eta^2 = .386$, $p < .001$).

Discussion The results of Experiment 2 replicate the positive results of Experiment 1 at a different range of loss levels, with the additional finding of a significant relationship between Loss and V-B, under conditions such that *Compressed* is evaluated separately from *High* and *Low*. This renders concerns about the pragmatics of the task less plausible.

General Discussion

These experiments provide evidence that when evaluating more and less compressed causal descriptions of the same process, we engage in a trade-off between compression on the one hand and information loss on the other. Our findings support a unified account of proportionality and stability, and take a first step towards understanding the relationship between causal cognition and compression. The trade-off we observe may also be relevant to people's evaluations of non-causal claims, such as descriptions of statistical patterns.

Nevertheless, alternative explanations of our findings remain plausible. To illustrate, consider the causal power measure due to Cheng (1997). For two events c and e, the causal power of c with respect to e is $p(e|c) - p(e|\neg c)$. If we let e be the development of bumpy stems in Drol and let c be either planting Drol in high-mineral, high-sodium soil or planting Drol in high-mineral soil and watering them with salty water, then as Loss increases in both of the current experiments, so too does the implicit causal power of c with respect to e in either version of the claim *High*. In both experiments, the correlation between all three dependent variables and Loss was driven by an increased absolute evaluation of *High* as

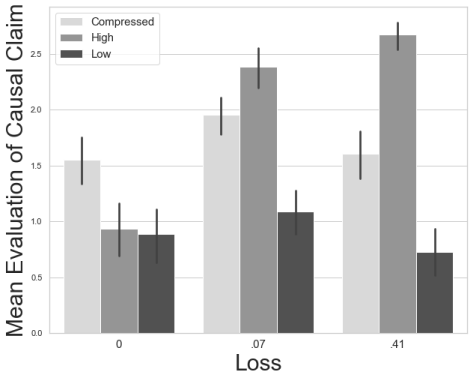

Figure 2: Mean evaluations of claims in Experiment 2, with bars showing 95% CIs. 'Loss' corresponds to information loss due to compression inherent in choosing *Compressed* over *High* and *Low*.

Loss increased. Thus, our results are consistent with the interpretation that participants evaluate *High* more favorably as the power of the causal relationship that it describes increases, while their absolute evaluations of *Compressed* and *Low* remain fixed. However, this interpretation also predicts an increasingly positive evaluation of *Compressed* as Loss increases, which we do not see in our results. Also, causal power does not explain why, when Loss=0, participants prefer more compressed causal claims. That said, further testing is needed to fully rule out a causal power interpretation.

Another direction for future work is to investigate the determinants of people's tolerance for information loss when evaluating compressed causal claims. The exploratory analyses reported in the results of both experiments show that at least *some* participants prefer more compressed causal claims to less compressed ones, even when the more compressed claim leads to information loss. Theoretical work by Brodu (2011), Kinney (2019), and Kinney and Watson (2020) argues 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.

We also plan to test how both information loss and agency impact participants' open-ended summaries of causal patterns. These studies will measure the extent to which participants penalize more detailed causal claims because they are read as ruling out alternatives (e.g., participants may assume that the claim 'planting Drol in high-mineral, high-sodium soil causes them to develop bumpy stems' implies that planting Drol in high-mineral, low-sodium soil does *not* cause them to develop bumpy stems). While pragmatic factors are sure to play a role in communication, we nevertheless anticipate that the trade-off between informativeness and compression describes causal representation more generally, in both intra- and interpersonal contexts.

References

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.

Beckers, S., & Halpern, J. Y. (2019). Abstracting causal models. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 33, pp. 2678–2685).

Bourrat, P. (2021). Measuring causal invariance formally. *Entropy*, *23*(6), 690.

Brodu, N. (2011). Reconstruction of epsilon-machines in predictive frameworks and decisional states. *Advances in Complex Systems*, *14*(05), 761–794.

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

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

Griffiths, P. E., Pocheville, A., Calcott, B., Stotz, K., Kim, H., & Knight, R. (2015). Measuring causal specificity. *Philosophy of science*, *82*(4), 529–555.

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

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. In *International conference on probabilistic graphical models* (pp. 257–268).

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

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.

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.

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

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

Pacer, M., & Lombrozo, T. (2017). Ockham's razor cuts to the root: Simplicity in causal explanation. *Journal of Experimental Psychology: General*, *146*(12), 1761.

Pearl, J. (2000). *Causality*. Cambridge University Press.

Pocheville, A., Griffiths, P. E., & Stotz, K. (2017). Comparing causes—an information-theoretic approach to specificity, proportionality and stability. In *Proceedings of the 15th congress of logic, methodology and philosophy of science* (pp. 93–102).

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

Rosch, E. (1999). Principles of categorization. *Concepts: core readings*, *189*, 312–322.

Spellman, B. A. (1997). Crediting causality. *Journal of Experimental Psychology: General*, *126*(4), 323.

Spirtes, P., Glymour, C. N., & Scheines, R. (2000). *Causation, prediction, and search*. MIT press.

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

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

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. (2018a). Causal cognition: Physical connections, proportionality, and the role of normative theory. In *Philosophy of psychology: Causality and psychological subject* (pp. 105–138). De Gruyter.

Woodward, J. (2018b). Explanatory autonomy: the role of proportionality, stability, and conditional irrelevance. *Synthese*, 1–29.

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