Cognition 237 (2023) 105464

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Original articles

Explanations in the wild

Justin Sulik [a,*], Jeroen van Paridon [b], Gary Lupyan [b]

[a] Cognition, Values & Behavior, Ludwig Maximilian University of Munich, Gabelsbergerstrasse 62, Munich 80333, Germany
[b] Department of Psychology, University of Wisconsin–Madison, 1202 West Johnson Street, Madison, WI 53706, USA

ARTICLE INFO

Dataset link: https://osf.io/wbxcj/

Keywords:
Explanations
Creativity
Metacognition
Insight
Perspective-taking

ABSTRACT

Why do some explanations strike people as highly satisfying while others, seemingly equally accurate, satisfy them less? We asked laypeople to generate and rate thousands of open-ended explanations in response to 'Why?' questions spanning multiple domains, and analyzed the properties of these explanations to discover (1) what kinds of features are associated with greater explanation quality; (2) whether people can tell how good their explanations are; and (3) which cognitive traits predict the ability to generate good explanations. Our results support a pluralistic view of explanation, where satisfaction is best predicted by either functional or mechanistic content. Respondents were better able to judge how accurate their explanations were than how satisfying they were to others. Insight problem solving ability was the cognitive ability most strongly associated with the generation of satisfying explanations.

1. Introduction

Explanations are central to the human experience, from serving as answers to children's 'Why?' questions (Legare, 2012; Mills et al., 2019), to motivating scientific theories (Deutsch, 2011; Gopnik, 2000). Historically, the question of what makes a good explanation has been addressed largely by philosophers of science (Brewer et al., 1998; Cummins, 2000). This approach has focused on how well explanations subserve epistemic goals such as accurate prediction (Woodward, 2019) or how well they embody theoretic virtues such as parsimony or simplicity (Kuhn, 1977; Thagard, 1978). It has also influenced cognitive research on explanations, for instance leading to proposals for how Bayesian models of cognition might implement various explanatory virtues (Wojtowicz & DeDeo, 2020), or that test whether people prefer explanations that facilitate prediction (Lombrozo & Carey, 2006) or that are parsimonious (Lombrozo, 2007).

However, lay explanations frequently fall short of scientific standards (Horne et al., 2019; Keil, 2006). A narrow focus on normative standards – such as accurate prediction or theoretic virtue – is therefore likely to misrepresent what kinds of explanations laypeople consider good. Thus, rather than focusing on how scientific theories explain, we focus on how *people* explain. We do this by examining explanations generated by laypeople to answer 'Why?' questions about common phenomena; by having other laypeople evaluate the quality of those explanations; and by trying to understand what predicts these evaluations of quality.

Rather than offering a normative account of explanation quality, we ask a descriptive question: What makes an explanation satisfying? Satisfaction predicts children's preferences for and recall of explanations (Frazier et al., 2016); it plays a role in motivations for and perceptions of learning (Liquin & Lombrozo, 2022); and it also tracks people's evaluations of aforementioned virtues such as parsimony across different contexts (Lim & Oppenheimer, 2020). As explanations play a cognitive role in guiding discovery, exploration, generalization and learning (both when people produce explanations and request them from others, Frazier et al., 2016; Gopnik, 2000; Liquin & Lombrozo, 2022; Mills et al., 2019; Walker et al., 2014; Williams & Lombrozo, 2010), satisfaction offers a window into the psychology of explanation.

More specifically, we wish to understand what kind of content makes everyday explanations satisfying. How do explanations generated by laypeople strike other laypeople? Viewing explanations as communicative acts (Faye, 2007; Keil, 2006) expands the proposal that human reasoning is not – as traditionally thought – geared solely towards the production of accurate knowledge for the reasoner, but towards persuading the reasoner's audience (Mercier & Sperber, 2011; Mercier & Strickland, 2012). The central problem, then, is what predicts people's evaluations of how satisfying an explanation is, while controlling its perceived accuracy.

Our 'explanations in the wild' approach – where laypeople generate free-form responses to a range of ordinary 'Why?' questions and other laypeople rate how satisfying or accurate they find these responses – contrasts with or complements previous empirical studies of explanation.

One common approach is to have participants evaluate explanations that were carefully created by experimenters to contain specific features

[*] Corresponding author.
E-mail address: justin.sulik@gmail.com (J. Sulik).

https://doi.org/10.1016/j.cognition.2023.105464
Received 29 June 2022; Received in revised form 10 April 2023; Accepted 12 April 2023
Available online 3 May 2023

Table 1
Domains and example questions. A question may well fall within multiple domains.

Domain	Example question
Socio-cultural	Why are there so many languages in the world?
Psychology	Why do people bite their nails?
Neuroscience	Why do we need sleep?
Biology	Why are polar bears white?
Chemistry	Why are snowflakes hexagonal?
Physics	Why are there waves in the ocean?

of interest (Colombo et al., 2017; Hopkins et al., 2016; Lombrozo & Carey, 2006); another is to analyze participants' explanations of artificial scenarios crafted by experimenters for similar purposes (Lombrozo & Gwynne, 2014). Experimental control is certainly valuable, yet these studies offer limited insight into the kinds of explanations that are produced and shared by non-experts in everyday contexts. Such explanations are of interest because the majority of people are non-experts in any given field, hence many of the explanations people produce and encounter in their regular lives are lay explanations. Our approach is thus a useful complement to researcher-generated explanations.

Other studies have used explanations not generated by researchers, some of which we might call 'curated' rather than 'wild'. For instance, Liquin and Lombrozo (2022) harvested explanations from published Q&A books and textbooks. We assume that explanations appearing in published books went through *some* degree of quality control, whereas we have participants generate explanations with no explicit filter on quality (as long as they tried to answer the question), allowing us to examine natural variation in perceived quality. Relatedly, the explanations studied by Liquin and Lombrozo (2022) were at least 34 words long, but to the extent that people conversationally address 'Why?' questions in a few words in their daily lives ('Why are you late?' 'Traffic.'), our approach offers an informative glimpse into everyday explanations. If levels of – and variability in – quality differ between lay explanations and expert explanations or between everyday more conversational explanations and explanations curated for publishing, factors predicting explanation quality may behave differently 'in the wild'.

Zemla et al. (2017) analyzed the properties of explanations harvested from an online forum, with participants evaluating the explanations on properties such as internal coherence, generality, and scope. However, some of the explanations were provided by domain experts rather than by laypeople. For example, one explanation of why Ebola is hard to contain was provided by a biomedical scientist on an Ebola response team. The expertise of these explainers makes it difficult to generalize such results to the kind of explanations that laypeople generate and share, because laypeople may be designing their explanations based on entirely different considerations. A strength of that study was the sheer number of properties rated for each explanation, whereas we offer a complementary approach by analyzing a large number of unique explanations. Zemla et al. harvested 24 explanations (three each for eight *explananda*, focusing on socio-historical topics), whereas we had participants generate 2883 explanations across 50 questions representing a range of domains (illustrated in Table 1). These explanations were evaluated by 5367 unique raters, yielding 117359 individual ratings.

Other work considers an even larger set of explanations than we do, and harvests them from conversational interactions rather than texts published in books or internet forums (which is thus closer to our understanding of 'in the wild' rather than 'curated'). For instance, Hickling and Wellman (2001) extracted around 5000 explanations from a corpus of child speech to study the development of explanations in terms of what children seek to explain (e.g., people, animals, physical objects) and the modes of explanation they recruit (e.g., psychological, biological, physical). Hickling and Wellman report that children flexibly produce causal explanations from a young age. Complementing this with a slightly more structured approach (and in adults), we have multiple participants answer the same 'Why?' questions, thereby allowing us to explore variation within answers to the same question and variation across answers to different questions.

We use the 'explanations in the wild' approach to answer three questions about the psychology of everyday explanation. First, what features of explanations are associated with greater satisfaction? Second, how well are people calibrated in their rating of explanations? Third, what cognitive traits are associated with the ability to provide satisfying explanations?

Our first aim is to study the features of explanations spontaneously produced by non-experts across diverse topics in order to understand how they contribute to perceptions of satisfaction. To decide which features to measure, we begin with the following description of one kind of explanation common in everyday life: 'an attempt to understand a causal relation by identifying relevant functional or mechanistic information' (Legare, 2014). Other kinds of explanation exist, such as mathematical explanation (Mancosu, 2001; Mejía-Ramos et al., 2019) and formal explanation (Prasada, 2017). However, as causation seems to be central to how philosophers (Stalnaker, 1984) and children (Hickling & Wellman, 2001) understand the world around us, we think this is a reasonable starting point for our study of everyday explanation in lay adults.

To assess the causal, mechanistic and functional content of explanations, we recruited non-expert raters (who did not produce the explanations) to rate how much each explanation appeals to commonsense understandings of causation (e.g., World War II was *sparked* by an assassination), function (e.g., a bird has wings *in order to fly*), or mechanism (e.g., electricity makes a bulb glow *by heating the filament*). Additionally, as an early thread in the philosophy of science framed explanation as appeal to general laws such as gravity (Hempel, 1965), but as we suspected that appeals to universal laws like this would be infrequent in non-expert explanations, we posited that one aspect of such laws – their generality – would be both common-place and easy for non-experts to understand. Thus, we also had participants rate how general the explanations were.

Part of the point of assessing these content features in explanations in the wild is to understand whether (and if so, how) lay appeals to causation, function and mechanism differ from expert conceptions informed by the philosophy of science. From a normative or logical perspective, functional explanations are ultimately causal (Lombrozo & Wilkenfeld, 2019; Wright, 1976) and experiments with researcher-generated explanations confirm that participants are sensitive to the link between function and causation (Lombrozo & Carey, 2006). However, from a psychological perspective, it does not necessarily follow that people will always spontaneously mention the logical causal link when generating functional explanations in the wild, or that the audience will spontaneously attend to the causal link (if given) or take it as implied (if absent), or even that they will process this information (whether given or implied) when evaluating explanation satisfaction. Thus, we ask whether the expression of causal content is perceived to be necessary for functional explanations in the wild as it is in more controlled contexts. Similarly, this approach allows us to disentangle how people generate and evaluate mention of causes from how they generate and evaluate mention of the mechanisms whereby causes bring about their effects.

Our second aim is to probe a metacognitive question: Do laypeople know how good their own explanations are? If people are well calibrated, then explanation generators and raters should concur in their assessments; otherwise, explanation generators might overestimate their ability to explain. Work on the 'Illusion of Explanatory Depth' (Rozenblit & Keil, 2002) has shown that people overestimate their own understanding before explaining something technical, only to realize the limits of their knowledge once they attempt to generate an explanation. However, for the less-technical everyday explanations considered here, it may be that the illusion of understanding persists even once people explain. Further, people's tendency to overestimate their ability may depend on the level of that ability. If so, people who generate worse explanations may also be less well calibrated in assessing them. This is somewhat analogous to the Dunning–Kruger Effect (Kruger & Dunning, 1999), though we note that the latter is about people's estimation of their own ability percentile vs their actual percentile score, whereas we explore calibration between the person generating the explanation and the other people rating it for quality.

Our final aim is to understand who is most likely to produce accurate or satisfying explanations. This is important for understanding which individual differences matter for explanation quality, and thus for uncovering which psychological mechanisms are at work in generating explanations. Several cognitive traits might contribute. If producing a good explanation is a matter of knowing the right facts, then more knowledgeable people will generate better explanations. If the *search* for information is crucial, then explanation quality might be predicted by how deeply a participant searches through their knowledge, not just by the extent of that knowledge. If so, better explanations may be generated by people who engage in effortful or reflective processing, or who are more curious. If a challenge in generating a high-quality explanations is working out what is relevant to begin with (as explanations are ill-defined problems, Horne et al., 2019), the ability to generate good explanations will depend on *insight*, the ability to creatively form a relevant problem representation (Bowden et al., 2005; Durso et al., 1994; Sulik, 2018). Finally, as we are construing explanations as communicative social acts (Faye, 2007; Keil, 2006), a person's ability to generate an answer that satisfies the question-asker may depend on their ability to take the question-asker's perspective.

We begin, in Study 1, by asking how perceived features of explanations (causation, function, mechanism, generality) predict quality measured through perceived accuracy and satisfaction. We also evaluate metacognitive calibration by comparing ratings of explanation quality made by people who generated the explanations vs. other independent raters. In Study 2, we administer individual-differences measures, and identify which cognitive traits predict the ability to produce good explanations. In response to questions raised by an anonymous reviewer regarding Study 1, we also ran a follow-up study ('Study 1a') which, though last chronologically, appears here between Studies 1 and 2.

2. Study 1: What makes an explanation satisfying?

2.1. Methods

2.1.1. Participants

We recruited participants from Amazon's Mechanical Turk (MTurk) platform. Participation was limited to those with an IP address in the USA and over a 95% approval rating on MTurk.

In Phase 1 (explanation generation, N = 224) participants were paid $0.50 to produce explanations and provide basic demographics. We aimed to collect 1000 explanations, which would offer almost 90% power to detect a small correlation ($r = .1$). This meant a minimum of 200 participants, as we aimed to collect 20 explanations for each of 50 'Why?' questions, where each participant answered 5 questions. If, due to random assignment, a question had too few explanations, we recruited more participants to fill the gap, hence needing more than the minimum.

In Phase 2 (explanation rating, N = 3118) participants were paid $0.35 to $0.45 to assess explanations. We aimed to collect 10 ratings per explanation per feature (this number yields relatively stable regression coefficients for subjective judgments, Motamedi et al., 2019), which meant a minimum of 3000 participants. Again, due to gaps from random assignment, we ultimately needed to recruit more than the minimum.

The study was approved by the University of Wisconsin–Madison Education and Social/Behavioral Science IRB.

2.1.2. Procedure

All materials are available at https://osf.io/wbxcj/.

We first generated a list of 50 'Why?' questions that were intended to cover a range of domains (Table 1) covered by empirical sciences, including social sciences. On one hand, this ensures that our results with non-experts could serve as a useful complement to explanation in the philosophy of science, which often considers explanation in this context. On the other hand, it also offers the opportunity for future work to bridge our open data on explanation in lay adults with the developmental literature, as young children seek explanations in physical, biological and socio-cultural domains (Hickling & Wellman, 2001).

In Phase 1, after providing informed consent, participants were randomly assigned five 'Why?' questions from the list of 50. Participants were asked to provide as good an explanation for each question as they could, in a free-response text box (and were asked not to google the answers). Then, participants rated their own explanations according to how satisfying and accurate they were. All ratings described here were on 7-point Likert scales. Finally, participants provided their age, gender and highest education level.

In Phase 2, we had each explanation assessed on six features: two aspects of explanation quality (perceived accuracy and satisfaction) and four types of content (perceived causation, mechanism, function, generality). Full instructions for eliciting all the ratings are available at https://osf.io/wbxcj/, though we briefly describe them below.

For the accuracy ratings, participants were simply instructed 'Please rate each explanation on how accurate or correct you think it is.' For the satisfaction ratings, they were instructed 'Please rate each explanation on how satisfying you think it is,' where this was later unpacked as follows: 'The answer could be true or accurate, but still be unsatisfying. For instance, if someone explains why deer have antlers by simply saying "Evolution", then this answer is correct, but it would not satisfy someone who wonders why they evolved that way. So try think about how appealing you think the answer is as a whole, not just whether it is true.'

In the instructions to raters, examples of causation included cigarettes causing lung cancer, or a ball moving because it was kicked. Mechanistic information was described as *how* something happens, so the following explanation of light bulbs – 'The flow of electrons heats the wire, and hot things glow' – contains some information about mechanism as it can be paraphrased '*by heating the wire*, the flow of electrons causes it to glow.' Function was described in terms of a goal or purpose, so the function of hearts is to pump blood, and it is possible to paraphrase this as 'hearts are *for pumping* blood'. Finally, in describing generality, participants were given examples of statements that are not general as they hold rarely ('Today is June 29 2022', which is true for a limited time) and statements that are very general ('Triangles have 3 sides', which is always true).

After providing informed consent, participants were told that they would see about 20 answers to one 'Why?' question, and would have to rate these according to a given feature. Questions and features were between-subjects variables: No participant rated more than one question or more than one feature. For analyses below, we averaged the approximately 10 ratings per feature per explanation to yield a numeric value (see https://osf.io/wbxcj/ for details of data quality indexes, which were used to exclude careless responses prior to averaging). Thus, each explanation could be more-or-less causal, more-or-less functional, more-or-less mechanistic, etc., as some explanations contained multiple kinds of information.

2.2. Results

All data and full analysis scripts (including full model specifications) are available at https://osf.io/wbxcj/. For the regressions, we specified weakly informative priors ($\beta \sim \mathcal{N}(0,1)$). As some explainers likely produced less satisfying explanations than others, we included a random

Table 2
Example explanations and features.

Feature	Question	Bottom quartile	Top quartile
Mechanism	Why does thunder make a noise?	Because of sound waves.	I believe thunder is caused by lightning affecting the air around it. The air expands quickly, either quick enough for a crack or a rumbling sound, because the lightning increases air pressure and temperature causing the sound of thunder.
Generality	Why do our noses run when we eat spicy food?	When our body temperature in our mouth rises our body believes there is something harmful entering our system so it releases mucus in an effort to push that harmful substance out.	Spicy foods contain capsaicin and capsaicin irritates the mucus membranes in our nose. This irritation causes our noses to run.
Function	Why do we dream?	We dream when we are in a sleep stage during rapid eye movement and it is part of everyday normal life.	We dream to consolidate/solidify memories, emotions, etc.
Causation	Why are flowers colorful?	To attract bees and other pollinators to allow the flowers to reproduce.	Flowers are each made of their own DNA. The DNA of a flower determines many factors including the color. It is possible to alter the seeds planted to change the color of the flower to what you would like it to be.

intercept for explainer. As some questions may have been easier to provide satisfying answers for than others, we also included a random intercept for question. As the ratings were averaged per feature per explanation, the data did not include a row for each rating, and we thus did not include a random intercept for each rater.

Details of rater agreement are available in SM1, as well as at the above OSF link.

2.2.1. Descriptive overview

Table 2 illustrates the content features with examples from the top and bottom quartiles of the ratings of each feature.

Overall, our explanations-in-the-wild approach produced explanations that were judged to be moderately accurate (M = 4.72, SD = 1.13), as well as satisfying (M = 4.31, SD = 1.13). Importantly for predicting explanation quality, there was a reasonable amount of variance in both variables (Fig. 1a).

Although perceived satisfaction and perceived accuracy were strongly correlated ($r = 0.65$, $p < 0.001$), they were nonetheless distinct as indices of explanation quality. Fig. 1a illustrates this with two explanations where the two quality features align (near the red reference line) and two explanations where they do not (lying outside the cyan reference lines). It seems that it is hard for an explanation to be satisfying if it is not perceived as accurate (there are few cases where satisfaction is more than one Likert rating higher than accuracy) but it is easy to be perceived as accurate without being satisfying (there are many cases where accuracy is more than one Likert rating higher than satisfaction).

2.2.2. What features of explanations predict greater satisfaction?

Fig. 1b shows zero-order correlations between quality and content features. It also includes explanation length, operationalized as the number of unique words in the explanation, excluding those found in the question and also excluding grammatical words (e.g., 'and', 'the' or 'is'). All correlations were significantly positive ($p < 0.001$) except the correlation between length and generality ($r = 0.06$, $p = 0.059$). Of the content features, mechanism had the strongest correlation with satisfaction, and causation the weakest. Explanation length correlated most strongly with mechanism and satisfaction.

We then built a series of regression models predicting ratings of satisfaction. As satisfaction predicts perceptions of learning but not necessarily accuracy of learning (Liquin & Lombrozo, 2022), and as satisfaction's relationship with parsimony varies according to context (Lim & Oppenheimer, 2020), people need not *always* value explanations according to how accurate and simple they perceive them to be. Thus, we wanted to understand how content features predicted satisfaction both with and without accuracy and explanation length as controls. Fig. 1c shows standardized coefficients from a Bayesian multiple linear regression predicting satisfaction from just the content features. All variables predicted unique variance in satisfaction, together accounting for half of the variance ($R^2 = 0.50\,[0.46, 0.53]$).

With perceived accuracy added as a covariate (Fig. 1d), some regression coefficients dropped substantially, with generality showing the largest decrease. A Bayesian mediation analysis shows that most of the effect of generality on satisfaction was via perceived accuracy ($\beta = 0.32\,[0.29, 0.36]$), though this still left a smaller direct effect ($\beta = 0.08\,[0.02, 0.14]$).

Next we added explanation length as a covariate (Fig. 1e). Longer explanations were judged to be more satisfying ($\beta = 0.22\;[0.17, 0.26]$). Importantly, the predictive effect of perceived accuracy was relatively unchanged ($\beta = 0.38\;[0.33, 0.43]$), suggesting that the relationship between satisfaction and perceived accuracy is not confounded by explanation length.

The positive relationship between length and satisfaction replicates a finding by Zemla et al. (2017, though that paper talks about 'quality' generally rather than satisfaction or accuracy as distinct hallmarks of quality). It is noteworthy that this relationship is positive, as Zemla et al. propose that the number of words (or in our case, the number of new content word types) tracks level of detail, and is thus one way to operationalize explanation complexity vs simplicity. Yet simplicity is commonly held to be an explanatory virtue (Kuhn, 1977).

Given the large drop in the effect of mechanism on satisfaction when length is added (from Fig. 1d to e), we conjecture that extra detail about mechanism – *how* a cause brings about its effect, rather than the mere presence of causal information – is one way to improve satisfaction at the expense of simplicity. This expands the list of potential reasons why longer explanations may be more satisfying, such as the finding that more complex phenomena require longer explanations (Lim & Oppenheimer, 2020).

2.2.3. Relationships between features of the explanations

As part of understanding what features are associated with ratings of satisfaction, it is worth examining how the various features (mechanism, causality, etc.) are related to one other.

Appeals to causes vs functions represent two common modes of explanation (Keil, 2006). Lombrozo (2010) describes these as 'backward-looking' and 'forward-looking' respectively. The zero-order correlation between function and causation ratings was $r = 0.16$ (see Fig. 1b), but regressing both of these on accuracy revealed a weakly negative residual correlation ($r = -0.08\,[-0.15, -0.01]$). Thus, controlling for perceived accuracy, the more functional an explanation, the lower its rating for causal content (and vice versa). While not speaking to any logical connection between the philosophical categories 'function' and 'causation', this negative association suggests that laypeople sometimes do not spell out causal connections in the context of a functional explanation.

In any case, whether positive or negative (i.e., whether holding accuracy constant or not), the association is small. Mention of causes and

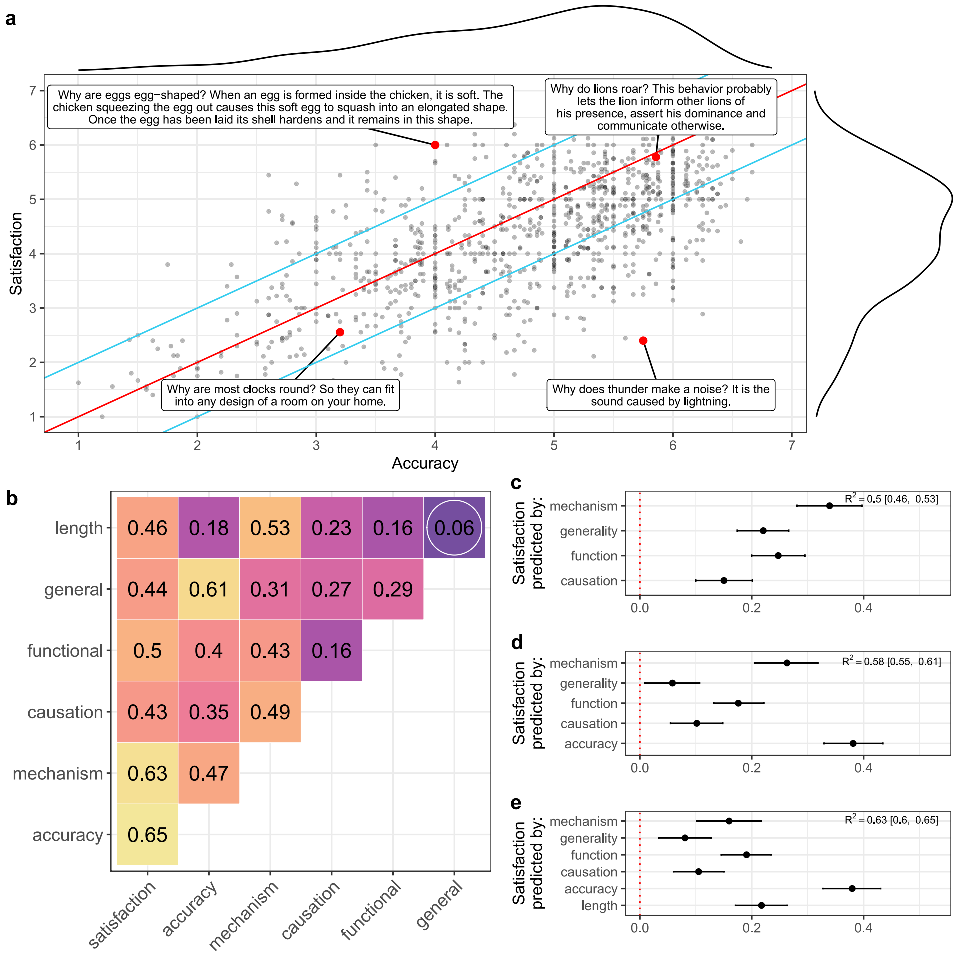

Fig. 1. Relationships between quality and content features

Note. (a) Scatterplot of accuracy and satisfaction, with reference lines shown at $y = x$ (red) and $y = x \pm 1$ (cyan) with distributions shown in the margins. Insets show four examples of provided explanations, where accuracy and satisfaction either align or diverge, and where satisfaction is either high or low. (b) Zero-order correlations between content features. All r's significant ($p < 0.001$) except when circled in white. (c–e) Standardized coefficients (βs with 95% CIs) from regressions predicting satisfaction from: (c) all content features, (d) content features plus accuracy, (e) content features, accuracy and explanation length. Model R^2s shown as insets. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

functions are thus not mutually exclusive: An explanation can contain elements of both. For instance, one response to the question 'Why do some people bully?' was 'Some people bully because they are having troubles of their own, or have low self-esteem, and picking on someone else makes them feel better.' The explanation (rated satisfaction: 5.18; accuracy: 5.64) mentions that low self-esteem can cause people to bully, and that a function of bullying is to make themselves feel better. Lombrozo and Gwynne (2014) found that causal and functional content were not in competition when it came to what generalizations participants drew from their explanations (their Experiment 1), or could even be positively associated, again in the context of generalization (their Experiment 2). In our explanations in the wild, it seems that the question of whether function and causation are positively associated or in competition can only be addressed in the context of other analytic decisions, such as whether to hold accuracy constant.

Apart from how functional and causal content are themselves related, how do they interact in predicting satisfaction? A study with researcher-generated explanations found that causal information was necessary for functional explanations to be considered acceptable (Lombrozo & Carey, 2006). Does this hold in the wild? If so, when satisfaction is regressed on both function and causation, there should be a low or null main effect of function, and there should be a positive interaction term. However, the main effect of function was positive ($\beta = 0.37\,[0.32, 0.42]$) and numerically larger than the main effect of causation ($\beta = 0.31\,[0.27, 0.36]$), whereas the interaction term was smaller and negative ($\beta = -0.14\,[-0.18, -0.10]$). Thus, an explanation rated highly for function does not need to have causal content for people to find it appealing.

We examine the relationship between causation and function in more depth in a follow-up study (Study 1a). Specifically, we probe whether causal and functional information are explicitly stated, might be implied, or are considered to be entirely absent from a given explanation; and whether these different response formats support our conclusions about satisfaction above.

If causation is an important aspect of explanation, it is surprising that ratings of casual content were not stronger predictors of satisfaction (Fig. 1c–e). A Bayesian mediation model shows that causation had an indirect effect on satisfaction via mechanism ($\beta = 0.2\,[0.17, 0.23]$) in addition to a direct effect ($\beta = 0.16\,[0.11, 0.22]$). There was moderate evidence (BF = 5.11) that the indirect effect is larger. Thus, it is not enough to merely name a cause — the explanation should also unpack how the cause brings about the effect.

2.2.4. Do explanations vary by domain?

Our main aim above was to understand what features of an explanation predicts ratings of satisfaction. However, explanation satisfaction may vary across domains (Hopkins et al., 2016; Weisberg et al., 2008), so we must also consider how domain might affect these relationships.

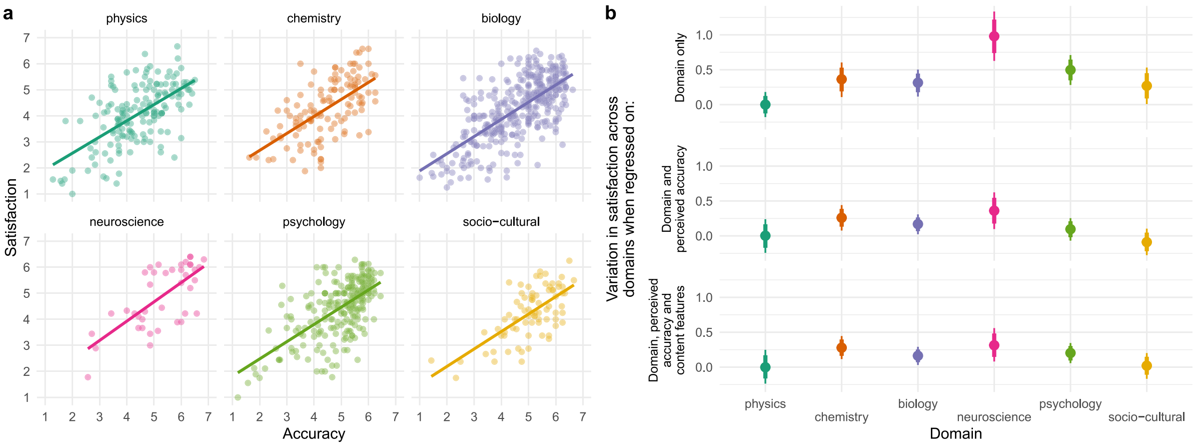

Fig. 2. Explanation quality across domains.
Note. (a) Scatterplot of accuracy and satisfaction, split by question domain, with linear fits. (b) Posterior predictions (with 83% and 95% CIs) for the effect of question domain on satisfaction, depending on whether satisfaction is regressed on domain only, on domain and accuracy, or on domain, accuracy and the four content features. The reference level – physics – is at 0 in each case.

The domain of the *explanandum* – the phenomenon to be explained – need not always match the domain of the explanation (Hopkins et al., 2016). For instance, a seemingly biological question can be answered by appealing to facts about chemistry. To ascertain the domain of the *explanandum* we presented the 50 ‘Why?’ questions to 14 research assistants, and asked them to tag each question with domain labels. Taggers could assign a question to multiple domains (e.g., the neuroscience question in Table 1 was also tagged as chemistry and biology, and the chemistry question was also tagged as physics). We assigned a question to a domain if it was tagged with that domain by at least 50% of the taggers.

For the domain of the *explanation*, we generated tags using language models (Python script available at https://osf.io/wbxcj/). Using published word embeddings (Grave et al., 2018), we computed explanation vectors by taking the mean of the word vectors for all the content words in each explanation (i.e., excluding auxiliaries, determiners, prepositions and similar grammatical words). We computed domain vectors by taking from the Wikipedia entry for each domain the words that were most specific to that domain, and then taking the mean of these domain-specific word vectors. Finally, we computed the cosine similarity between the explanation vectors and the domain vectors, assigning each explanation to the domain with the highest cosine similarity (Van Paridon & Thompson, 2020).

The relationship between perceived accuracy and satisfaction is consistent across domains (Fig. 2a), but how does satisfaction vary across domains? In answering this, we test two claims made in the literature concerning explanation domain: (1) that there is a ‘seductive allure’ of neuroscience explanations in that people find explanations especially compelling if they appeal to concepts from neuroscience (Weisberg et al., 2008), or (2) that there is a ‘reductive allure’ in that people find an explanation appealing if it reduces a phenomenon at one domain to principles at a lower-level domain Hopkins et al. (2016).

For manual tags of question domain, neuroscience had the highest satisfaction (Fig. 2b). There was strong evidence (BF=362) that neuroscience explanations were more satisfying than the second-highest domain, psychology. In contrast, for word-embedding derived tags of explanation domain, all CIs included 0 (or at least touched 0, in the case of the socio-cultural domain; for details see https://osf.io/wbxcj/). We note that future work using language models to tag explanation domains could improve on these null results, but for now there appears to be qualified support for the allure of neuroscience.

However, as we have shown that content features and perceived accuracy also predict satisfaction, does the qualified support for the seductive allure of neuroscience hold up, once these other variables are controlled for? Fig. 2b also shows the posterior predictions for the effect of question domain on satisfaction when perceived accuracy is included as a covariate, and when all four content features are included. Across these models, neuroscience remains numerically highest in satisfaction, yet the effect of domain is evidently contingent on perceived accuracy and the four content features. Not only does the relative ranking of the other domains change across models, but when content is included, the evidence that neuroscience is more satisfying than the now-second-highest domain, chemistry, is merely anecdotal ($BF_{10} = 1.61$). We thus caution against making claims about the effect of domain on satisfaction independently of explanation content.

To explore the ‘reductive allure’ claim, we created a new variable ‘reduction’ with value ‘true’ if the explanation domain was below that of the question domain – the domain of the *explanandum* – in Table 1 and ‘false’ otherwise. We excluded physics questions here, as there is no lower domain in our tagging system. There was no effect of reduction on satisfaction ($\beta = -0.11$ $[-0.24, 0.04]$, $BF_{01} = 4.27$).

Finally, to test whether the effects of perceived accuracy, mechanism, function, causation or generality on satisfaction vary across domains, we added by-domain random slopes to the model in Fig. 1d. None of these random slopes had CIs that excluded zero, regardless whether domain was represented as manual tags of question domain, or word-embedding derived tags for explanation domain (for full details, see https://osf.io/wbxcj/).

2.2.5. How well are people calibrated in their ratings of explanations?

Our second main aim was to understand how well explainers' ratings of their own explanations' quality was calibrated with those of other independent raters. To do so, we calculated for each explanation an ‘overestimation’ quantity, which is just the difference between the explainer's own rating and the average rating by others. We calculated overestimation separately for perceived accuracy and satisfaction. We also calculated ‘ability’, the average quality of each person's explanations (as rated by others), again separately for perceived accuracy and satisfaction, to see if calibration was related to ability.

We regressed overestimation on ability and variable type (perceived accuracy vs. satisfaction), including an interaction term. If people are well calibrated, overestimation for each variable type should be centered on zero, but if people tend to overestimate their ability, it will be positive. If the estimation of ability is independent of that ability, the slope for ability should be zero, but if ability estimation is worse for people with lower ability, then the slope will be negative.

Overall, people did not overestimate how accurately others would perceive their explanations ($b = 0.16$ $[-0.02, 0.34]$, $BF_{01} = 2.44$, though this counts as merely anecdotal evidence). However, they did overestimate how satisfying their explanations would be for other people ($b =$

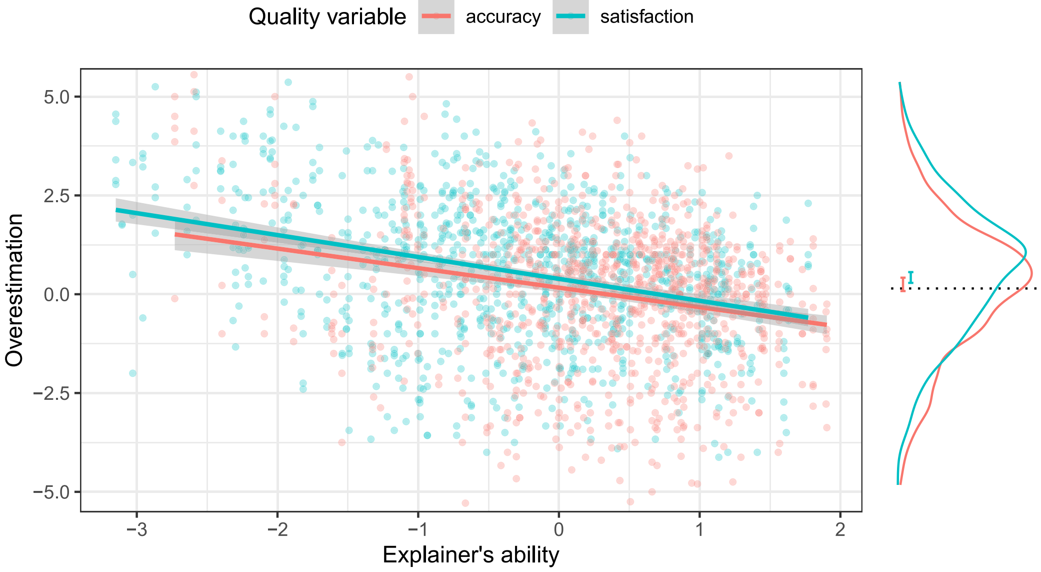

Fig. 3. Explainers' overestimation of the quality of their explanations.
Note. Overestimation (the difference between people's rating of how good their own explanations are, and the ratings of others) as predicted by explainer ability (the mean of others' ratings of quality per explainer), colored by quality variable. The marginal plot shows the distribution of overestimation, along with 95% CIs around the model estimate of the average of each quality variable. Satisfaction showed overestimation overall (its CIs in the marginal plot exclude zero, unlike accuracy) but lower ability was associated with greater overestimation for both accuracy and satisfaction.

0.25 [0.10, 0.39]). People with lower ability showed more overestimation of perceived accuracy ($b = -0.45$ $[-0.62, -0.30]$, Fig. 3), and the slope for satisfaction was not different from that of accuracy (interaction term $b = -0.05$ $[-0.20, 0.11]$, $BF_{01} = 10.32$).

SM2 examines whether the use of a Likert scale led to biased responding at the extremes of the scale (suggested by an anonymous reviewer). It shows that our conclusions are unchanged in two alternative analyses.

2.3. Discussion

We found that various kinds of content predicted explanation satisfaction, with mechanism and function emerging as the main drivers of satisfaction (though the effect of content varied depending whether accuracy and explanation length were included in the models).

Unlike previous studies with experimenter-generated explanations, we found that function predicted satisfaction independently of causation, and that function and causation had a weak positive or negative association (depending whether the common contribution of accuracy was accounted for). Explanations in the wild may thus operate differently to more formal explanations in the lab, depending what kind of information people spontaneously produce when generating explanations, or attend to (or infer) while rating explanations for content or quality. However, to check that this is not an artefact of our design, we pursue this question in more detail in the subsequent follow-up study.

We found that the relationship between content and satisfaction did not appear to differ across domains. While we note that a more sophisticated computational model of explanation domain may throw light on this in the future, we also caution against making claims about differences in satisfaction across domains without first accounting for explanation content.

Finally, we found that people overestimate the quality of their own explanations – at least when it comes to explanation satisfaction – but that people with lower average ability were worse calibrated on both dimensions of quality.

3. Study 1a (follow-up): Do laypeople view functional explanations as necessarily causal?

Function and causation had the weakest zero-order correlation among content features in Study 1, and were negatively associated once perceived accuracy was partialed out. This finding seems to conflict with philosophical claims that function logically involves causation (Wright, 1976). It also conflicts with empirical research testing those philosophical norms using experimenter-generated explanations: Lombrozo and Carey (2006) found that, for functional explanations to be accepted, the functions must play a causal role in bringing about the explanandum.

However, people are able to ascribe functions to objects without necessarily inferring that the functions relate to the objects' causal histories (Joo et al., 2023). Indeed, both laypeople (Kelemen & Rosset, 2009) and experts (Kelemen et al., 2013) may ascribe functions even when they are causally inappropriate. As Study 1 found that function predicted satisfaction independently of causation in explanations in the wild, it is at least psychologically plausible that people need not always construe functions causally.

Here we aim to probe our previous findings more robustly, in order to test (1a) whether a different way of asking about explanation content could reveal a stronger positive association between perceived function and causation (as predicted by aforementioned claims that functions are logically causal); or else (1b) whether laypeople spontaneously produce functional explanations that are not interpreted by other laypeople as containing causal information; (2) how consistent evaluations of causation and function are across response formats and study designs; and (3) whether it is necessary for a functional explanation to spell out a causal link to be rated as satisfying.

Whereas Study 1 involved Likert-scale responses rating confidence in the presence/absence of the relevant content, here we ask for a more categorical response (whether the relevant content is given, implied or absent — see Procedure for a description of response options). Further, whereas Study 1 used a between-subjects design (participants rated for function or causation but not both), this follow-up uses a within-subjects design (participants evaluated both causal and functional content, though question is still a between-subjects variable). It may be that prompting participants to look for causal information while they evaluate function could help them identify any causal connections between a function and the explanandum. Either way, if our findings are consistent across designs, this speaks to the informativeness of our data in Study 1 (cf. question 2 above).

3.1. Methods

3.1.1. Participants

We recruited participants from Amazon's Mechanical Turk (MTurk) platform. Participation was limited to those with an IP address in the USA and over a 95% approval rating on MTurk. We used CloudResearch/TurkPrime to manage participation as this platform tracks data quality in MTurk workers (Litman et al., 2017).

Participants rated the explanations produced in Study 1. As in Study 1, we aimed to have around 10 responses per explanation so we recruited 500 participants, 10 for each of the 50 'Why?' questions (though due to random assignment, some questions had slightly more or fewer than this target). Participants were paid $0.90–$1.00 to rate around 20 explanations from Study 1.

The study was approved by the University of Wisconsin–Madison Education and Social/Behavioral Science IRB.

3.1.2. Procedure

All materials are available at https://osf.io/wbxcj/.

Whereas ratings in Study 1 were on a Likert scale, for example ranging from 1 = 'Very confident it is not causal' to 7 = 'Very confident it is causal', here we sought a more categorical response. In particular, even if explanations did not contain causal or functional information, might participants take such information as implied? This could potentially reveal that functional explanations in Study 1 implied a causal link, even if they did not spell it out explicitly.

After participants provided informed consent, the nature of the task was explained to them, including the meaning of causation and function and the meaning of the response options (full instructions available at the above OSF link).

In the instructions, after causation and function were defined in lay terms, participants were told about the four response options: 'Yes - explicit', 'Yes — paraphrasable', 'Sort of — implied' and 'No — not even implied'.

Participants were told that an explanation counts as explicitly having causal information if it contains phrases such as 'cause', 'make', 'as a result' or words with similar meanings. It counts as having explicitly functional information if it contains phrases such as 'function', 'for', 'so that', 'in order to' or words with similar meanings. Example explanations containing such terms were provided.

They were told that if the explanation does not obviously contain keywords such as the above but could be paraphrased in a way that made these explicit, it should count as 'Yes — paraphrase'. The instructions provided examples of such paraphrasing.

If the explanation seemed to hint as something causal or functional, but it was not obvious how to paraphrase it to make that clearer, they could choose 'Sort of — implied'.

Finally, if none of the above applied, then they should choose the final option, 'No — not even implied'.

Participants were assigned one of the 50 'Why?' questions from Study 1, and were shown the approximately 20 explanations previously provided to answer that question. On the survey page, participants were reminded of the meanings of causation, function and the response options. They then rated the provided explanations for causation and for function, with one set of response options per content feature. The order of causation/function was randomized per participant, as was the order of explanations.

In addition, participants were given an 'opt-out' check box for each explanation which they could select to indicate that the given text was not an answer to the question. This was intended to filter out uninformative responses such as "I don't know" or nonsense responses in the Study 1 dataset of explanations. Selecting this option disabled the causation/function responses. Four of the 1013 explanations were excluded from further analysis on the basis that this check box was selected by a majority of raters. For the sake of consistency, we have retroactively applied this minor exclusion to Study 1, and can confirm that it makes no difference to our conclusions.

As a data-quality check, we deliberately added two non-answers to the survey page. One was "Don't know, sorry" and the other was a random string of numbers and letters. If a participant missed these (failed to select the opt-out on *both* occasions), we counted them as inattentive and excluded them from further analysis. 41 participants (8%) were excluded on this basis.

3.2. Results

All data and full analysis scripts (including full model specifications) are available at https://osf.io/wbxcj/. For the regressions, we specified weakly informative priors ($\beta \sim \mathcal{N}(0,1)$). In all plots of regression predictions, dots indicate medians of the expected value of the posterior predictive distribution and error bars indicate 95% CIs.

As these are categorical responses (unlike in Study 1), we do not average them. Where we enter this raw data into regressions, we can thus include a random intercept for rater in addition to a random intercept for question. However, for some analyses below we wish to know what the modal response category was for each explanation (the option chosen by a plurality of raters; in case of a tie, all winners were counted). Unlike models of the raw responses, we could not include a random intercept for rater as modal responses involve aggregating per explanation.

Fig. 4 illustrates the responses to one question, 'Why do people hiccup?' (equivalent plots for all 50 questions can be found at https://osf.io/wbxcj/). Some explanations were rated as both causal and functional (e.g., explanation 19: 'We hiccup because a bubble of air gets trapped in our throats. This sets off a reflex that cause our throats to spasm to dislodge the bubble.'). Some were rated as functional but not causal (e.g., explanation 1: 'To get the gas out.'). Others were rated as causal but not functional (e.g., explanation 2: 'It's air bubbles caught in the lungs.' Participants recognized that this did not explicitly label the bubbles as a cause, but most realized that it still counts as having causal information because it can be paraphrased to make this explicit, for instance as 'Air bubbles cause hiccups.'). Some were rated as neither (explanation 12, which did not meet the threshold for being excluded as not answering the question). Explanation 17 ('Hiccups are a muscle spasm within the diaphragm.') illustrates how, given the inevitability of noise in such data, it is worth also considering modal responses: Despite some idiosyncratic individual responses, the modal response was that it is paraphrasably causal but not functional.

3.2.1. Were functional explanations necessarily interpreted as containing or implying causal information?

Fig. 5a, b shows the counts of how often each combination of response categories appeared in the data, both for raw responses (Fig. 5a) and modal responses (Fig. 5b). Both for raw and modal responses, the commonest explanations evoked by our 50 questions were rated explicitly causal and not functional.

However, when an explanation was rated as explicitly functional, it was mostly either not causal or explicitly causal. Thus, it seems that functional explanations in the wild need not convey causal information. Spearman rank correlations show a negative association between function and causation (raw responses: $r_s = -0.213$, $p < .001$; modal responses: $r_s = -0.18$, $p < .001$).

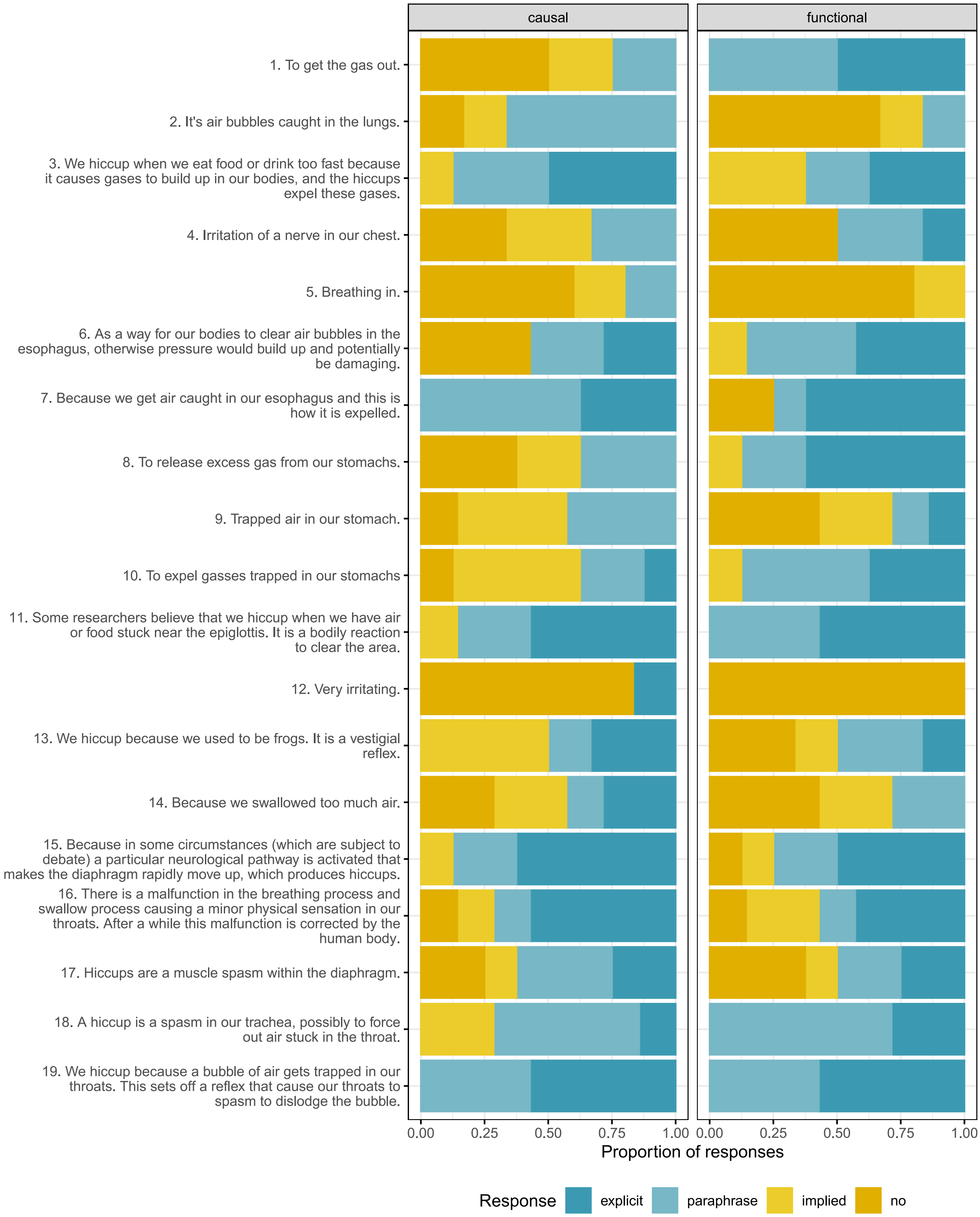

Fig. 4. Categorical responses across explanations answering the question ‘Why do people hiccup?’.

We regressed causal responses on functional responses with a Bayesian ordinal mixed-effects model (Bürkner & Vuorre, 2019). Both outcome and predictor were modeled as ordered categories ranging from ‘no’ to ‘explicit’. Thus, references to ‘less’ or ‘lower’ in the following mean “towards the ‘no’ end of the scale” while references to ‘more’ or ‘higher’ mean “towards the ‘explicit’ end of the scale’.

Given an ordered predictor with four levels, the regression estimates linear, quadratic and cubic coefficients. This allows, for instance, that the distance between ‘paraphrase’ vs ‘explicit’ levels of function can be closer than the distance between ‘no’ vs ‘implied’. We specified the model family as cumulative. A cumulative model assumes that the ordered outcome categories reflect an underlying ‘causalness’ variable, and it estimates intercepts representing thresholds in that variable, corresponding to transitions between adjacent overt category levels (Bürkner & Vuorre, 2019).

A parsimonious form of this model estimates intercepts/thresholds for the data overall. This parsimonious model yielded a negative linear coefficient for function ($b = -0.621$ $[-0.717, -0.526]$) and a positive quadratic coefficient ($b = 0.265$ $[0.174, 0.357]$; there was no cubic effect: $b = -0.037$ $[-0.122, 0.046]$). Thus, the more functional an explanation was, the lower its probable causal category, but this negative effect tapered off for higher levels of function. Fig. 5c illustrates these effects by plotting the model-estimated probability that a given level of causation will co-occur with each level of function. For higher levels of function, it is more likely that the explanation will rated ‘no’ or ‘implied’ for causation, and less likely that it will be rated ‘explicit’ or ‘paraphrase’.

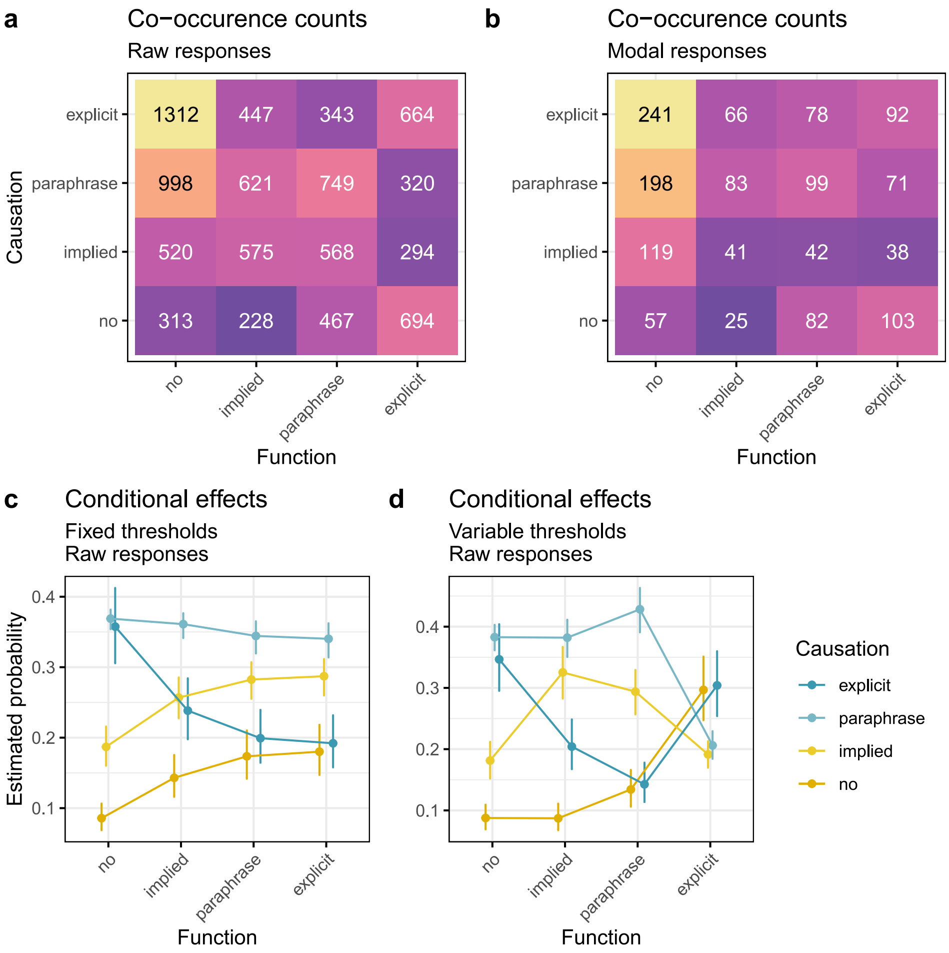

Fig. 5. Counts and ordered-regression estimates of the co-occurrence of categorical causation and function responses.
Note. (a) Counts of how often each combination of causation and function responses occurred in the data. (b) Counts of how often each combination of causation and function response was the modal response to an explanation. (c) Ordered regression conditional effects for the raw responses, assuming one set of thresholds across levels of function; (d) Ordered regression conditional effects for the raw responses, allowing variable thresholds across different levels of function. For equivalent models for the modal category responses, see https://osf.io/wbxcj/.

However, because it estimated one set of thresholds, this model was not sensitive to differences in the distribution of causation responses across levels of function (whereby, for instance, counts in Fig. 5a increased with levels of causation for function='no', but were concentrated at the extremes for function='explicit'). Informally, the model smoothed out differences across columns in Fig. 5a. But these are precisely what we want to track. The model has thus radically underestimated how likely it is that an 'explicit' response for function co-occurred with a 'no' response for causation, compared to the counts in Fig. 5a.

A less parsimonious version of the model can estimate different intercepts/thresholds for each level of function. Fig. 5d illustrates the estimated probabilities for this model, showing a more dramatic increase in the likelihood of a 'no' response for causation for higher levels of function. However, in using function responses to model greater complexity in the intercepts, the model is less able to detect overall fixed effects (linear: $b = -0.221$ $[-1.867, 1.39]$, quadratic: $b = 0.048$ $[-1.608, 1.683]$, cubic: $b = 0.01$ $[-1.58, 1.61]$).

Either way, an explicitly functional explanation did not necessarily contain causal information, and this can be seen in an overall negative association in the more parsimonious model, or in the dramatic increase in 'no' causation responses for higher levels of function in the more complex model.

3.2.2. How did these categorical responses compare with the ratings in study 1?

Despite differences in response format (Likert rating of confidence vs categorical response) and study design (between-subject vs within-subject), if the responses are tracking causal or functional content in explanations, we would expect a meaningful relationship between new responses here and the previous responses in Study 1. Even if not linear, the relation should at least be monotonic: If an explanation was more likely to be rated 'explicit' than 'paraphrase' here, then we should see a higher confidence rating in Study 1. If the responses are tracking different information, however, the relationship might not even be monotonic.

For an overview of the different datasets, Fig. 6a, b combine density plots with Loess smooths. Each panel focuses on one category of the function or causation responses in the current study. The x-axis indicates what proportion of the responses to a given explanation represented that category. The y-axis reflects the average confidence rating for that explanation in Study 1. The higher the proportion of 'explicit' or 'paraphrase' responses here, the more confident participants in Study 1 were that an explanation contained the relevant type of information. The higher the proportion of 'implied' or 'no' responses here, the less confident participants in Study 1 were. It is thus reasonable to think of 'explicit' and 'paraphrase' as broadly positive responses (the explanation text contains the relevant information) and 'implied'

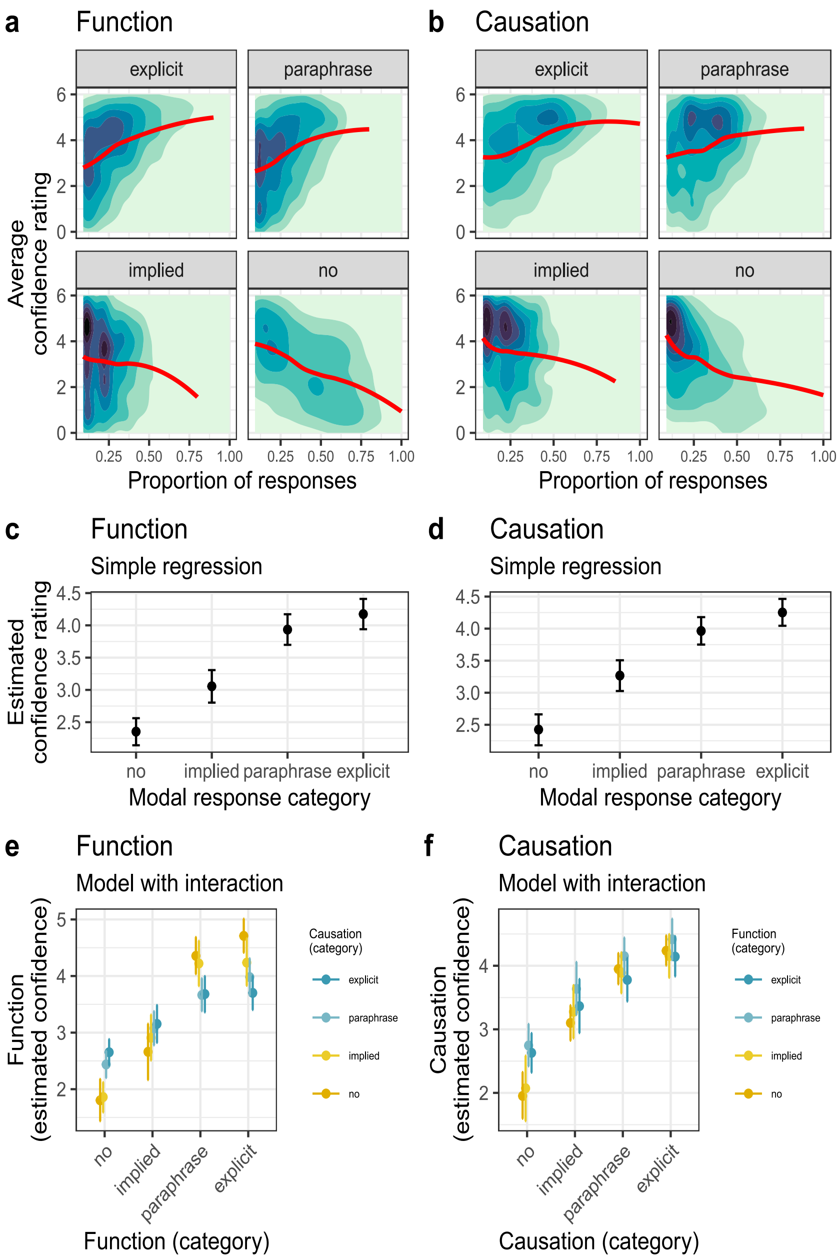

Fig. 6. Distributions and regression predictions comparing responses in the current study with confidence ratings from Study 1.
Note. (a, b) Density plots with Loess smooths relating the categorical responses here with confidence ratings from Study 1. (c, d) Regression predictions of confidence ratings for each level of categorical response within each content type. (e, f) Function and causation confidence ratings predicted by the modal category responses, including interaction terms between content types.

and ‘no’ as broadly negative responses (the explanation text does not actually contain the information, though in the former case, it could be inferred). To test these associations, we regressed Study 1’s function ratings on the present study’s modal function categories (Fig. 6c) and regressed Study 1’s causation ratings on the modal causation categories (Fig. 6d). Higher levels of the function category were associated with greater confidence that an explanation contained functional information (linear: $b = 1.419$ $[1.282, 1.557]$) though this association tailed off at higher levels (quadratic: $b = -0.231$ $[-0.377, -0.083]$; cubic: $b = -0.181$ $[-0.334, -0.027]$). Thus, as illustrated in Fig. 6c, high confidence ratings in Study 1 could reflect a paraphrasable response almost as much as an explicit response. The same pattern is observable for causation ratings (linear: $b = 1.382$ $[1.233, 1.533]$; quadratic: $b = -0.278$ $[-0.422, -0.133]$; though no cubic: $b = -0.06$ $[-0.204, 0.081]$).

To test for an interaction between causal and functional categories, we used both modal category responses (including interaction terms) to predict causal and functional confidence ratings from Study 1. The model replicates the linear and quadratic main effects reported above, but given the number of interactions between linear, quadratic and cubic effects (and the opacity of their interpretation), we turn to plots of the model conditional effects for further illumination (Fig. 6e, f). For coefficients, see https://osf.io/wbxcj/.

The most striking difference between the outcome variables is that higher functional confidence ratings in Study 1 are associated with lower causation categories. Specifically, in Fig. 6e, for explanations categorized as paraphrasably/explicitly functional, explanations without causal content (no/implied) were associated with *higher* function confidence ratings. The reverse is not visible for causal confidence ratings in Fig. 6f, where varying functional category levels do not seem to make much difference for explanations charaterized as paraphrasably/explicitly causal (apart from the ‘no’ response on the *x*-axes).

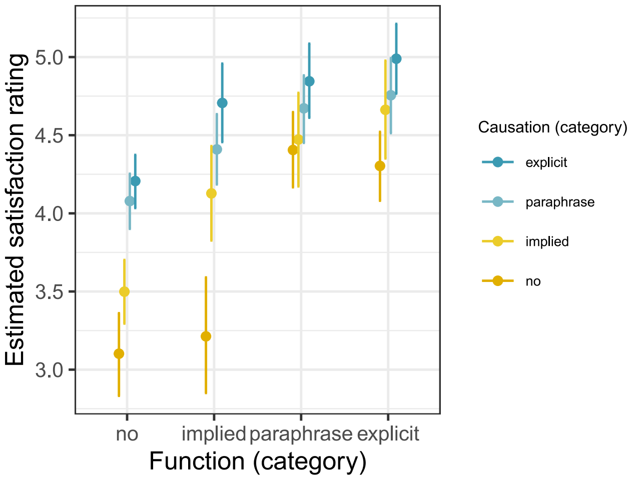

Fig. 7. Satisfaction ratings as predicted by modal causation and function categories, including an interaction term.

In sum, this follow-up study confirms that, when an explanation received high functional and low causal ratings in Study 1, this was likely because raters did not think it expressed causal content.

3.2.3. Is it necessary for functional explanations to convey a causal link to be rated as satisftying?

Study 1 found that function predicted satisfaction independently of causation, but perhaps that was driven by functional explanations that *implied* something causal. We regressed satisfaction ratings from Study 1 on the categorical responses from the current study (including an interaction term), to test whether the previously observed effect of function is really due to the expression of functional content.

Satisfaction was predicted both by function (linear: $b = 0.75$ $[0.643, 0.859]$) and causation (linear: $b = 0.689$ $[0.571, 0.807]$), and in both cases a quadratic term meant that these effect tailed off for high levels of each predictor (function quadratic: $b = -0.157$ $[-0.273, -0.043]$; causation quadratic: $b = -0.113$ $[-0.226, -0.002]$).

There was also a negative interaction between the linear effect of function and the linear effect of causation ($b = -0.423$ $[-0.621, -0.222]$). Fig. 7 illustrates the conditional effects. Higher levels of causation category are always associated with greater satisfaction. However, there is a dramatic jump in satisfaction ratings when 'causation = no' between 'function=implied' and 'function = paraphrase'. While having no causal information leads to very low satisfaction for lower levels of function (no/implied), it does not do so for higher levels of function (paraphrase/explicit).

In sum, even though spelling out – or even implying – causal information is a good thing in terms of explanation satisfaction, a functional explanation can be satisfying even without causal content.

4. Discussion

Using categorical responses (indicating whether an explanation explicitly mentioned causal or functional information, could be paraphrased to do so, only seemed to imply it, or did not contain it), this follow-up study has provided additional support for three main claims.

First, participants were able to categorize an explanation as functional without needing to see or infer a causal link. The normative philosophical claim that function logically involves a causal link does not seem to translate into psychological necessity in terms of what content is expressed in explanations in the wild.

Second, despite a number of differences between the studies, the responses here are monotonically related with responses in the previous study. Not only does this support the claim that responses are tracking lay perceptions of functional and/or causal content, it also means that conclusions drawn from ratings in Study 1 do not just reflect the design of that study.

Third, functional explanations can lack causal content and still be rated as satisfying. This supports our general claim that function and causation contribute independently to explanation satisfaction, though the model revealed some complex interactions worth exploring in future research.

5. Study 2: Individual differences in cognitive traits

As progress on the cognitive science of explanation requires a better understanding of relevant cognitive mechanisms, our final research question was: Which cognitive traits are associated with explainers' ability to produce high quality explanations? We pre-registered four hypotheses (https://osf.io/qw8ut). The first of these is a self-replication of an aspect of Study 1 concerning explanation quality. Cognitive drivers of explanation quality are then examined in the remaining hypotheses.

H1 Satisfaction and perceived accuracy will correlate positively (replicating Study 1).

H2 Explanation satisfaction will correlate positively with measures of cognitive ability or cognitive style (i.e., with higher verbal intelligence, insight ability, perspective-taking ability, reflective cognitive style, epistemic curiosity, and science literacy)

H3 These measures of cognitive ability/style will positively predict unique variance in explanation satisfaction.

H4 These measures of cognitive ability/style will still predict unique variance in satisfaction, controlling for perceived accuracy.

5.1. Methods

5.1.1. Participants

As in Study 1, we recruited participants from MTurk using the same inclusion criteria, in two phases.

For Phase 1, we pre-registered a sample size of 200 (for details, based on correlations from a pilot study, see https://osf.io/qw8ut), and we pre-registered that we would re-recruit participants from a previous unrelated study, as this included several individual-differences measures that we require here. However, we were only able to re-recruit 187 of those participants. They were paid $4.00 to generate 10 explanations (yielding 1870 explanations) and to respond to various individual-differences measures of cognitive ability and cognitive style.

In Phase 2, 1879 participants were paid $0.40 to rate the Phase 1 explanations for satisfaction or perceived accuracy. As for Study 1 Phase 2, we aimed to have 10 ratings per explanation.

The study was approved by the University of Wisconsin–Madison Education and Social/Behavioral Science IRB.

5.1.2. Materials

All materials, including rating instructions, are available at https://osf.io/wbxcj/. In Phase 1, in addition to generating 10 explanations, participants responded to the following scales:

Insight ability: 20 Compound Remote Associate (CRA) problems (sampled from Bowden & Jung-Beeman, 2003). Each problem consists of three cue words (e.g., 'cane', 'daddy' and 'plum'). The aim is to think of a fourth word that can be combined with all three to produce common words or phrases (here, 'sugar', yielding 'sugar cane', 'sugar daddy' and 'sugar plum'). These problems index participants' ability to creatively make connections between sometimes distantly associated concepts.

Science literacy: 12 multiple-choice items asking about general science knowledge, such as a true or false question about whether the center of the earth is hot (National Science Board, 2018; Shtulman & Valcarcel, 2012).

Perspective-taking ability: 20 items from a communication game (Sulik & Lupyan, 2018). In each trial, participants are given a target word (e.g., 'bank') and their aim is to generate a single word as a signal that would help someone else guess the target, based on the signal alone. For instance, if they generate the signal 'teller', it turns out that people are very likely to guess 'bank' correctly, but if they generate the signal 'money', few people are likely to guess 'bank' on the basis of this signal alone. The challenge is to think of a signal that is informative from the audience's point of view. See https://osf.io/wbxcj/ for details of scoring, and of the distinction between test and distractor items.

Epistemic curiosity: 10 items, such as 'I enjoy learning about subjects which are unfamiliar' (Litman & Spielberger, 2003). Participants rate their agreement (on a 4-point Likert scale) with each item.

In addition, as we re-recruited Phase 1 participants from a previous study, we already had data for the following individual-differences measures.

Vocabulary: 14 multiple-choice vocabulary test items. Cor et al. (*Wordsumplus*, 2012). Participants are given a word and need to pick from a list of options the meaning that best matches it. Vocabulary knowledge is an aspect of crystalized verbal intelligence (Malhotra et al., 2007).

Cognitive reflection: We combined 3 Cognitive Reflection Test (CRT) items from Shenhav et al. (2012) and 4 items from Thomson and Oppenheimer (2016). Each involves a question that has an intuitive but wrong answer, such as 'A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost?' (Frederick, 2005, though this classic example was not among the aforementioned 7 items). A commonly given but wrong answer is 10 cents. The correct answer is 5 cents. This scale thus indexes participants' ability to reflect on a problem sufficiently to go beyond the obvious solution.

Verbal reasoning: 4 items testing deductive reasoning (Condon & Revelle, 2014). These items were included, first, because they lack obvious yet misleading answers and thus serve as distractors for the above CRT items (and were presented along with them); second, because they measure verbal reasoning and thus index another aspect of verbal intelligence (in addition to vocabulary); and third, because deductive reasoning is central to explanation according to early accounts from the philosophy of science (Hempel, 1965).

5.1.3. Procedure

In Phase 1, after providing informed consent, participants undertook a simple English test (10 sentences that they had to identify as grammatical or not) to ensure they could comprehend the instructions and were able to provide explanations. They then provided explanations to 10 'Why?' questions drawn from Study 1 (with similar instructions) and undertook the individual-differences measures in the order listed above.

Phase 1 included three attention checks, and we excluded 24 participants from analysis on the basis that they either failed two or more attention checks, or got less than 70% correct on the English test.

In Phase 2, after providing informed consent, participants were randomly assigned 20 explanations from a single question, and were asked to rate each according to a single criterion (either perceived accuracy or satisfaction, with similar instructions to Study 1).

Each explanation was also accompanied by a check box labeled 'This is not even an answer,' which participants could click to filter out spurious or joke answers. We dropped 40 explanations from analysis that had been judged as not an explanation by at least five raters. Further, one of the given 'explanations' was in fact an attention check, asking participants to click on a specific response if they were reading carefully. If a participant failed to click the indicated response, we did not include their ratings in calculating the average satisfaction or accuracy score. 395 raters (21%) were dropped on this basis.

5.2. Results

All data and full analysis scripts (including full model specifications and control variables age, gender and education) are available at https://osf.io/wbxcj/. For the regressions, we specified weakly informative priors ($\beta \sim \mathcal{N}(0,1)$). Regressions included a random intercept for explainer and for question. As the ratings were averaged per feature per explanation, the data did not include a row for each rating, and we thus do not include a random intercept for each rater.

5.2.1. Pre-registered analyses

We depart from the pre-registered analysis by using Bayesian instead of frequentist regressions.

Fig. 8a displays the zero-order correlations between both indices of explanation quality (perceived accuracy and satisfaction) and our individual-difference measures: epistemic curiosity, vocabulary, perspective taking ability, general science literacy, insight problem solving ability, verbal reasoning, and cognitive reflection. As in Study 1, perceived accuracy correlated strongly with satisfaction (H1). All of the individual-differences measures correlated significantly (all $p < .005$) and positively with satisfaction (H2), except for epistemic curiosity ($r = .08, p = .276$).

Epistemic curiosity was not significantly related to any other measure, though its numerically strongest correlation was with general science knowledge ($r = .149, p = .055$). Otherwise, there were significant small-to-moderate correlations between the other variables. Of the individual-differences measures, epistemic curiosity was the only true self-report measure. It is therefore possible that the small size of epistemic curiosity's correlations merely reflects participants' inability to reflect accurately on their own epistemic curiosity, rather than a true lack of a relationship between having greater epistemic curiosity and generating more satisfying explanations.

Fig. 8b shows standardized coefficients from a Bayesian multiple linear regression, predicting satisfaction from the various individual-differences measures. This and the following models include demographic control variables: age, gender, and education, though these had no effect in any of the models reported here (for their coefficients, see SM3; for further details, see the full analysis at https://osf.io/wbxcj/).

Three individual-differences measures predicted unique variance in satisfaction in this multiple regression (H3): science knowledge ($\beta = 0.11$ $[0.02, 0.20]$); perspective taking ($\beta = 0.10$ $[0.01, 0.18]$) and insight ($\beta = 0.15$ $[0.07, 0.24]$). The others did not: curiosity ($\beta = 0.01$ $[-0.07, 0.09]$ $BF_{01} = 25$), verbal reasoning ($\beta = 0$ $[-0.08, 0.09]$ $BF_{01} = 22.1$), and cognitive reflection ($\beta = 0.08$ $[-0.01, 0.16]$ $BF_{01} = 4.8$).

Fig. 8c shows the coefficients with perceived accuracy included as a co-variate (H4). Now, only insight ($\beta = 0.12$ $[0.05, 0.20]$) and science knowledge ($\beta = 0.08$ $[0, 0.16]$) had positive effects. The CIs for perspective taking just overlapped 0 ($\beta = 0.06$ $[-0.01, 0.13]$, with an estimated posterior probability of .95 for a positive effect). For the others, there was moderate evidence that they had no effect (all $BF_{01} > 6.5$; see https://osf.io/wbxcj/ for details).

5.2.2. Exploratory analyses

As the value of searching more carefully through one's knowledge may depend on the extent of one's knowledge, one immediate question is whether there might be an interaction between either of the measures of knowledge that we included (vocabulary and science literacy) and the measures of people's tendency to search or reflect on that knowledge (curiosity and cognitive reflection). We added four interaction terms (one for each pairwise combination of knowledge and search variables) to the model in Fig. 8c. None of these interaction terms had an effect (all $BF_{01} > 10$; see https://osf.io/wbxcj/ for details).

As some cognitive measures correlated more strongly with perceived accuracy than with satisfaction, and as several effects on satisfaction dropped out when accuracy was included in the model, we

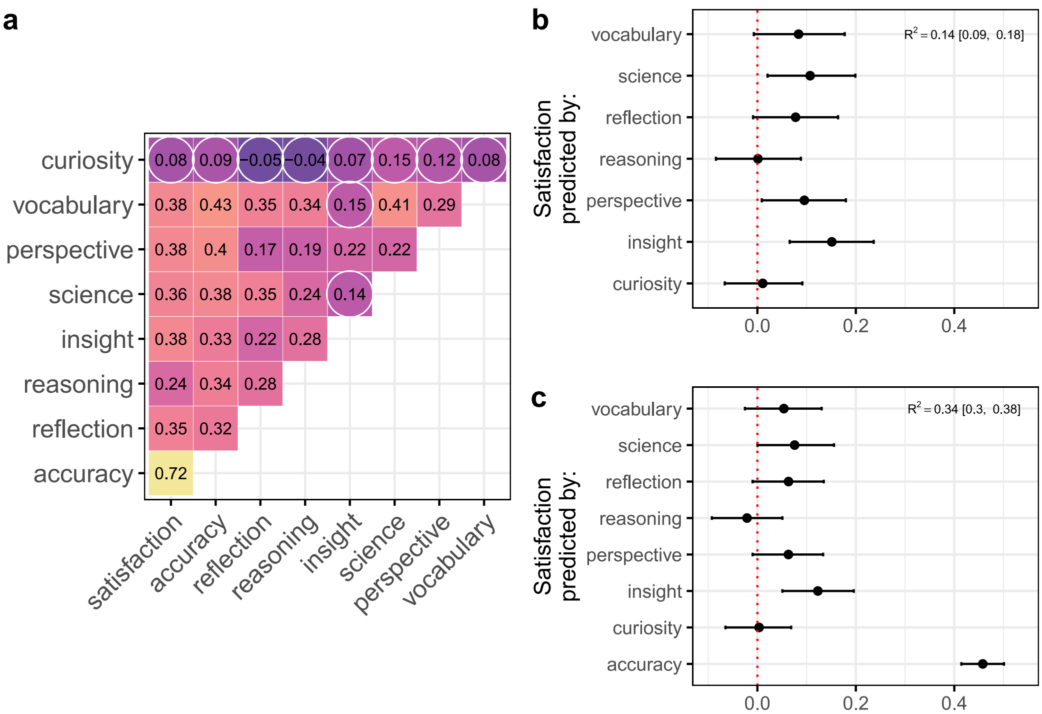

Fig. 8. Explanation quality and measures of cognitive ability.
Note. (a) Zero-order correlations between explanation quality variables and individual-differences measures. All r's significant ($p < .05$) except those circled in white. (b) Standardized coefficients (βs, with 95% CIs) with satisfaction regressed on all cognitive measures; (c) Accuracy added as a co-variate to the model in (b). Model R^2s (for fixed effects) are shown as insets. For covariates age, education and gender (all not predictive), see SM3.

investigated the extent to which each variable directly predicts satisfaction vs. predicting it as mediated via perceived accuracy. The following ignores predictors with coefficients near 0 in Fig. 8b (reasoning and curiosity).

We conducted a path analysis using Bayesian simultaneous regressions, with direct pathways from the predictors to satisfaction, as well as indirect pathways via perceived accuracy. Fig. 9a shows the model structure, along with standardized coefficients for all paths (with 95% CIs). Fig. 9b illustrates the total effect of each cognitive variable on satisfaction, shaded to reflect what proportion of the total effect is direct vs. indirect via perceived accuracy. There was strong evidence for all total effects (all evidence ratios > 23.5) except the demographic covariates (age, gender, education — see SM3).

Considering the total effects, the ability to provide a satisfying explanation depended unsurprisingly on having the relevant knowledge (science literacy) and verbal intelligence (as measured by vocabulary). The direct effect for science literacy was larger than its indirect effect ($BF_{10} = 5.86$) but the direct and indirect effects of vocabulary were the same size ($BF_{01} = 28.7$). The ability to produce satisfying explanations was also predicted by the disposition to engage one's cognitive ability to look beyond the obvious or intuitive contributions of that knowledge (cognitive reflection).

Most interesting, in our view, are the results for insight problem solving and perspective taking ability. The latter implies that a good explanation is a communicative act, benefiting from the ability to take others' perspective. The largest total effect was for insight problem solving. Indeed, the direct effect of insight on satisfaction was larger than the other variables' total effects. In short, of the abilities we measured, the one most critical for explanation satisfaction was the capacity to make insightful connections, retrieving and putting together distantly related information in one's knowledge, to form a non-obvious representation of the problem.

5.3. Discussion

As one way to advance the psychology of explanations, we tested which cognitive traits were associated with the rated quality of generated explanations, to better understand which cognitive mechanisms might be involved in producing satisfying explanations.

We found that insight problem solving was the cognitive trait most strongly associated with explanation satisfaction. Insight – commonly experienced as an 'Aha!' moment – is about discovering a relevant representation of a problem. It involves finding or making connections between pieces of world knowledge, especially when these connections are non-obvious, are novel or are otherwise not made salient by provided cues or contextual constraint.

Unsurprisingly, scientific general knowledge was another contributing factor, but we also found that perspective-taking ability directly predicted explanation satisfaction. In as far as explanations are answers to 'Why?' questions, one may be better able to generate an explanation if one is better able to see others' perspectives on the question.

6. General discussion

What kinds of explanations do people judge as being good? We solicited thousands of explanations from laypeople in response to 'Why?' questions. Analyzing these explanations has allowed us to identify several predictors of quality.

Holding perceived accuracy constant, causation, function, and mechanism all predicted unique variance in rated satisfaction. Of these, functional and mechanistic information provided the greatest boost to satisfaction. Although having a function logically implies some causal connection, functional explanations in the wild did not consistently express causal information, and they did not have to do so to be regarded as satisfying. This coheres with the observation that 'functional explanations . . . don't wear their causal commitments on their sleeves' (Lombrozo & Wilkenfeld, 2019).

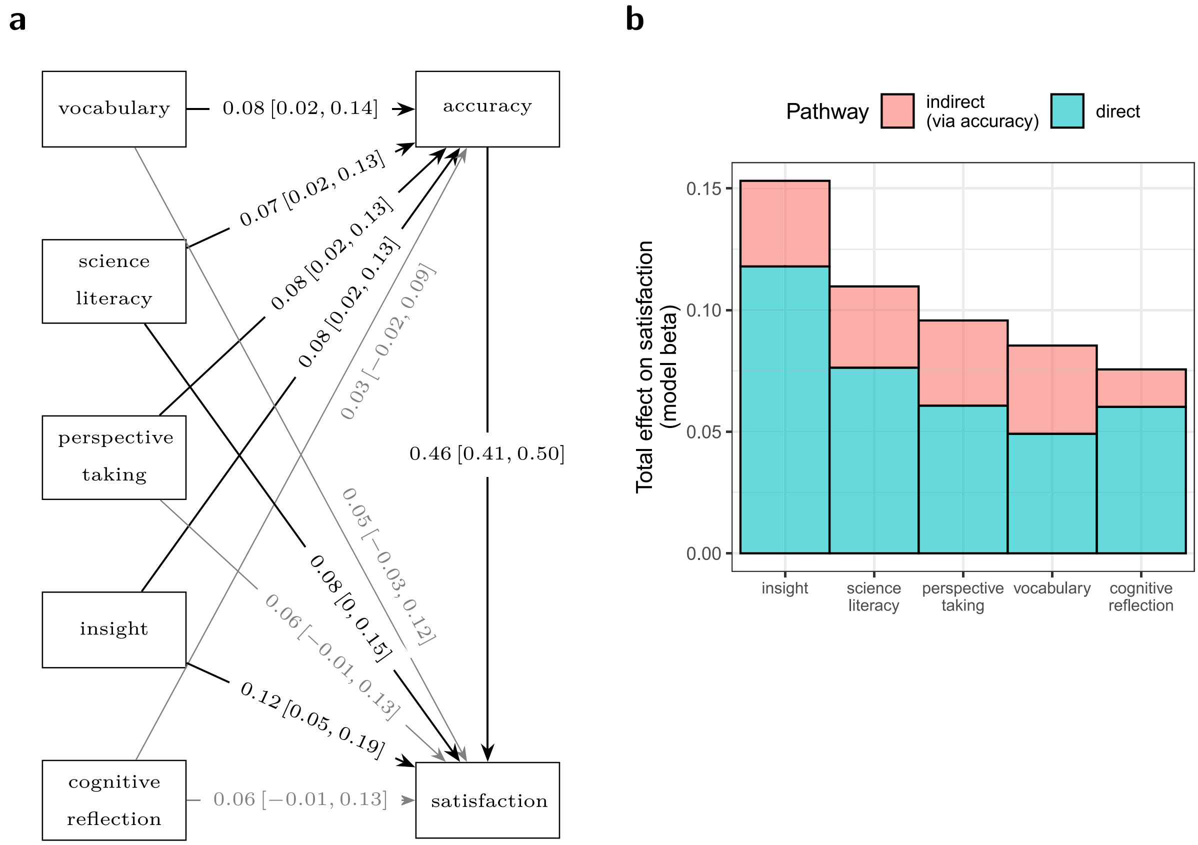

Fig. 9. Results of the Bayesian path analysis.
Note. (a) Pathways in Bayesian simultaneous regressions. Pathway labels are standardized coefficients and 95% CIs. Paths are gray if their CIs include 0. (b) Regression coefficients from the same analysis showing how the total effect in each case is divided between the direct effect of each cognitive variable on satisfaction and the indirect effect as mediated via accuracy. For covariates age, education and gender (all not predictive), see SM3.

How well are people calibrated in their rating of explanations? Explainers generally overestimated how satisfying their explanations were, though on average they did not overestimate the perceived accuracy of their explanations. Nonetheless (as Fig. S1.1b in SM1 illustrates), agreement among raters regarding satisfaction was highest out of all the rated variables. For both accuracy and satisfaction, those participants who generated worse explanations also tended to overestimate the quality of their explanations more strongly. Which cognitive abilities help people produce good explanations? The most important abilities for producing satisfying explanations were insight problem solving, science knowledge, and perspective taking. A good explanation goes beyond just including correct facts: It is also about *leveraging* the relevant knowledge, connecting the dots and doing so in a way that is useful from the audience's perspective.

6.0.1. Implications for the cognitive science of explanation

Our results support a pluralistic view of explanation (Colombo, 2017), with mechanism (*how* something occurs) and function (something's *purpose*) being dominant features in predicting how satisfied people are with a given explanation. These results are consistent with findings that, although 'Why?' questions are often semantically ambiguous, people can pragmatically infer whether a given 'why?' is really more a matter of 'how?' or 'for what purpose?' (Joo et al., 2021).

In contrast to research that offers a normative account of explanation (how explanations *ought* to work, for instance given an epistemic goal such as accurate prediction), our aims are directed at a more descriptive account of explanation (how explanations work, given their role in daily mental life). In comparison to philosophical accounts of explanation, cognitive accounts of explanations – and especially cognitive theories that are not motivated by concerns directly inherited from the philosophy of science – are relatively new. Complementing previous cognitive work (Bechlivanidis et al., 2017; Gopnik, 2000; Hopkins et al., 2016; Horne et al., 2019; Keil, 2006; Kelemen & Rosset, 2009; Liquin & Lombrozo, 2022; Lombrozo, 2006; Zemla et al., 2017, among others), our explanations-in-the-wild approach highlights several desiderata of a cognitive theory of explanations.

The first desideratum is considering lay explanation to be a communicative, interactive phenomenon (Faye, 2007; Hilton & Erb, 1996; Keil, 2006), rather than only as a process subserving internal theory formation. In Study 1, we showed that it was more challenging to judge how satisfying one's explanations were than to judge their perceived accuracy. Thus, the generation of truly satisfying explanations needs to take others' perspectives into account. In Study 2, we showed that perspective-taking ability predicted satisfaction. Both findings suggest that a cognitive theory of explanation must account for how people generate and evaluate explanations as they interact with one another.

The second desideratum is accounting for what pieces of information are relevant when generating an explanation (Hilton & Erb, 1996). Our explanations-in-the-wild approach focuses on how explainers spontaneously generate relevant information given their background knowledge, whereas traditional laboratory approaches (e.g., Hilton & Erb, 1996) have tended to present participants with contrastive cases where relevance is evaluated rather than generated. Further, insight, the strongest predictor of satisfaction in Study 2, involves finding a relevant representation of a problem (Durso et al., 1994; Gilhooly & Murphy, 2005). Generating an explanation is an ill-defined problem (Horne et al., 2019) in that working out what information is relevant to the problem is part of the problem. A cognitive theory of explanation must account for how the generation of explanations involves inferences about relevance.

6.0.2. Where to go from here?

What, concretely, does a feeling of satisfaction indicate? Satisfaction has been studied as a predictor of further inquiry, including learning or generalization (Gopnik, 2000; Liquin & Lombrozo, 2022), and here we have examined content features that predict satisfaction, but what does the experience of satisfaction mean in concrete terms and in the moment it is felt? Does it signal that one has had enough information? Enough information to stop wondering? Enough information of a certain kind that can be integrated or connected in specific ways to yield insight? One promising direction, which may tie together a number of these possibilities, is to explore whether (and if so, how) satisfaction relates to the feeling of clarity. Nguyen (2021) proposes clarity as an organizing principle that plays a role in managing our cognitive resources. Framing satisfaction as a subjective feeling of clarity would allow for a gap between actual understanding and an illusion of understanding (cf. miscalibration in our Study 1, or a distinction between actual learning and perceptions of learning in Liquin and Lombrozo 2022). Further, insofar as clarity is a seductive feeling, it opens people's cognitive systems to misdirection or even exploitation (Nguyen, 2021), which suggests a direction for further research on cognitive explanation in the context of science denial.

We identified a number of cognitive individual differences that contributed to explanation quality (though age, gender and education did not), but what other factors, including social or cultural ones, relate to the generation and evaluation of explanations? Social conservatism is associated with greater reliance on the inherence heuristic in explanation (Cimpian & Salomon, 2014; Hussak & Cimpian, 2018). It is also associated with lower trust in science (Gauchat, 2012; Sulik et al., 2021), so this raises a further angle on the issue mentioned at the end of the previous paragraph.

Satisfaction was predicted by function independently of causation — further supported in a follow-up study. In contrast, in-lab studies with experimenter-generated explanations have found appeals to function to be dependent on causal connections (Lombrozo & Carey, 2006). What accounts for this difference? The relationship between these two content features is likely complex (Liquin & Lombrozo, 2018; Lombrozo & Gwynne, 2014; McCarthy & Keil, 2022) but an 'explanations in the wild' approach – as a complement to traditional lab studies – can further help identify the circumstances in which causation and function contribute to explanation quality.

The issue may turn out to depend on lay vs expert conceptions of causation and/or function. Or maybe the crucial difference is between experimenter-generated explanations and more informal participant-generated explanations. It could hinge on the use of counterfactuals in Lombrozo and Carey (2006).

However, it could also tell us something about the affordances of different study designs to probe different aspects of cognition. Work on reasoning and problem solving has shown that humans behave more normatively when a contrast class is explicitly demanded by experimenters, compared to when they solve the problem in a more unconstrained fashion, in which case they typically do not spontaneously generate a contrast class on their own (Gale & Ball, 2006; Gorman et al., 1987). In a perspective-taking task, participants do poorly when open-endedly generating signals, but are able to do better (and make more conventional decisions) when choosing from a list of experimenter-generated signals (Sulik & Lupyan, 2018). Chin-Parker and Bradner (2010) found differences in causal vs functional explanation when participants generated explanations compared to when they evaluated them.

The overarching theme is that different cognitive processes may be involved in open-endedly generating something and evaluating the same thing in a more constrained context, despite these both connecting with the same phenomenon. If explanation plays a complex role in cognition, it is worth understanding it from both perspectives, especially if these sometimes yield apparently conflicting results.

This matters for the bigger picture because we highlighted connections to pragmatic communication and insight above. A core feature of pragmatic communication is its open-endedness or lack of contextual constraint (Sperber & Wilson, 1995), and one characterization of insight problems (in addition to the famous 'Aha!' moment) is that relevant problem representations are hard to generate but easy to evaluate (Bowden & Jung-Beeman, 2007).

To advance the cognitive science of explanation in a way that engages with naturalistic approaches to pragmatics and insight, it is necessary to study open-ended generation, which is an aspect of explanation in the wild.

6.0.3. Limits to generality

The explanations produced by our participants were typically short, so we may not be capturing properties (e.g., logical soundness) that might only appear in longer and more formal explanations. We have also focused on features of content, rather than on more structural properties such as consistency (Zemla et al., 2017), but we have made our data open (https://osf.io/wbxcj/) and encourage others to take up that challenge.

We followed Zemla et al. (2017) in treating explanation length (here, by counting new content word types) as indicating level of detail, and thus one way to measure the relative complexity or simplicity of an explanation. However, this is not how explanatory simplicity is usually meant, limiting the generality of our conclusions. For instance, Lombrozo (2007) counts causes rather than word types. Sober (2002) points out that it is not entirely clear what to count: causes, types, changes, predicates, assumptions, implications, and so on. Even if we know what to count, simplicity is not merely a matter of counting. Thagard (2006) states that 'Simplicity is a matter of explaining a lot with few assumptions,' highlighting the relative or contextually-sensitive nature of simplicity. In any case, whereas researcher-generated explanations may make it easier to vary (and thus count) the number of causes mentioned, our explanations-in-the-wild approach has produced a great deal of variation in sentence structures and levels of clarity or vagueness, sometimes explicitly mentioning causes and sometimes merely implying them, so it was not always possible to distinguish the number of causes referred to.

We chose to focus on certain kinds of content – such as causation, function and mechanism – that are ubiquitous in real-world phenomena studied by empirical sciences (including social sciences) but that are familiar to laypeople (Keil, 2006). However, our conclusions about these content features will not extend to other kinds of explanation such as mathematical explanation, where causal relations do not obviously apply (Keil, 2006), and where different theoretic virtues may be valued (Inglis & Aberdein, 2015). However, it is intriguing that the latter study, which included an exploratory factor analysis of mathematicians' appraisals of mathematical proofs, found that 'insightful' proofs scored highly on two latent dimensions: aesthetics and utility. This raises the question whether our results relating satisfaction and insight in Study 2 could be extended to show something similarly two-dimensional.

Causation is a complex concept but our instructions to participants provided relatively simple descriptions of it. Despite the simplicity of the instructions, participants rated diverse things as causal: abstract ultimate causes (God, nature, evolution); chemical composition and molecular structure; physical properties such as shape, size, or material strength; biological phenomena such as DNA and anatomical differences; environmental or historical contexts; and internal states (sensory perceptions, desires, beliefs, etc.). Our open data will allow others to conduct conceptual analyses or machine learning, including how varying conceptions of causation map onto ratings of mechanism or function.

Our data show whether a generated explanation appeals to causes. However, causation plays a role in other cognitive phenomena which goes well beyond this presence/absence dimension. These phenomena include generic sentences (Prasada et al., 2013), inherent essences

(Cimpian & Salomon, 2014; Cimpian & Steinberg, 2014), category judgments (Rehder & Hastie, 2001), and even high-level epistemic frameworks such as intuitive theories (Gerstenberg & Tenenbaum, 2017) or models of understanding (Lombrozo & Wilkenfeld, 2019). Across these phenomena, the concerns include philosophical theories of causality itself; how humans represent causal relations; the role causation plays in structuring our world knowledge; and how the nature of causal representation impacts cognitive processes such as inference. Our content-based ratings of causation are limited in that they do not directly track these concerns, yet they may still complement current approaches to causal representation and causal reasoning.

For example, Johnson and Ahn (2015) were interested in how causation structures our world knowledge. In particular, they tried to identify when a causal chain such as $A \rightarrow B \rightarrow C$ is stored in world knowledge as this three-part schema vs when it is stored as two chunks $A \rightarrow B$ and $B \rightarrow C$. One way they addressed this question was to ask participants to rate the extent to which they would explicitly mention intermediate cause B if they were to explain to someone else how A led to C. In contrast to this rated hypothetical, our approach could reveal when people actually mention such intermediate causes as (anecdotally) high mechanism ratings in our data often reflected appeals to intermediate causes.

Further, Wolff (2007) investigated various models of how people represent causes in a more physics sense, building on a linguistic analysis of how different causal verbs map onto different causal relations (Wolff & Song, 2003). The latter adopted a top-down approach that first identified causal verbs, harvested example sentences containing those verbs from a linguistic corpus, and then used these sentences as stimuli for similarity ratings that fed into multidimensional scaling. In contrast, our corpus of everyday explanation texts could offer a more bottom-up way to conduct a similar linguistic analyses as it reflects words that people spontaneously used to explain, where these explanations are rated for causal content (among other things).

Finally, Dündar-Coecke et al. (2022) identified three 'mechanism domains' (mechanical, chemical, and electromagnetic) from a clustering analysis of 42 artifacts where participants grouped items based on mechanism or function or similarity. They subsequently showed that mechanism domain influenced participants' causal reasoning. Dündar-Coecke et al. (2022) described mechanism to participants as 'how something works' whereas our instructions described it as 'how something happens'. However, the examples in our instructions overlapped heavily with the cases they identified, so our data may reflect a compatible – if broader – concept. If so, and with the addition of some distance metrics, our data could support a broader range of mechanism domains including, for instance, psychological mechanisms or cultural evolution mechanisms.

6.0.4. Conclusions

Overall, our studies extend a recent call for a more cognitive view of explanation (Horne et al., 2019) motivating two proposals for what such a theory of explanation must account for: explanation as a communicative, interactive phenomenon, and explanation as an open-ended relevance-deciding problem. Even though a normative, epistemic account of explanation is important for scientific progress, people require significant training to develop the expertise necessary for explanation in that formal sense. By understanding how laypeople evaluate lay explanations, we will better understand both the cognitive abilities that modern scientific theorizing emerged from, or how scientific explanations, when communicated to the public, can be made to feel more satisfying.

CRediT authorship contribution statement

Justin Sulik: Developed the study concept and design, Data collection and analysis, Drafted the manuscript. **Jeroen van Paridon:** Computational model for explanation domain tagging was developed and validated, Provided critical revisions. **Gary Lupyan:** Developed the study concept and design, Provided critical revisions.

Data availability

All data are available at https://osf.io/wbxcj/.

Acknowledgments

This research was supported by an award from The Templeton Foundation to the Metaknowledge Research Network, USA, and by NSF INSPIRE, USA 1344279 and NSF BCS, USA 2020969 to GL. JS was funded by the NOMIS Foundation, Switzerland through the project "Diversity in Social Environments". The funders had no role in the conduct of this research.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.cognition.2023.105464.

References

Bechlivanidis, C., Lagnado, D. A., Zemla, J. C., & Sloman, S. (2017). Concreteness and abstraction in everyday explanation. *Psychonomic Bulletin & Review*, 1–14.

Bowden, E. M., & Jung-Beeman, M. (2003). Normative data for 144 compound remote associate problems. *Behavior Research Methods, Instruments, & Computers*, *35*(4), 634–639.

Bowden, E. M., & Jung-Beeman, M. (2007). Methods for investigating the neural components of insight. *Methods*, *42*, 87–99.

Bowden, E. M., Jung-Beeman, M., Fleck, J., & Kounios, J. (2005). New approaches to demystifying insight. *Trends in Cognitive Sciences*, *9*(7), 322–328.

Brewer, W. F., Chinn, C. A., & Samarapungavan, A. (1998). Explanation in scientists and children. *Minds and Machines*, *8*(1), 119–136.

Bürkner, P.-C., & Vuorre, M. (2019). Ordinal regression models in psychology: A tutorial. *Advances in Methods and Practices in Psychological Science*, *2*(1), 77–101.

Chin-Parker, S., & Bradner, A. (2010). Background shifts affect explanatory style: how a pragmatic theory of explanation accounts for background effects in the generation of explanations. *Cognitive Processing*, *11*(3), 227–249.

Cimpian, A., & Salomon, E. (2014). The inherence heuristic: An intuitive means of making sense of the world, and a potential precursor to psychological essentialism. *Behavioral and Brain Sciences*, *37*, 461–527.

Cimpian, A., & Steinberg, O. D. (2014). The inherence heuristic across development: Systematic differences between children's and adults' explanations for everyday facts. *Cognitive Psychology*, *75*, 130–154.

Colombo, M. (2017). Experimental philosophy of explanation rising: The case for a plurality of concepts of explanation. *Cognitive Science*, *41*(2), 503–517.

Colombo, M., Bucher, L., & Sprenger, J. (2017). Determinants of judgments of explanatory power: Credibility, generality, and statistical relevance. *Frontiers in Psychology*, *8*(1430).

Condon, D. M., & Revelle, W. (2014). The international cognitive ability resource: Development and initial validation of a public-domain measure. *Intelligence*, *43*, 52–64.

Cor, M. K., Haertel, E., Krosnick, J. A., & Malhotra, N. (2012). Improving ability measurement in surveys by following the principles of IRT: The wordsum vocabulary test in the general social survey. *Social Science Research*, *41*(5), 1003–1016.

Cummins, R. (2000). How does it work? versus what are the laws?: Two conceptions of psychological explanation. In F. C. Keil, & R. A. Wilson (Eds.), *Explanation and cognition* (pp. 117–144). Cambridge, MA: MIT Press, chapter 5.

Deutsch, D. (2011). *The beginning of infinity: explanations that transform the world*. Penguin UK.

Dündar-Coecke, S., Goldin, G., & Sloman, S. A. (2022). Causal reasoning without mechanism. *Plos One*, *17*(5), Article e0268219.

Durso, F. T., Rea, C. B., & Dayton, T. (1994). Graph-theoretic confirmation of restructuring during insight. *Psychological Science*, *5*(2), 94–98.

Faye, J. (2007). The pragmatic-rhetorical theory of explanation. In J. Persson, & P. Ylikoski (Eds.), *Rethinking explanation* (pp. 43–68). Springer.

Frazier, B. N., Gelman, S. A., & Wellman, H. M. (2016). Young children prefer and remember satisfying explanations. *Journal of Cognition and Development*, *17*(5), 718–736.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19*(4), 25–42.

Gale, M., & Ball, L. J. (2006). Dual-goal facilitation in wason's 2–4–6 task: What mediates successful rule discovery? *The Quarterly Journal of Experimental Psychology*, *59*(05), 873–885.

Gauchat, G. (2012). Politicization of science in the public sphere: A study of public trust in the United States, 1974 to 2010. *American Sociological Review*, *77*(2), 167–187.

Gerstenberg, T., & Tenenbaum, J. B. (2017). Intuitive theories. In M. R. Waldmann (Ed.), *The oxford handbook of causal reasoning*. Oxford University Press.

Gilhooly, K. J., & Murphy, P. (2005). Differentiating insight from non-insight problems. *Thinking & Reasoning*, *11*(3), 279–302.

Gopnik, A. (2000). Explanation as orgasm and the drive for causal knowledge: The function, evolution and phenomenology of the theory formation system. In F. C. Keil, & R. A. Wilson (Eds.), *Explanation and cognition* (pp. 299–323). MIT Press, chapter 12.

Gorman, M. E., Stafford, A., & Gorman, M. E. (1987). Disconfirmation and dual hypotheses on a more difficult version of wason's 2–4–6 task. *The Quarterly Journal of Experimental Psychology Section A*, *39*(1), 1–28.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). Learning word vectors for 157 languages. arXiv preprint arXiv:1802.06893.

Hempel, C. (1965). *Aspects of scientific explanation and other essays in the philosophy of science*. New York: Free Press.

Hickling, A. K., & Wellman, H. M. (2001). The emergence of children's causal explanations and theories: evidence from everyday conversation. *Developmental Psychology*, *37*(5), 668.

Hilton, D. J., & Erb, H.-P. (1996). Mental models and causal explanation: Judgements of probable cause and explanatory relevance. *Thinking & Reasoning*, *2*(4), 273–308.

Hopkins, E. J., Weisberg, D. S., & Taylor, J. C. (2016). The seductive allure is a reductive allure: People prefer scientific explanations that contain logically irrelevant reductive information. *Cognition*, *155*, 67–76.

Horne, Z., Muradoglu, M., & Cimpian, A. (2019). Explanation as a cognitive process. *Trends in Cognitive Sciences*, *23*(3).

Hussak, L. J., & Cimpian, A. (2018). Investigating the origins of political views: Biases in explanation predict conservative attitudes in children and adults. *Developmental Science*, *21*(3), Article e12567.

Inglis, M., & Aberdein, A. (2015). Beauty is not simplicity: An analysis of mathematicians' proof appraisals. *Philosophia Mathematica*, *23*(1), 87–109.

Johnson, S. G., & Ahn, W.-K. (2015). Causal networks or causal islands? The representation of mechanisms and the transitivity of causal judgment. *Cognitive Science*, *39*(7), 1468–1503.

Joo, S., Yousif, S. R., & Keil, F. C. (2021). Understanding 'why': How implicit questions shape explanation preferences. PsyArXiv.

Joo, S., Yousif, S. R., & Knobe, J. (2023). Teleology beyond explanation. *Mind & Language*, *38*, 20–41.

Keil, F. C. (2006). Explanation and understanding. *Annual Review of Psychology*, *57*, 227–254.

Kelemen, D., & Rosset, E. (2009). The human function compunction: Teleological explanation in adults. *Cognition*, *111*(1), 138–143.

Kelemen, D., Rottman, J., & Seston, R. (2013). Professional physical scientists display tenacious teleological tendencies: purpose-based reasoning as a cognitive default. *Journal of Experimental Psychology: General*, *142*(4), 1074–1083.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77*(6), 1121.

Kuhn, T. S. (1977). Objectivity, value judgment, and theory choice. In T. S. Kuhn (Ed.), *The essential tension: select studies in scientific tradition and change* (pp. 74–86). Chicago, IL: University of Chicago Press.

Legare, C. H. (2012). Exploring explanation: Explaining inconsistent evidence informs exploratory, hypothesis-testing behavior in young children. *Child Development*, *83*(1), 173–185.

Legare, C. H. (2014). The contributions of explanation and exploration to children's scientific reasoning. *Child Development Perspectives*, *8*(2), 101–106.

Lim, J. B., & Oppenheimer, D. M. (2020). Explanatory preferences for complexity matching. *PLoS One*, *15*(4), Article e0230929.

Liquin, E. G., & Lombrozo, T. (2018). Structure-function fit underlies the evaluation of teleological explanations. *Cognitive Psychology*, *107*, 22–43.

Liquin, E. G., & Lombrozo, T. (2022). Motivated to learn: An account of explanatory satisfaction. *Cognitive Psychology*, *132*, Article 101453.

Litman, L., Robinson, J., & Abberbock, T. (2017). Turkprime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, *49*(2), 433–442.

Litman, J. A., & Spielberger, C. D. (2003). Measuring epistemic curiosity and its diversive and specific components. *Journal of Personality Assessment*, *80*(1), 75–86.

Lombrozo, T. (2006). The structure and function of explanations. *Trends in Cognitive Sciences*, *10*(10), 464–470.

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

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

Lombrozo, T., & Carey, S. (2006). Functional explanation and the function of explanation. *Cognition*, *99*, 167–204.

Lombrozo, T., & Gwynne, N. Z. (2014). Explanation and inference: mechanistic and functional explanations guide property generalization. *Frontiers in Human Neuroscience*, *8*, 700.

Lombrozo, T., & Wilkenfeld, D. (2019). Mechanistic versus functional understanding. In S. R. Grimm (Ed.), *Varieties of understanding: new perspectives from philosophy, psychology, and theology*. Oxford University Press.

Malhotra, N., Krosnick, J. A., & Haertel, E. (2007). The psychometric properties of the gss wordsum vocabulary test. *11*, *GSS methodological report*.

Mancosu, P. (2001). Mathematical explanation: Problems and prospects. *Topoi*, *20*(1), 97–117.

McCarthy, A., & Keil, F. (2022). A right way to explain? function, mechanism, and the order of explanations. In J. Culbertson, A. Perfors, H. Rabagliati, & V. Ramenzoni (Eds.), *Vol. 44*, *Proceedings of the annual meeting of the cognitive science society*. page Retrieved from https://escholarship.org/uc/item/6666c13t.

Mejía-Ramos, J. P., Alcock, L., Lew, K., Rago, P., Sangwin, C., & Inglis, M. (2019). Using corpus linguistics to investigate mathematical explanation. *Methodological Advances in Experimental Philosophy*, 239–264.

Mercier, H., & Sperber, D. (2011). Why do humans reason? arguments for an argumentative theory. *Behavioral and Brain Sciences*, *34*, 57–111.

Mercier, H., & Strickland, B. (2012). Evaluating arguments from the reaction of the audience. *Thinking & Reasoning*, *18*(3), 365–378.

Mills, C. M., Sands, K. R., Rowles, S. P., & Campbell, I. L. (2019). I want to know more!: Children are sensitive to explanation quality when exploring new information. *Cognitive Science*, *43*.

Motamedi, Y., Little, H., Nielsen, A., & Sulik, J. (2019). The iconicity toolbox: empirical approaches to measuring iconicity. *Language and Cognition*, *11*(2), 188–207.

National Science Board (2018). Science & engineering indicators 2018.

Nguyen, C. T. (2021). The seductions of clarity. *Royal Institute of Philosophy Supplements*, *89*, 227–255.

Prasada, S. (2017). The scope of formal explanation. *Psychonomic Bulletin & Review*, 1–10.

Prasada, S., Khemlani, S., Leslie, S.-J., & Glucksberg, S. (2013). Conceptual distinctions amongst generics. *Cognition*, *126*(3), 405–422.

Rehder, B., & Hastie, R. (2001). Causal knowledge and categories: the effects of causal beliefs on categorization, induction, and similarity. *Journal of Experimental Psychology: General*, *130*(3), 323.

Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: an illusion of explanatory depth. *Cognitive Science*, *26*, 521–562.

Shenhav, A., Rand, D. G., & Greene, J. D. (2012). Divine intuition: Cognitive style influences belief in god. *Journal of Experimental Psychology: General*, *141*(3), 423–428.

Shtulman, A., & Valcarcel, J. (2012). Scientific knowledge suppresses but does not supplant earlier intuitions. *Cognition*, *124*(2), 209–215.

Sober, E. (2002). What is the problem of simplicity. In A. Zellner, H. A. Keuzenkamp, & M. McAleer (Eds.), *Simplicity, inference, and modelling*. Cambridge University Press.

Sperber, D., & Wilson, D. (1995). *Relevance: communication and cognition* (second ed.). Malden, MA: Blackwell Publishing.

Stalnaker, R. (1984). *Inquiry*. Cambridge University Press.

Sulik, J. (2018). Cognitive mechanisms for inferring the meaning of novel signals during symbolisation. *PLoS One*, *13*(1), Article e0189540.

Sulik, J., Deroy, O., Dezecache, G., Newson, M., Zhao, Y., El Zein, M., & Tunçgenç, B. (2021). Facing the pandemic with trust in science. *Humanities and Social Sciences Communications*, *8*(1), 1–10.

Sulik, J., & Lupyan, G. (2018). Perspective taking in a novel signaling task: effects of world knowledge and contextual constraint. *Journal of Experimental Psychology: General*, *147*(11), 1619–1640.

Thagard, P. (1978). The best explanation: Criteria for theory choice. *The Journal of Philosophy*, *75*(2), 76–92.

Thagard, P. (2006). Evaluating explanations in law, science, and everyday life. *Current Directions in Psychological Science*, *15*(3), 141–145.

Thomson, K. S., & Oppenheimer, D. M. (2016). Investigating an alternate form of the cognitive reflection test. *Judgment and Decision Making*, *11*(1), 99–113.

Van Paridon, J., & Thompson, B. (2020). Subs2vec: Word embeddings from subtitles in 55 languages. *Behavior Research Methods*, 1–27.

Walker, C. M., Lombrozo, T., Legare, C. H., & Gopnik, A. (2014). Explaining prompts children to privilege inductively rich properties. *Cognition*, *133*, 343–357.

Weisberg, D. S., Keil, F. C., Goodstein, J., Rawson, E., & Gray, J. R. (2008). The seductive allure of neuroscience explanations. *Journal of Cognitive Neuroscience*, *20*(3), 470–477.

Williams, J. J., & Lombrozo, T. (2010). The role of explanation in discovery and generalization: Evidence from category learning. *Cognitive Science*, *34*(776–806), 776–806.

Wojtowicz, Z., & DeDeo, S. (2020). From probability to consilience: How explanatory values implement bayesian reasoning. *Trends in Cognitive Sciences*, *24*(12), 981–993.

Wolff, P. (2007). Representing causation. *Journal of Experimental Psychology: General*, *136*(1), 82.

Wolff, P., & Song, G. (2003). Models of causation and the semantics of causal verbs. *Cognitive Psychology*, *47*(3), 276–332.

Woodward, J. (2019). Scientific explanation. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy (winter 2019 edition)*.

Wright, L. (1976). *Teleological explanations: an etiological analysis of goals and functions*. Univ of California Press.

Zemla, J. C., Sloman, S., Bechlivanidis, C., & Lagnado, D. A. (2017). Evaluating everyday explanations. *Psychonomic Bulletin & Review*, *24*, 1–13.