Explanations in the wild

Justin Sulik[a*], Jeroen van Paridon[b], and 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

Author Note

*Corresponding author: justin.sulik@gmail.com

Abstract

Why do some explanations strike people as highly satisfying while others, seemingly equally accurate, satisfy them less? We asked lay-people 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.

Keywords: Explanations, Creativity, Metacognition, Insight, Perspective-taking

Explanations in the wild

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 prediction (Woodward, 2019) or how well they embody theoretic virtues such as parsimony or simplicity (Kuhn, 1977; Thagard, 1978). It has also influenced psychological research on explanations, for instance leading to proposals for how Bayesian models of cognition might implement various explanatory virtues (Wojtowicz and DeDeo, 2020), or that test whether people prefer explanations that facilitate prediction (Lombrozo and 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 and theoretic virtues — is therefore likely to misrepresent what kinds of explanations lay-people consider good. Here, rather than focusing on how scientific theories explain, we focus on how *people* explain. We do this by examining explanations generated by lay people to answer 'Why?' questions about common phenomena; by having other lay people evaluate the quality of those explanations; and by trying to understand what predicts these evaluations of quality.

As an alternative to understanding explanation quality in terms of epistemic goals or theoretic virtues, we ask how satisfying they are (Liquin and Lombrozo, 2022; Lombrozo, 2007). Specifically, what distinguishes satisfying explanations from less satisfying ones? This approach to quantifying explanation quality builds on the notion of 'explanation as orgasm' (Gopnik, 2000), which proposes that part of the reason people generate explanations is the satisfaction this brings them.

However, as a complement this idea that explanation is a satisfying process for the explainer, we wish to understand what makes explanations satisfying to their audience. Specifically, how do explanations generated by lay-people strike other lay-people? 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 and Sperber, 2011; Mercier and 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 lay-people generate free-form responses to a range of ordinary 'Why?' questions and other lay-people rate how satisfying or accurate they find these responses — contrasts with previous empirical studies of explanation. One common approach is to have participants evaluate explanations that were carefully created by experimenters to contain specific features of interest (Colombo et al., 2017; Hopkins et al., 2016; Lombrozo and Carey, 2006); another is to analyse participants' explanations of artificial scenarios crafted by experimenters for similar purposes (Lombrozo and Gwynne, 2014). Such experimental control is certainly valuable, yet these studies offer limited insight into the kinds of explanations that are produced and shared by non-experts. 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.

One study has, like ours, examined explanations for familiar phenomena that were not generated by the researchers. 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. A limitation of this study is that many of the explanations were generated by domain experts rather than by lay people. 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 lay people provide to each other, because lay people may be designing their explanations based on entirely different considerations. A second limitation is the use of just 24 explanations focusing on socio-historical topics. This contrasts with our analysis of 1000 explanations across various domains in Study 1 (example domains shown in Table 1). Finally, unlike Zemla et al., we impose no filter on explanation quality, allowing us to examine natural variation in perceived quality, which we expect to be larger in lay explanations than in expert explanations.

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?

We use the ‘explanations in the wild’ approach to answer three questions about the psychology of 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. To decide which features to measure, we start with Legare’s definition of an explanation as ‘an attempt to understand a causal relation by identifying relevant functional or mechanistic information’ (2014)[1]. To this end, we used

[1] We choose to focus on this definition because it strikes a good balance: its scope is broad enough to cover

independent non-expert raters (who did not produce the explanations) to rate how much each explanation appeals to common-sense 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*). In addition, we had participants rate how general the explanations were. An early thread in the philosophy of science framed explanation as appeal to general laws such as gravity (Hempel, 1965). Although we suspected that appeals to universal laws like this would be infrequent in non-expert explanations, we posited that one aspect of such laws — generality — would be both common-place and easy for non-experts to understand.

Our second aim is to probe a metacognitive question: Do lay people know how good their own explanations are? If people are well calibrated, then explanation generators and raters should concur in their assessments. Otherwise, people might overestimate their ability to explain (an 'Illusion of Explanatory Depth', Rozenblit and Keil, 2002). Alternatively, people's tendency to overestimate their ability may depend on the level of that ability. People who generate worse explanations may also be less well calibrated in assessing them (the 'Dunning-Kruger Effect', Kruger and Dunning, 1999).

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

many instances of explanations in people's lives, yet it is still relatively straightforward. Nonetheless, we note that it does not cover all possible cases of explanation. On the more cognitive side, there are some explanations that appeal to inherence (Cimpian and Salomon, 2014; Cimpian and Steinberg, 2014) or to something's form (Prasada, 2017). On the more philosophical side, some explanations appeal subsumption, or unifying disparate elements under a common umbrella (Hempel, 1965). However, we think it is worth keeping the above definition as our starting point for understanding explanations in the wild, because it focuses on factors that are common to both philosophical and psychological accounts.

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 what kinds of 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.

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).

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. 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 (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 wouldn’t 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. No participant rated more than one question or more than one feature. Ratings for each feature were averaged for each explanation.

2.2 Results

All data and full analysis scripts (including model specification, priors, and random effects structures) are available at https://osf.io/wbxcj/.

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.

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.

Overall, our explanations-in-the-wild approach produced explanations that were judged to be moderately accurate (M=4.74, SD=0.88), as well as satisfying (M= 4.33, SD=0.99). Importantly for predicting explanation quality, there was a reasonable amount of variance in both variables (Fig. 1a).

Although satisfaction and perceived accuracy were strongly correlated ($r = 0.67$, $p < 0.001$), 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).

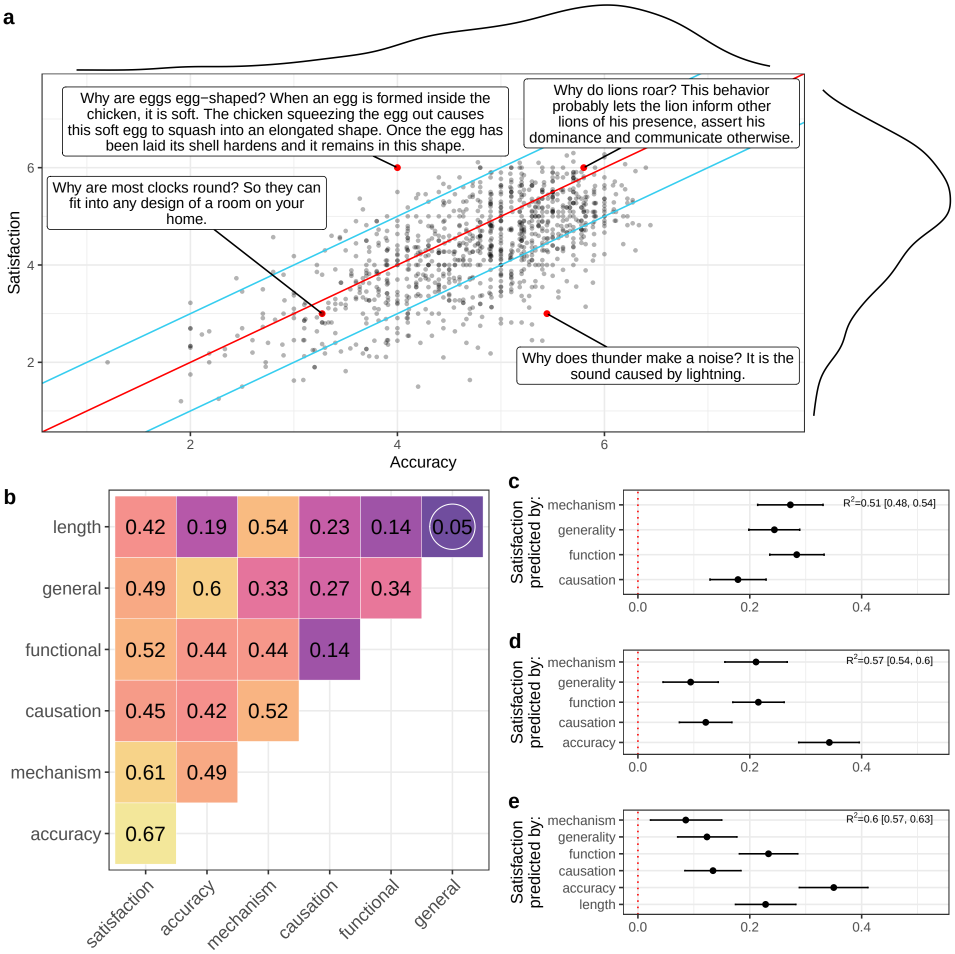
2.2.2 What features of explanations predict greater satisfaction?

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 ($r = 0.05$, $p = 0.15$). 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. It is by no means certain that people always value explanations according to how accurate and simple they are: The prevalence of conspiracy theories is merely one compelling counterexample, and part of the point of an 'explanations in the wild' approach is to look beyond the narrow constraints of normative views of explanation. Thus, we wanted to understand how content features predicted satisfaction both with and without accuracy and explanation

Figure 1

Relationships between quality and content features



Why are eggs egg−shaped? When an egg is formed inside the chicken, it is soft. The chicken squeezing the egg out causes this soft egg to squash into an elongated shape. Once the egg has been laid its shell hardens and it remains in this shape.

Why do lions roar? This behavior probably lets the lion inform other lions of his presence, assert his dominance and communicate otherwise.

Why are most clocks round? So they can fit into any design of a room on your home.

Why does thunder make a noise? It is the sound caused by lightning.

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.

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 about half of the variance ($R^2 = 0.51\,[0.48, 0.54]$). Of note, mechanism, despite having the largest zero-order correlation with satisfaction, was not the strongest predictor in this multiple regression.

With perceived accuracy added as a covariate (Fig. 1d), the 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 ($\beta = 0.32\,[0.28, 0.35]$), though this still left a smaller direct effect ($\beta = 0.14\,[0.08, 0.19]$). We conjecture that appeals to general truths mostly improve an explanation's satisfaction by improving its perceived accuracy.

Adding explanation length as an additional covariate (Fig. 1e) — one way of construing simplicity across the wide range of explanations produced here[2] — further reduced the predictive power of the content features. Longer explanations were judged to be more satisfying ($\beta = 0.23\,[0.17, 0.28]$). Importantly, the predictive effect of perceived accuracy was unchanged ($\beta = 0.35\,[0.29, 0.41]$), 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 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, not necessarily the *number* of causes — is one way to improve satisfaction at the expense of simplicity. This expands the list of potential reasons why longer explanations

[2] Naturally, there are several senses of 'simplicity' and some of these (e.g., Thagard, 1978) are not easily captured by our 'explanations in the wild' approach.

may be more satisfying, such the finding that more complex phenomena require longer explanations (Lim and 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.

Causal and functional explanations are often described as two distinct styles of explanation (Chin-Parker and Bradner, 2010). The zero-order correlation between function and causation was $r = 0.14$ (see Fig. 1b), but regressing both of these features on accuracy reveals a residual correlation between causation and function to be negative ($r = -0.15\,[-0.22, -0.09]$). Thus, controlling for perceived accuracy, the more functional an explanation is, the less likely it is to be causal (and vice versa). However, this correlation is small, meaning that an explanation can contain elements of both types. 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 (mean satisfaction rating: 5.18; mean accuracy rating: 5.64) mentions that low self-esteem can cause people to bully, and that a function of bullying is to make themselves feel better.

A related question is how function and causation interact in predicting satisfaction. A study with researcher-generated explanations found that causal information was necessary for functional explanations to be considered acceptable (Lombrozo and 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 ($\beta = 0.43\,[0.38, 0.48]$) and the interaction term was smaller and negative ($\beta = -0.07\,[-0.11, -0.02]$). Thus, causation is not a necessary condition for people to find a functional explanation appealing, and the more an explanation leverages one of these to be satisfying, the less it typically leverages the other.

If causation represents one major style of explanation, it is surprising that it has a consistently weak effect on satisfaction (Fig. 1c–e). A Bayesian mediation model shows that causation had an indirect effect on satisfaction via mechanism ($\beta = 0.21\,[0.18, 0.24]$) in addition to a direct effect ($\beta = 0.17\,[0.12, 0.23]$). There was moderate evidence (BF=4.19) that the indirect effect is larger. Thus, a causal explanation is more of a *how?* than purely a *why?*

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.

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 and Thompson, 2020).

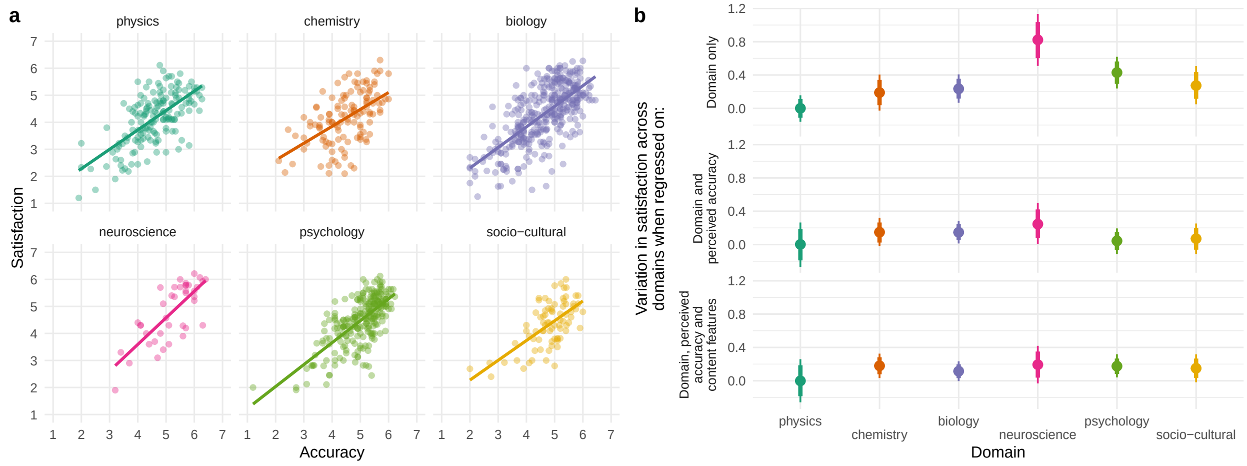
The relationship between perceived accuracy and satisfaction is consistent across domains (Fig. 2a), but how does satisfaction vary across domains? In answering this, we test two claims made in the literature concerning explanation domain: (1) that there is a 'seductive allure' of neuroscience explanations 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=163) 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 samples 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 neuroscience is no longer more satisfying than the other categories (all $BF_{01} > 7$). We thus caution against making

Figure 2

Explanation quality across domains



Note. (a) Scatterplot of accuracy and satisfaction, split by question domain, with linear fits. (b) Posterior samples (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.

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 ($\beta = -0.06$ $[-0.16, 0.12]$, $BF_{01} = 9.9$).

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 (as predicted by the Illusion of Explanatory Depth), 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 (as per the Dunning-Kruger effect[3]), then the slope will be negative.

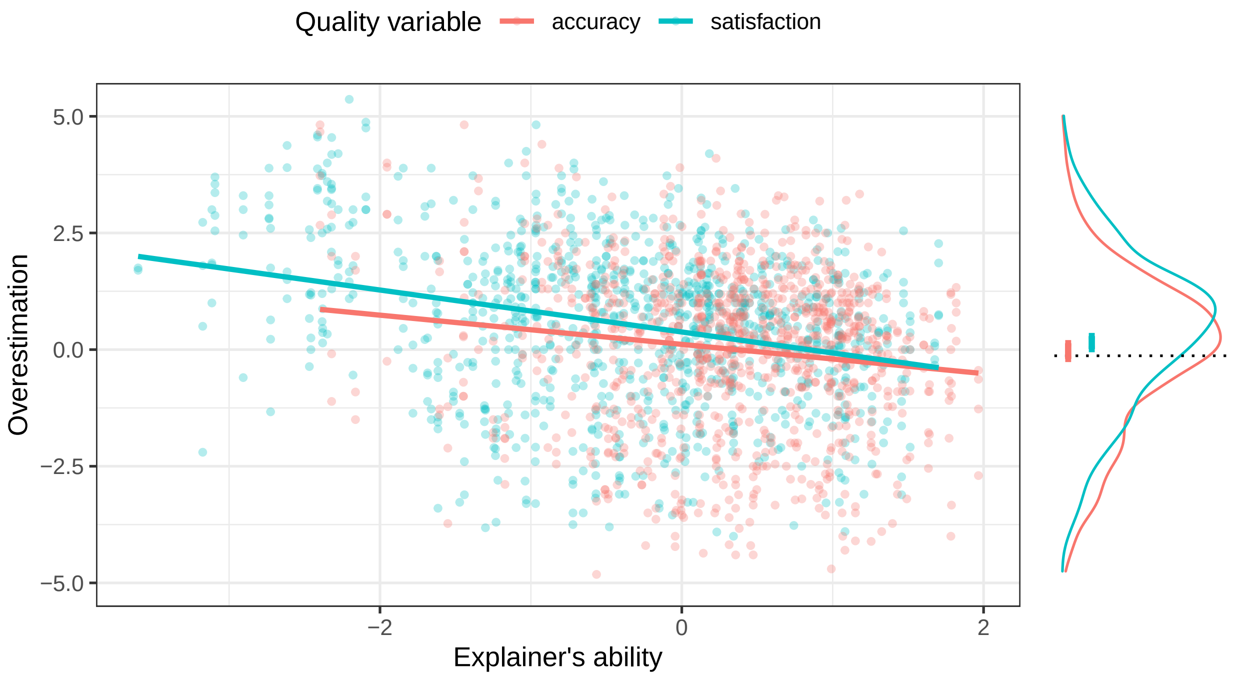
Overall, people did not overestimate how accurately others would perceive their explanations ($b = 0.11$ $[-0.07, 0.29]$, $BF_{01} = 5.6$). However, they did overestimate how satisfying their explanations would be for other people ($b = 0.29$ $[0.39, 0.64]$). People with lower ability showed more overestimation of perceived accuracy ($b = -0.29$ $[-0.44, -0.14]$, Fig. 3), and the slope for satisfaction was not different from that of accuracy (interaction term $b = -0.09$ $[-0.23, 0.05]$, $BF_{01} = 6.19$). Unlike the Illusion of Explanatory Depth, the

[3] We note that there has been some recent debate about this effect (e.g., Gignac and Zajenkowski, 2020). However, these technical nuances are beyond our scope, and we mention this name only as a point of reference: Our results here may stand on their own. Further, whereas the Dunning-Kruger effect is usually discussed in terms of percentile ratings, here we refer to absolute values.

Dunning-Kruger effect was visible both in perceived accuracy and satisfaction.

Figure 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. The Illusion of Explanatory Depth for satisfaction is evident in the fact that its CIs in the marginal plot exclude zero, unlike those of accuracy. The Dunning-Kruger effect is evident in the negative slopes in the main plot. The slopes for accuracy and satisfaction do not differ significantly.

Other than calibration between explainers and raters, how well were raters calibrated with one another? Were they equally consistent in rating satisfaction and perceived accuracy? We regressed the raw ratings on variable type (accuracy vs. satisfaction) along with random intercepts for question and explanation, and simultaneously regressed the model 'sigma' or 'scale' parameter on the same predictors. The sigma for satisfaction ($1.41\,[1.37, 1.45]$) was higher than for accuracy ($1.35\,[1.31, 1.39]$, difference in sigma=$0.06\,[0.05, 0.07]$), indicating that people were more varied in their ratings of satisfaction than of accuracy.

3 Study 2: Individual differences in cognitive traits

Our final research question was which cognitive mechanisms are associated with explainers' ability to produce high quality explanations. We pre-registered four hypotheses (https://osf.io/qw8ut):

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.

3.1 Methods

3.1.1 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 explanations and 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.

3.1.2 Materials

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

Insight ability: 20 Compound Remote Associate (CRA) problems (sampled from Bowden and 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 and Valcarcel, 2012).

Perspective-taking ability: 20 items from a communication game (Sulik and 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 and 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. (*Wordsumplus*, Cor et al., 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 and 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).

3.1.3 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 74 explanations from analysis that had been judged as not an explanation by more than one rater. 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.

3.2 Results

All data and full analysis scripts (including model specification, priors, random effects structures, and control variables such as age, gender and education) are available at https://osf.io/wbxcj/.

3.2.1 Pre-registered analyses

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

Fig. 4a 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 ($r = .099, p = .205$).

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.

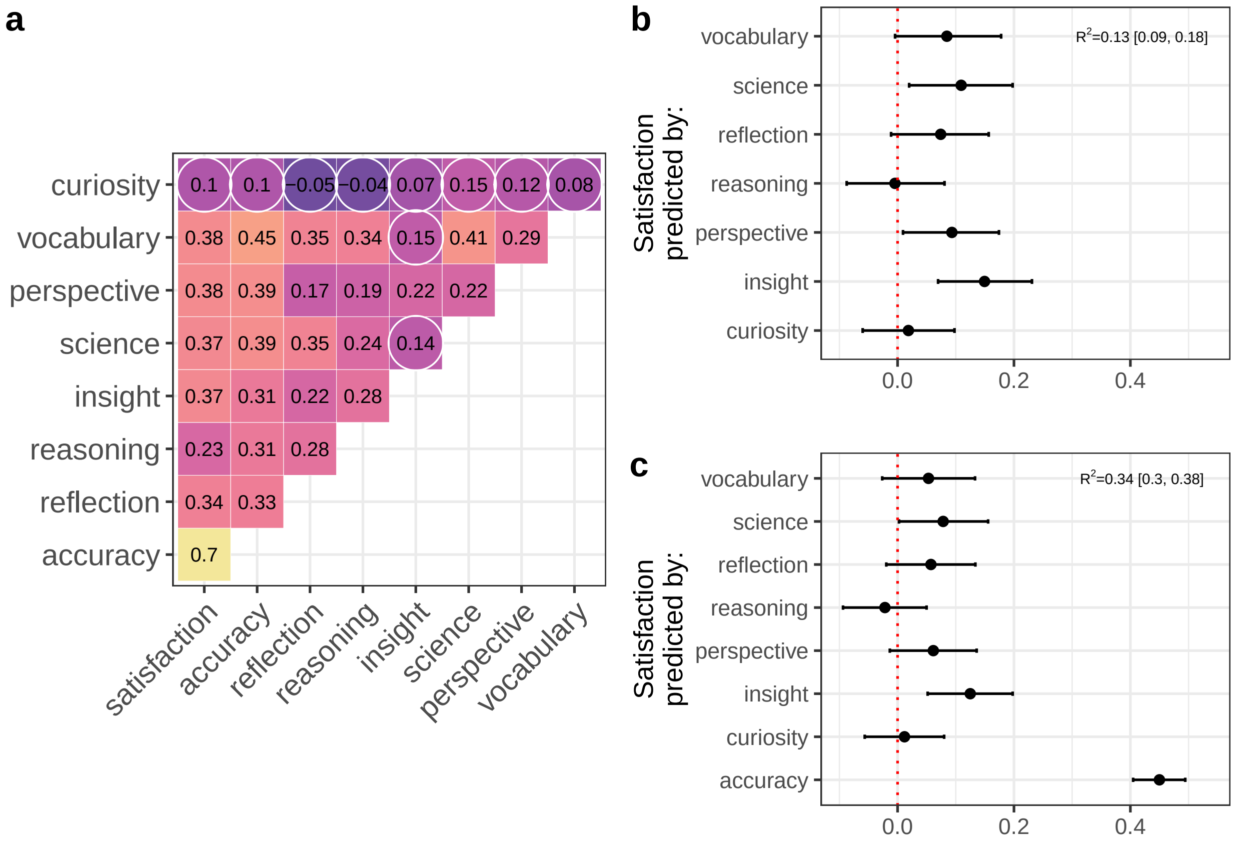
Fig. 4b shows standardized coefficients from a Bayesian multiple linear regression, predicting satisfaction from the various individual-differences measures. This and the following model include demographic control variables: age, gender, and education, though these had no effect in any of the models reported here (for numeric 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 ($\beta = 0.11$ $[0.04, 0.2]$); perspective taking ($\beta = 0.09$ $[0.01, 0.17]$) and insight ($\beta = 0.15$ $[0.07, 0.23]$). The others did not: curiosity ($\beta = 0.02$ $[-0.06, 0.10]$ $BF_{01} = 23$), verbal reasoning ($\beta = 0$ $[-0.09, 0.08]$ $BF_{01} = 23]$), and cognitive reflection ($\beta = 0.07$ $[-0.01, 0.16]$ $BF_{01} = 5$).

Fig. 4c 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 CIs that did not include 0. The CIs for perspective taking just overlapped 0 ($\beta = 0.06$ $[-0.01, 0.14]$, with an estimated posterior probability of .95 for a positive effect). For the others, there was moderate evidence that they had no effect (all $BF_{01} > 8$; see https://osf.io/wbxcj/ for

Figure 4

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 individual-differences measures; (c) Accuracy added as a co-variate to the model in (b). Model R^2s (for fixed effects) are shown as insets.

details).

3.2.2 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. 4c. None of these interaction terms had an effect (all $BF_{01} > 8$; 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 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. 4b (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. 5a shows the model structure, along with standardized coefficients for all paths (with 95% CIs). Fig. 5b 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).

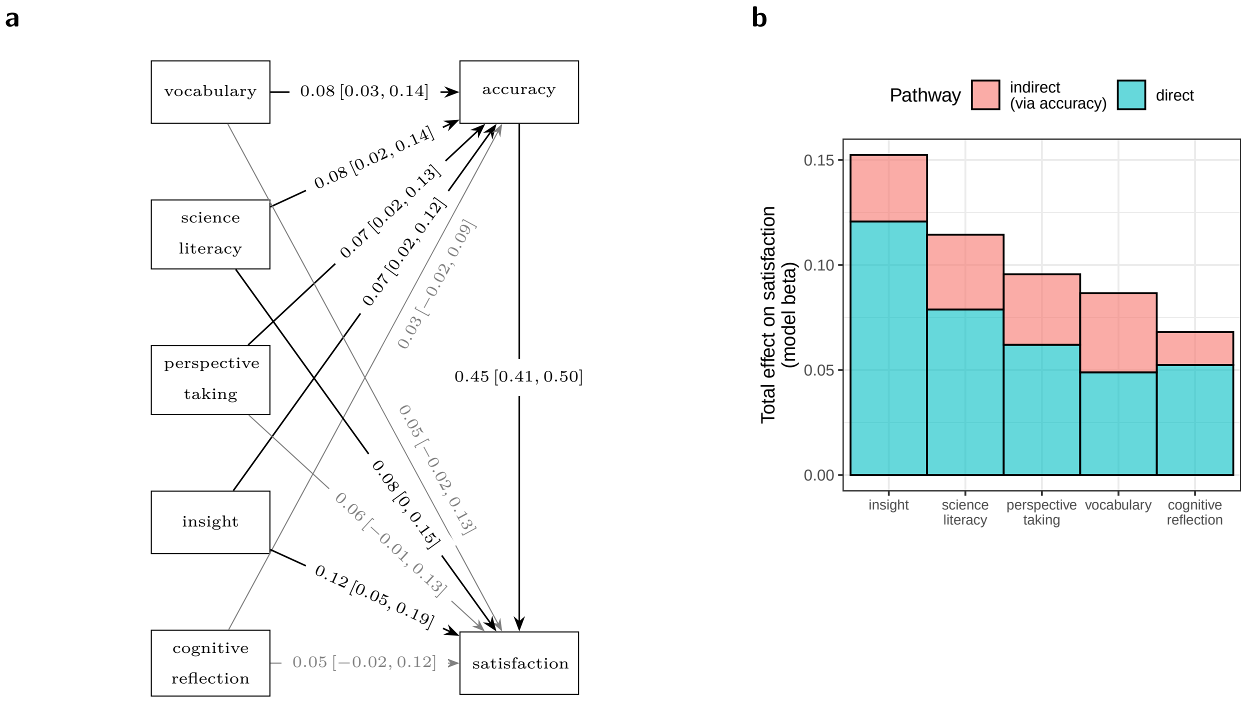
Considering the total effects, the ability to provide a satisfying explanation depended unsurprisingly on having the relevant knowledge (science literacy) and verbal intelligence (as measured by vocabulary). The direct effect for science literacy was larger than its indirect effect ($BF_{10} = 5.86$) 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.

Figure 5

Results of the Bayesian path analysis



Note. (a) Pathways comprising Bayesian simultaneous regressions. Pathways are labelled with standardized coefficients and 95% CIs. Paths are in grey if their CIs include 0. (b) Regression coefficients from the same Bayesian path 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 of each cognitive variable on satisfaction as mediated via accuracy.

4 General Discussion

What kinds of explanations do people judge as being good? We solicited thousands of explanations from lay-people in response to ‘Why?’ questions. Analyzing these explanations has allowed us to identify several predictors. Holding perceived accuracy constant, causation, function, and mechanism predicted unique variance in satisfaction. Explanation generality was more predictive of perceived accuracy than satisfaction. Explainers generally overestimated how satisfying their explanations were, though on average they did not overestimate the perceived accuracy of their explanations. For both accuracy and satisfaction, those participants who generated worse explanations also tended to overestimate the quality of their explanations more strongly. Ratings of explanation satisfaction were more varied than those for perceived accuracy.

The most important cognitive abilities for producing satisfying explanations were insight problem solving, science knowledge, and perspective taking. A good explanation goes beyond just including appropriate 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. 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 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). However, our finding that satisfaction is predicted by function *independently* of causation raises questions for future research: Why is it that function is sometimes independent of causation in predicting explanation equality, as here, and sometimes dependent on it, as in previous studies using experimenter-generated explanations Lombrozo and Carey (2006)? The relationship between these two content features is likely complex (Liquin and Lombrozo, 2018; Lombrozo and Gwynne, 2014; McCarthy and Keil, 2022) but an ‘explanations in the wild’ approach can further help identify the circumstances in which both causation and function contribute to explanation quality.

It is well established that people value functional explanations highly (Kelemen and Rosset, 2009; Kelemen et al., 2013; McCarthy and Keil, 2022; Wagner-Egger et al., 2018), but why was mechanism the other dominant feature? One view of explanation ('explanation for export', Lombrozo and Carey, 2006) holds that an explanation is about knowing how one would intervene on a system if one wanted to affect the outcome. Emphasizing mechanism (rather than just causation, which mostly affected satisfaction via mechanism) may be a way to do that. Whereas 'X causes Y' just states the existence of a relationship between cause and effect, a reference to mechanism ('X causes Y by doing Z') indicates how this occurs, highlighting what information is relevant if one wants to intervene on the system. However, another view holds that mechanism aids with the discovery of higher-order or more abstract principles (Keil, 2019). The precise reason for the importance of mechanism is thus another question for future research.

One limitation is that the explanations produced by participants were typically short, so we may not be capturing properties (e.g., logical soundness, circularity) that might only appear in longer or 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.

What implications do our results have for the psychology of explanation? In contrast to research that has focused on an epistemic theory of explanation (how explanations *ought* to work, given some epistemic goal such as prediction), our aims are directed at a more cognitive theory of explanation (how explanations work, given their role in daily mental life). In comparison to epistemic theories of explanations, cognitive theories of explanations are still in their infancy. Our results highlight several desiderata of a cognitive theory of explanations.

The first desideratum is considering lay explanation first and foremost as a communicative, interactive phenomenon (Faye, 2007; Keil, 2006), rather than as a process subserving internal theory formation (the somewhat solipsistic explanation-as-orgasm view). In Study 1, we showed that it was more challenging to judge how satisfactory 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 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. We attributed the importance of mechanism in Study 1 to its role in picking out what information is relevant, but there are other reasons to think that relevance must be central to the psychology of explanation. Insight, the strongest predictor of satisfaction in Study 2, involves finding a relevant representation of a problem (Durso et al., 1994; Gilhooly and 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.

Overall, our studies extent a recent call for a more cognitive view of explanation (Horne et al., 2019) by calling for a more socio-cognitive view, motivating two proposals for what such a theory of explanation must account for: explanation as a communicative, interactive phenomenon, and explanation as a 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 lay people evaluate lay explanations, we will better understand both the cognitive abilities that modern scientific theorizing emerged from, and how scientific explanations, when communicated to the public, can be made to feel more satisfying.

Author contributions

JS & GL developed the study concept and design. Data collection and analysis was performed by JS. The computational model for explanation domain tagging was developed and validated by JvP. JS drafted the manuscript, and GL and JvP provided critical revisions. All authors approved the final version of the manuscript for submission.

Acknowledgements

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

5 References

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

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

Chin-Parker, S. and 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. and 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. and 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., and Sprenger, J. (2017). Determinants of judgments of explanatory power: Credibility, generality, and statistical relevance. *Frontiers in Psychology*, 8:1430.

Condon, D. M. and 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., and 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 Keil, F. C. and Wilson, R. A., editors, *Explanation and Cognition*, chapter 5, pages 117–144. MIT Press, Cambridge, MA.

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

Durso, F. T., Rea, C. B., and 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 Persson, J. and Ylikoski, P., editors, *Rethinking Explanation*, pages 43–68. Springer.

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

Gignac, G. E. and Zajenkowski, M. (2020). The Dunning-Kruger effect is (mostly) a statistical artefact: Valid approaches to testing the hypothesis with individual differences data. *Intelligence*, 80:101449.

Gilhooly, K. J. and 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 Keil, F. C. and Wilson, R. A., editors, *Explanation and Cognition*, chapter 12, pages 299–323. MIT Press.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and 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.* Free Press, New York.

Hopkins, E. J., Weisberg, D. S., and 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., and Cimpian, A. (2019). Explanation as a cognitive process. *Trends in Cognitive Sciences*, 23(3).

Joo, S., Yousif, S. R., and Keil, F. C. (March 31 2021). Understanding ‘why’: How implicit questions shape explanation preferences. *PsyArXiv.*

Keil, F. (2019). The challenges and benefits of mechanistic explanation in folk scientific understanding. In Wilkenfeld, D. A. and Samuels, R., editors, *Advances in Experimental Philosophy of Science*, pages 41–57. Bloomsbury Publishing.

Keil, F. C. (2006). Explanation and understanding. *Annual Review of Psychology*, 57:227–54.

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

Kelemen, D., Rottman, J., and 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.

Kruger, J. and 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 Kuhn, T. S., editor, *The Essential Tension: Select Studies in Scientific Tradition and Change*, pages 74–86. University of Chicago Press, Chicago, IL.

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. and Oppenheimer, D. M. (2020). Explanatory preferences for complexity matching. *PloS One*, 15(4):e0230929.

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

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

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

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

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

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

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

McCarthy, A. and Keil, F. (2022). A right way to explain? function, mechanism, and the order of explanations. In Culbertson, J., Perfors, A., Rabagliati, H., and Ramenzoni, V., editors, *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44, page Retrieved from https://escholarship.org/uc/item/6666c13t.

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

Mercier, H. and 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., and 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., and 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.

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

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

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

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

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

Sulik, J. and 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.

Thomson, K. S. and Oppenheimer, D. M. (2016). Investigating an alternate form of the Cognitive Reflection Test. *Judgment and Decision Making*, 11(1):99–113.

Van Paridon, J. and Thompson, B. (2020). subs2vec: Word embeddings from subtitles in 55 languages. *Behavior Research Methods*, pages 1–27.

Wagner-Egger, P., Delouvée, S., Gauvrit, N., and Dieguez, S. (2018). Creationism and conspiracism share a common teleological bias. *Current Biology*, 28(16):R867–R868.

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

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

Woodward, J. (2019). Scientific explanation. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy (Winter 2019 Edition)*.

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