Running Head: EVALUATING EVERYDAY EXPLANATIONS

Evaluating Everyday Explanations

Jeffrey C. Zemla and Steven Sloman

Brown University

Christos Bechlivanidis and David A. Lagnado

University College London

Correspondence concerning this article should be addressed to Jeffrey Zemla, Department of Cognitive, Linguistic, and Psychological Sciences, Brown University, Providence, RI 02912.

Contact: jzemla@brown.edu

Abstract

People frequently rely on explanations provided by others to understand complex phenomena. A fair amount of attention has been devoted to the study of scientific explanation, and less on understanding how people evaluate naturalistic, everyday explanations. Using a corpus of diverse explanations from Reddit's "*Explain Like I'm Five"* and other online sources, we assessed how well a variety of explanatory criteria predict judgments of explanation quality. We find that while some criteria previously identified as explanatory virtues do predict explanation quality in naturalistic settings, other criteria such as simplicity do not. Notably, we find that people have a preference for complex explanations that invoke more causal mechanisms to explain an effect. We propose that this preference for complexity is driven by a desire to identify enough causes to make the effect seem inevitable.

Evaluating Everyday Explanations

People are explanatory creatures. We often seek to generate explanations based on our own knowledge of how the world works. However our ability to generate complete explanations on our own is frequently inadequate. We may not have all of the evidence or the expertise to be able to form accurate models of complex phenomena. So we use the knowledge of experts, friends, and communities to piece together explanations. Our beliefs about science are not limited to intuitive preconceptions, but are also derived from scientists who inform us of how things work. Our beliefs about the economy are affected not only by our own experiences, but also by what economists and politicians tell us about large-scale financial systems. We rely on the explanations of others to form our own beliefs. How, then, do we evaluate the explanations of others?

Explanatory Criteria

A common view has emerged that the quality or value of an explanation can be determined by how well it satisfies a set of criteria known as *explanatory virtues* (Lipton, 2004; Thagard, 1978; Harman, 1965; Mackonis, 2003; Glymour, 2014; Lombrozo, 2011). However, there is disagreement about what counts as an explanatory virtue, how these virtues are defined and measured, and how they are weighted when we evaluate an explanation. Two commonly proposed virtues are simplicity and coherence. For example, a good explanation should be simple, requiring the fewest number of causes to explain a phenomenon (e.g., Lombrozo, 2007). A good explanation should also be coherent; it should be compatible with our existing beliefs and consistent with the evidence and with itself (e.g., Thagard, 1989).

We may also evaluate an explanation using other criteria, such as the credibility of the explainer or how well the explanation is articulated, that do not reflect the intrinsic value of an explanation. These criteria are useful in satisfying goals beyond identifying the information inherent to an explanation (Patterson, Operskalski, & Barbey, 2015). For instance, a well-articulated explanation can be useful for pedagogical reasons. The perceived credibility of an explainer may affect whether or not one believes the explanation, regardless of its intrinsic merit. While these criteria do not affect the inherent quality of an explanation, they may still serve important pragmatic functions and can be useful indicators of explanation quality.

Everyday Explanations

Philosophers have examined features central to scientific explanation that may improve our understanding. Does an explanation need to appeal to general laws (Hempel, 1965)? Should an explanation aim to unify the widest range of phenomena (Kitcher, 1989)? A common method in philosophical inquiry is to analyze existing scientific explanations: what types of explanations do scientists provide, and what makes them good or bad explanations?

However, many of the criteria used for evaluating explanations in a scientific context may differ from the criteria that are important for explaining everyday events. Explanations in non-scientific domains may require a different set of explanatory criteria because they are structured differently. For example, historical explanations are more likely to appeal to a narrative, and less likely to invoke general laws (Dray, 2000).

Although some philosophical theories suggest that abstract explanations are desirable (e.g., Strevens, 2007), people sometimes prefer explanations that are more

concrete (Bechlivanidis, Lagnado, Zemla & Sloman, under review) and less generalizable (Khemlani, Sussman, & Oppenheimer, 2011). Despite philosophical claims that explanations should be simple (e.g., Thagard, 1978), people tend to explain inconsistencies by positing additional causes rather than disputing a premise, resulting in a more complex causal structure (Khemlani & Johnson-Laird, 2011; Johnson-Laird, Girotto, & Legrenzi, 2004). Non-scientific explanations may also serve different explanatory goals. For instance, Newtonian mechanics is a source of good explanations for pedagogical and most practical purposes, even though Einstein's relativistic mechanics provides a more faithful explanation of how the world works.

In contrast to the philosophical literature on explanations, psychologists have tended to study short and simple explanations (e.g., Kelemen & Rosset, 2009; Weisberg, Keil, Goodstein, Rawson, & Gray, 2008; Cimpian & Salomon, 2014). These explanations have minimal causal structure, often only a single causal relation. Some experiments of this type rely on causal inference (e.g., Lombrozo, 2007; Khemlani et al., 2011); they ask participants to identify the cause or causes that best explain the observed effects, often holding constant the probability of an effect given its cause. We intend to test whether results obtained with these paradigms also apply to explanations that are more naturalistic.

Explanations that do consist of multiple causal relations require people to consider additional criteria, such as whether there are gaps in the causal structure (Keil, 2006). For example, it is undoubtedly true that leaves change color in Autumn because chlorophyll in the leaves breaks down. However this explanation omits parts of the causal model, such as why chlorophyll causes leaves to be green, and what causes chlorophyll to break down.

People may be sensitive to this omission, leading them to evaluate the explanation negatively even if they agree on the primary cause.

In more natural settings, we sometimes construct complex causal explanations in order to explain many pieces of evidence. Pennington and Hastie (1986, 1988) found that people explain complex events by constructing stories around the evidence, and that these stories can differ depending on the order that evidence is presented. These stories can be evaluated by how well they cohere with the available evidence (Byrne, 1995) using a set of coherence principles (Thagard, 1989). It is generally taken for granted that these principles are desirable, and subsequent work has provided some empirical support for these principles (Read & Marcus-Newhall, 1993; Schank & Ranney, 1992).

We should also consider how an explanation fits with our broader knowledge of the world. When evaluating a single explanation, we should consider possible alternative explanations (Fernbach, Darlow, & Sloman, 2010) and counterfactuals (Woodward & Hitchcock, 2003). When explanations provide evidence in support of a causal mechanism (Sloman, 2005), that evidence should be evaluated independently to determine whether it is credible and relevant (Kuhn, 1991).

Real-world explanations are typically more nuanced than experimental stimuli, and thus provide a more ecologically valid way of understanding the explanatory criteria people use to evaluate explanations. Experimental stimuli used to test explanatory criteria are often focused on a narrow subset of explanation types—for instance, explanations that explain token events or explanations that explain classes of events (types), but not both. Though many explanatory criteria have been established for evaluating scientific explanations, we test whether those same criteria are seen as virtues in everyday contexts.

In addition, evaluating explanations can require us to engage in a number of processes simultaneously, including dialectical reasoning (resolving inconsistencies), probabilistic reasoning (finding the most likely causes, or the causes that make the effect most likely), and didactic methods (educating the reader). We observe whether previously touted explanatory virtues endure in the face of these multiple goals.

Experiment 1

To investigate how people evaluate everyday explanations, we compiled a small corpus of explanations that were generated in a non-scientific and non-experimental context. Specifically, we gathered explanations from Reddit's *Explain Like I'm Five* (ELI5; www.reddit.com/r/explainlikeimfive), an Internet community that receives roughly 7 million unique visitors per month. The explanations in our corpus were rated by participants on a host of explanatory criteria that have been proposed in prior literature.

Method

Participants. 240 participants located in the United States were recruited using Amazon's Mechanical Turk (Paolacci, Chandler, & Ipeirotis, 2010). Five participants were removed from the data set prior to analyses for failing an attention check question[1] (Oppenheimer, Meyvis, & Davidenko, 2009). Of the remaining 235 participants, 131 were male and 104 were female, aged 18 to 69 (median age of 34).

Materials. Eight explananda[2] (see Table 1) were selected from ELI5 with three explanations for each, for a total of 24 explanations. The explananda were selected to fit

[1] Participants were shown a Likert scale with the prompt: "To ensure you are following instructions, please leave this question blank and hit next."
[2] Explanandum (pl. explananda) refers to the "thing being explained," or the subject of inquiry.

into one of four categories: historical, public health, legal, and social policy. These categories were chosen to contrast with scientific explanation and also reflect topics of interest to the general public. By selecting explanations from several categories, we sought to identify whether explanatory criteria are domain-general rather than apply only in certain domains. The explanations also varied in style, including a mixture of token and type explanations, as well as teleological and mechanistic explanations. We selected explananda from ELI5 that had a high level of engagement (i.e., many unique explanations and many "votes" from the site's users). For each explanandum, we chose three different explanations that proposed distinct mechanisms or offered different evidence in support of a given mechanism. In addition, the specific explanations were chosen because they varied *prima facie* on several explanatory criteria, such as appeals to expertise and evidence, complexity, and generality. An example explanation is shown in Table 2, and all of the explanations are provided in the Supplementary Material.

Procedure. Each participant was shown one explanandum with one corresponding explanation. After reading the explanation in full, participants assessed the *quality* of the explanation by rating whether the text constitutes a "good explanation." Afterwards, participants rated the remainder of the attributes (see Table 3) in a randomized order. To prevent participants from referring to their previous ratings, each attribute was rated on its own page, with two exceptions. *Generality* was rated on the same page as *principle consensus* because the latter question refers to the former. *Evidence credibility* and *evidence relevance* were rated last and on the same page, after participants were asked to highlight any evidence in the explanation. All attributes were rated using a 7-point Likert scale ranging from *Strongly Disagree* to *Strongly Agree*.

Table 1. Explananda used in Experiment 1

Explananda
• Why has the price of higher education skyrocketed in the US, and who is profiting from it?
• Why don't opponents of illegal immigration go after the employers who hire illegal immigrants?
• How can Malt-O-Meal blatantly rip off every brand-name cereal while Apple and Samsung have been in legal issues since the beginning of time?
• Why do the FBI and CIA use polygraph ("lie detector") tests on their employees, if polygraph tests are considered pseudoscience and so unreliable that US courts don't allow them as evidence?
• If Ebola is so difficult to transmit (direct contact with bodily fluids), how do trained medical professionals with modern safety equipment contract the disease?
• Why are so many people up in arms over "you have to have health insurance" initiatives, but are okay with mandated car insurance?
• Why is Ronald Reagan held to be one of the best US presidents, even among scandals like Iran-Contra that would have destroyed the reputation of other presidents?
• How has Switzerland managed to stay in a neutral position during times of conflict like WWII?

Table 2. An example explanation used in Experiment 1 to explain "If Ebola is so difficult to transmit (direct contact with bodily fluids), how do trained medical professionals with modern safety equipment contract the disease?

I am a biomedical scientist and part of the Ebola response team at a large and prestigious hospital on the east coast.

1) The most recent persons to get it is a doctors without borders doc. What people don't realize is that these doctors go into "battle" vastly under supplied in these foreign countries. They do not have Tyvek coveralls, respirators, gloves, and proper sterilization equipment. A lot of them because of supplies are forced to use the same pair of gloves on multiple patients for the day. Some don't use gloves at all.

2) Taking care of someone with Ebola is hell. There are literally body fluids everywhere. Imagine bloody decomposed fluid oozing out of every pore in your body, plus gallons of diarrhea and vomit. The protective equipment people are wearing here is good, but only if it stays intact and it doffed correctly. 90% of the infections occur because the person contaminates themselves when removing the soiled equipment.

In other words, taking the protective gear off improperly contaminates you, and 3rd world country doctors don't have the proper supplies.

Table 3. List of attributes rated in Experiment 1

Attribute	Statement
Alternatives	There are probably many other reasonable alternative explanations
Articulation	Regardless of accuracy, this explanation is well articulated
Complexity	This explanation is complex
Desired complexity	A good explanation of this issue is likely to be complex
Evidence credibility†	The facts and data are accurate
Evidence relevance†	This explanation is much better than it would be if the facts and data were not mentioned
Expert	This explanation refers to an expert
External coherence	This explanation fits with what I already know
Generality	This explanation appeals to a general principle (that is, a general rule that applies to many things)
Incompleteness	There are gaps in this explanation
Internal coherence	The parts of this explanation fit together coherently
Novelty	I learned something new from this explanation
Perceived expertise	This explanation was written by an expert in this topic
Perceived truth	I believe this explanation to be true
Possible explanation	This issue can be explained
Principle consensus†	Most people would agree with this general principle [following *generality*]
Prior knowledge	I knew a lot about this issue beforehand
Quality	This is a good explanation
Requires explanation	This issue requires an explanation
Scope	This explanation helps explain other issues besides what was asked
Visualization	It is easy to visualize what this explanation is saying

†These attributes were rated conditionally upon responses to other questions, and as such have missing values for some participants. *Evidence credibility* and *evidence relevance* were required only if the participant indicated that the explanation did appeal to some evidence. Only participants who indicated that an explanation appealed to a general principle (*generality*) were required to rate whether most people would agree with that general principle (*principle consensus*).

Table 4. Means and SDs for each attribute as well as partial correlations between each attribute and explanation quality, controlling for explanandum as a random effect. Adjusted p-values are computed using a full Bonferroni correction for multiple comparisons.

Attribute	M (sd)	R	p	Adjusted p
Alternatives	5.2 (1.4)	-.61	.001	.03
Articulation	5.3 (1.6)	.69	<.001	.007
Complexity	3.6 (1.8)	.49	.03	.66
Desired-complexity	5.1 (1.7)	.28	.20	>.99
Evidence-credibility	5.2 (1.3)	.08	.72	>.99
Evidence-relevance	3.8 (2.0)	.03	.91	>.99
Expert	3.0 (1.9)	.38	.07	>.99
External-coherence	4.5 (1.7)	.47	.02	.41
Generality	4.8 (1.5)	.58	.002	.06
Incompleteness	4.1 (1.8)	-.65	<.001	.01
Internal-coherence	5.4 (1.4)	.82	<.001	<.001
Novelty	4.6 (2.1)	.44	.03	.62
Perceived-expertise	3.7 (1.8)	.57	.004	.09
Perceived-truth	5.2 (1.6)	.90	<.001	<.001
Possible-explanation	5.8 (1.2)	.52	.009	.18
Principle-consensus	5.0 (1.3)	.71	<.001	.002
Prior-knowledge	3.6 (1.8)	.08	.71	>.99
Requires-explanation	5.2 (1.7)	.18	.47	>.99
Scope	4.3 (1.7)	.43	.04	.85
Visualization	5.7 (1.3)	.62	.009	.17

Results

Overview

We first examined the relation between explanation quality and each attribute without controlling for the other attributes. A mean score was computed for each attribute for each of the 24 explanations. Partial correlations were computed using a mixed effect model for each attribute, in each case treating *quality* as the dependent variable, one of the twenty remaining attributes as a fixed effect, and *explanandum* as a random effect. A partial correlation was used in place of a simple Pearson correlation because the 24 data points are not truly independent (there are three explanations for each explanandum). All

subsequent correlations reported for Experiment 1 reflect a partial R after controlling for explanandum as a random effect.

Fourteen of twenty attributes significantly predicted explanation quality, as shown in Table 4. Those that did not include: the *desired complexity* of an explanation, whether the evidence was credible (*evidence credibility*), whether the evidence was relevant (*evidence relevance*), whether the explanation referred to an *expert*, whether the participant had a lot of *prior knowledge* in the domain, and whether the explanandum required an explanation (*requires explanation*). To aid in interpretation, we also corrected for multiple comparisons using a full Bonferroni correction, though it is likely that this correction is overly conservative: all tests were planned *a priori*, and the tested hypotheses are often complementary rather than orthogonal. Nonetheless, six attributes survived the multiple comparison correction, suggesting that their relation with explanation quality may be particularly strong: whether the explanation had a number of possible *alternatives*, the *articulation* of the explanation, whether there were gaps in the explanation (*incompleteness*), whether the parts of the explanation fit together (*internal coherence*), whether the explanation was regarded as true (*perceived truth*), and whether most people agree with the general rule provided in the explanation (*principle consensus*).

Though many of the attributes were able to predict explanation quality, we also observed substantial covariance between the attributes. The attribute correlation matrix (Figure 1) depicts the magnitude of the correlation between all attributes pairwise, including explanation quality. It is likely that many of these attributes are not independent predictors of explanation quality, but instead reflect a smaller number of latent factors. For further discussion of how these attributes group together, see the Supplementary Material.

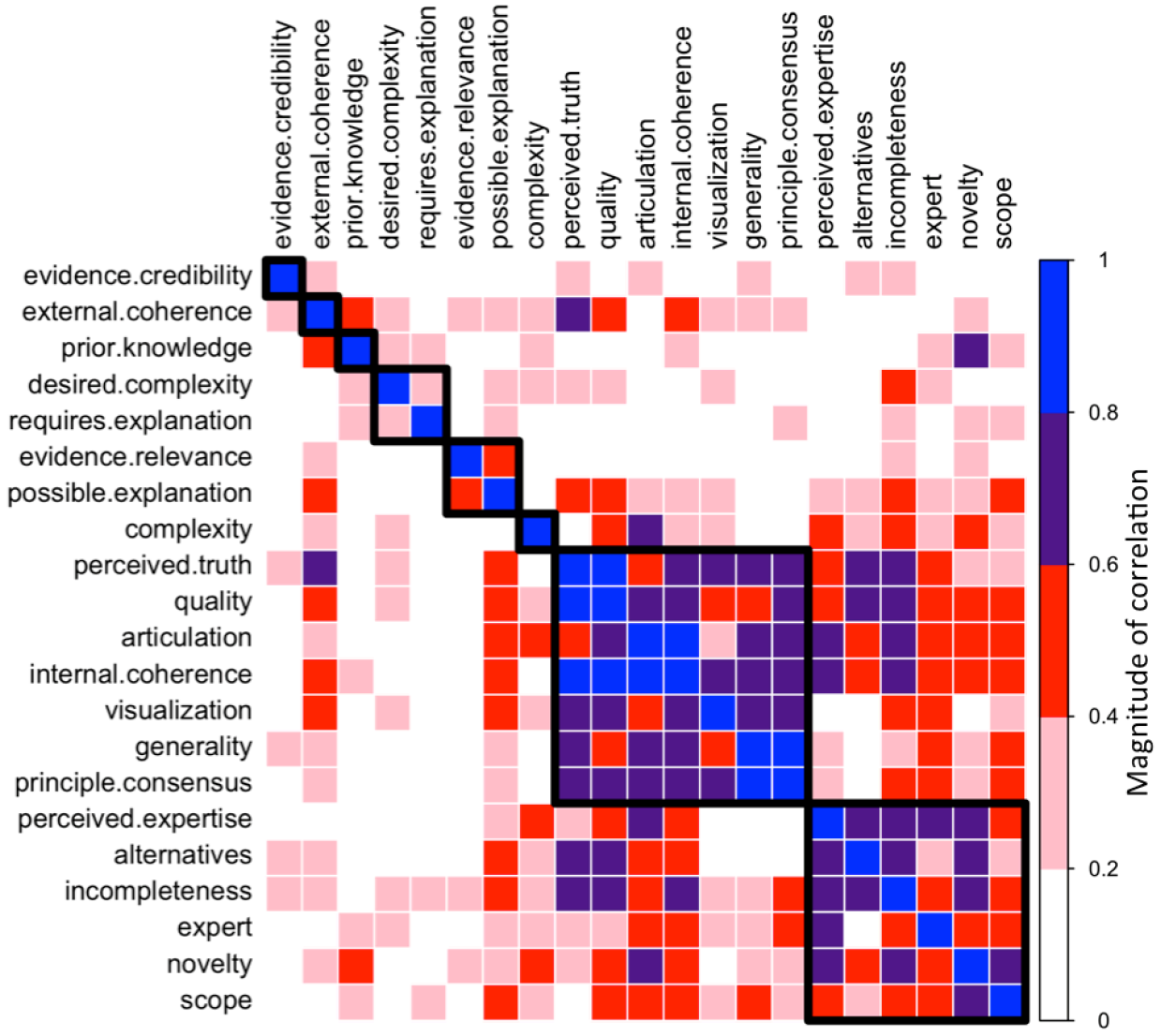

Figure 1. The magnitude (absolute value) of each pairwise correlation is shown after controlling for explanandum as a random effect. A hierarchical clustering procedure using average-linkage is shown to aid visualization.

Expertise

When evaluating an explanation, it can be helpful to assess the credibility of the explainer. Our knowledge about an individual can be used to predict what else that person is likely to know (Keil, Stein, Webb, Billings, & Rozenblit, 2008), which could play a role in judging whether the premises of an explanation are true. However it is not always clear what cues are used to judge expertise. For example, while we might expect experts to use

more technical language, using long words needlessly can make an author appear less intelligent (Oppenheimer, 2006). Similarly, scientific jargon does not always affect ratings of explanation quality (Weisberg**,** Taylor, & Hopkins, 2015; though see Eriksson, 2012).

Additionally, classifying the explainer as an expert can make the explanation more credible, but a good explanation does not have to be constructed by an expert. If two explanations are identical except for their source, they are presumably equivalent in their explanatory power even if they are not assigned the same degree of belief.

We assessed expertise using two dependent measures: whether the explanation referred to an expert (*expert*) and whether the participant believed the explanation was written by an expert (*perceived expertise*). The two factors are not identical, though they are related, R = .61, p = .002. An explanation can refer to an expert by self-identifying the explainer as an expert, or by citing an authoritative source. In contrast, someone may judge an explanation to be written by an expert through the quality of the language and level of technical sophistication. Both factors positively predict explanation quality, however *perceived expertise* is a stronger predictor (see Table 4). One possibility is that identifying an *expert* primarily serves to increase the *perceived expertise* of the explainer. A mediation model lends support to this hypothesis: although both *expert* and *perceived expertise* are positive predictors of *quality* (and each other), only *perceived expertise* is a significant predictor of quality when using multiple regression. Sobel's test (Sobel, 1982) confirms that *perceived expertise* mediates *expert* and *quality*, z(24) = 1.87, p = .06.

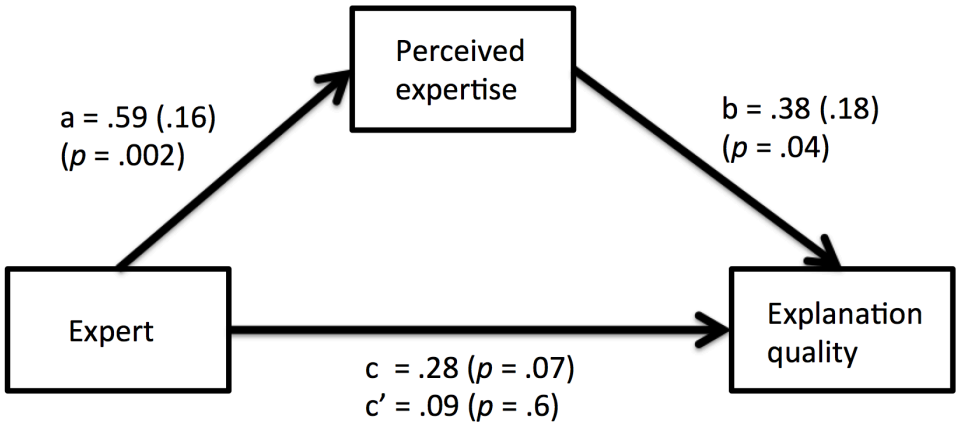

Figure 2. Perceived expertise mediates the relation between expert (reference to an expert or self-identification) and explanation quality.

Coherence

One of the most often cited explanatory virtues is *coherence*. Despite having received much attention in the literature, the term has been defined in several different contradictory ways. While some authors use coherence to refer to consistency with prior knowledge and beliefs (Murphy & Medin, 1985; Mackonis, 2013), other authors use it to refer to whether the components of an explanation are compatible or complement each other (Thagard, 1989; Bovens & Olsson, 2000; Keil, 2006).

We distinguish between internal and external coherence. External coherence refers to how much of the explanation overlaps or "fits" with what the reader already knows. Internal coherence refers to "how well the parts of the explanation fit together." We found that internal coherence is nearly twice as predictive as external coherence (R_{int} = .82, R_{ext} = .47, see Table 4). Using multiple regression, after accounting for internal coherence, external coherence did not significantly correlate with quality judgments (R_{int} = .73, p_{int} < .001, R_{ext} = .21, p_{ext} = .33). Previous research has suggested that people may not spontaneously generate or consider possible alternatives when evaluating an explanation

(Hirt & Markman, 1995). This failure to take an outside view when reasoning (Sloman & Lagnado, 2015) may explain why internal coherence takes precedence over fit with background knowledge.

Articulation

Despite providing no epistemic value, the articulation of an explanation was a strong predictor of perceived explanation quality (R = .79). We examined several linguistic markers to determine if surface features could explain perceived articulation and, by extension, predict explanation quality.

Articulation was correlated with a multitude of surface features, such as the number of words in an explanation (R = .64, p = .002), the median word frequency[3] in an explanation (R = -.54, p = .02), and the average word length (R = .45, p = .056). Perceived articulation also correlated with two related well-known readability metrics (Flesch, 1948; Kincaid, Fishburne, Rogers, & Chissom, 1975), Flesch-Reading Ease (R = -.54, p = .018), and Flesch-Kincaid Grade Level (R = .63, p = .003). Additionally, the proportion of nouns in an explanation predicted articulation (R = .54, p = .016).

Oddly, none of these metrics were significantly correlated with judgments of explanation quality (all R < .31, all p > .17), with the exception of word count (R = .60, p = .003). This finding is peculiar, given that articulation was highly correlated with explanation quality. As such, it is not entirely clear whether explanations are rated highly because they are articulate, or whether this correlation is the result of a third variable. For

[3] Word frequency was calculated using the English GigaWord corpus and excluding the top 1,000 most common words (most closed-class words).

instance, an intelligent person might be skilled at both writing and explaining (identifying the causal structure), even if one does not directly impact the other.

Simplicity

A guiding principle in explanatory reasoning is that of Occam's Razor: All things being equal, the simplest hypothesis should be preferred. Thus, we initially predicted a negative correlation between subjective complexity and explanation quality. Surprisingly, we observed a positive correlation, with explanations that were rated as more complex also rated as better explanations ($R = .49$, $p = .03$).

To further investigate this relationship, we examined other measures of complexity. Explanations may be deemed complex for many reasons, and it is not immediately clear what aspect of complexity our subjective measure is capturing. One possibility is that an explanation may be complex because it appeals to a large number of mechanisms. That is, the explanation suggests the explanandum occurred as a result of many causal pathways. Alternatively, an explanation may be complex because it is very detailed. An explainer may go into great detail about even a single mechanism. We test both of these hypotheses.

Causal pathways. One reason an explanation may be judged complex is because its underlying causal structure appeals to a large number of mechanisms. The four authors jointly identified the causal model for each of the 24 explanations in our corpus (see Figure 3 for an example; all of the causal models are provided in the Supplementary Material) by identifying causal language in the explanation (Sloman, 2005). Each node in a causal model represents a cause or effect, or both. Node labels were used as shorthand to represent the underlying cause or effect. A directed link from node A to node B indicates that A is a cause of B, though not necessarily a sufficient cause. Non-causal information, such as the

credibility of the speaker or flowery language, was not included in the causal model. Simple facts that are not causally related to the rest of the explanation were also excluded. Specific anecdotes and evidence used in the explanation were represented in the causal model by virtue of the fact that causal relations were distilled from more concrete examples. The causal models were constructed to be acyclic, consistent with Bayesian graphical models used elsewhere (e.g., Sloman, 2005). Though this process is somewhat subjective, we converged on a single causal model for each explanation.

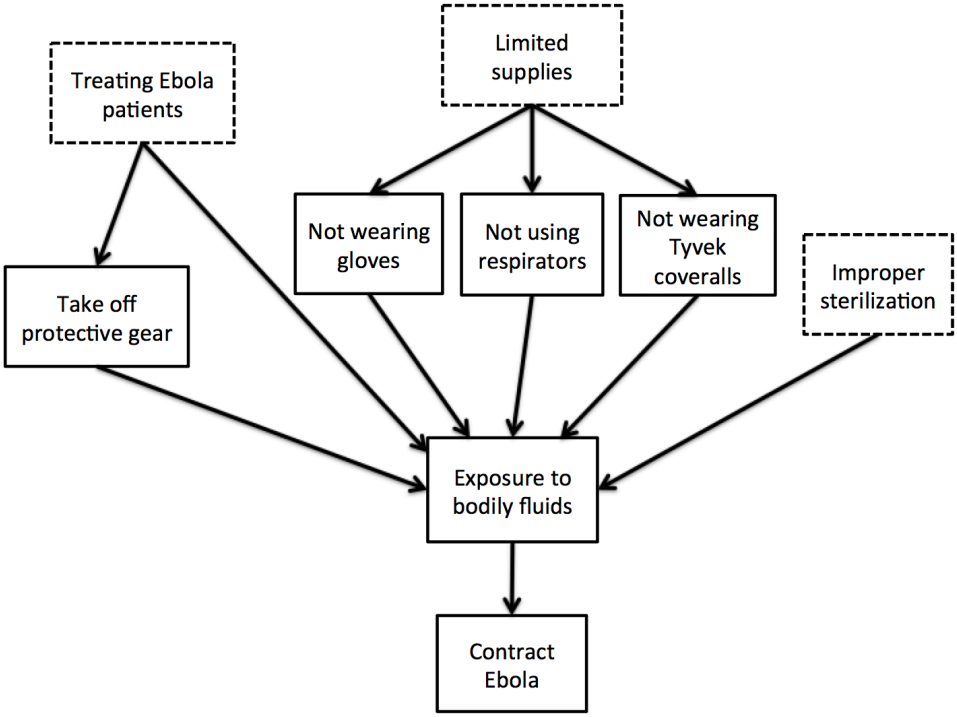

Figure 3. An example causal model reflecting the explanation shown in Table 2. The causal model shows three root causes (depicted with a dashed border) that are connected to the explanandum (contracting Ebola).

We estimated the number of causal mechanisms in an explanation by counting the number of root causes in the model (nodes without a parent node and connected by some pathway to the explanandum). This measure is consistent with previous research that suggests the number of *unexplained* causes, rather than the absolute number of causes, has

an impact on explanation judgments (Lombrozo & Vasilyeva, in press). As predicted, the number of root causes significantly predicts explanation quality, R = .64, p = .005. A reasonable objection might be that explanations that appeal to more causes also explain more effects. However, the correlation remains significant even when we control for the number of final effect nodes in the model and a subjective measure of how much the explanation explains (*scope*), R = .63, p = .015.

Explanation length. Another reason an explanation may be complex is because it contains a lot of details. One way to operationalize this is to simply count the number of words in an explanation. Those explanations that use more words to describe the causal system can be seen as more detailed. Indeed, we found that as the length of an explanation increased, so did its perceived quality, R = .60, p = .004. Furthermore, it appears that the number of causal mechanisms and explanation length are independent predictors of explanation quality, as both are significant predictors in a multiple regression, R_{words} = .52, (p = .02), R_{root_nodes} = .51 (p = .03). One caveat is that explanation length is also correlated with articulation, as reported earlier. When subjective articulation was included in the regression analysis, explanation length no longer predicted perceived quality, R_{words} = .26 (p = .27), R_{root_nodes} = .43 (p = .08), $R_{articulation}$ = .44 (p = .04), indicating shared variance between the three attributes.

Complexity and Expertise. Complexity may also have indirect effects on ratings of explanation quality. For example, it is possible that a complex explanation may make the explainer seem knowledgeable, which in turn increases the quality of an explanation. Explanation quality is significantly correlated with both perceived expertise and judgments of complexity (see Table 4). In addition, subjective complexity is strongly correlated with

perceived expertise, R = .55, p = .008. We tested the hypothesis that perceived expertise mediates the relationship between complexity and explanation quality. However Sobel's test for mediation (Figure 4) does not reach significance (z = 1.7, p = .088).

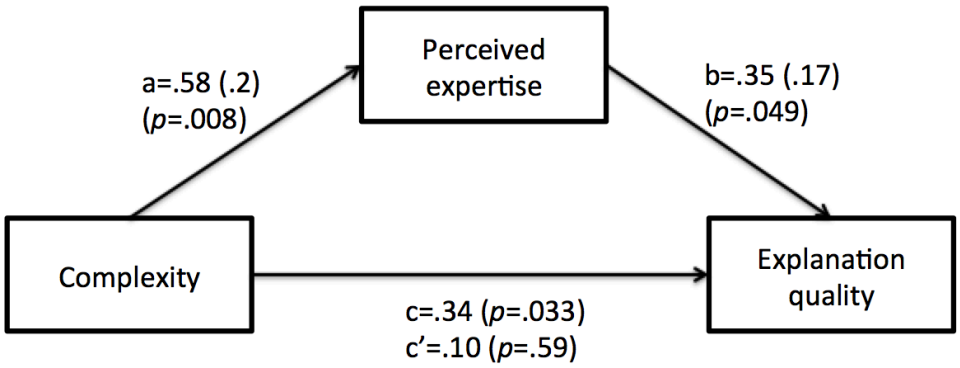

Figure 4. Perceived expertise is strongly related to both complexity and explanation quality, though it is not a significant mediator.

We also conducted a multiple regression analysis to see if perceived expertise could explain variance in explanation quality ratings independent of other measures of complexity (subjective complexity, explanation length, and number of root causes). Although subjective complexity is no longer significant in this analysis (see Table 5), the other predictors remain significant. This finding suggests that although all of the factors in Table 5 reflect measures of complexity, these factors are not interchangeable.

Table 5. Using multiple regression analysis, explanation length, number of root causes, and perceived expertise each significantly predict explanation quality ratings.

	Partial R	*p*
Explanation length	.50	.02
Root causes	.50	.046
Subjective complexity	-.33	.14
Perceived Expertise	.50	.02

Incompleteness

We expected that explanations containing gaps in the proposed causal mechanisms would be rated lower than explanations that did not contain any gaps. That is, if an

explanation suggests that A causes B, but it is not immediately clear *how* A causes B, participants will be sensitive to this omission. In support of this, we found that ratings of *incompleteness* (whether "there are gaps in the explanation") significantly correlated with explanation quality ($R = -.65$, $p < .001$).

We explored this further by examining the average path length in each of the causal models, measuring the average number of steps from a root cause to the explanandum. Pathways that contain more steps are likely to contain fewer gaps, and could be rated higher. However, this was not the case, $R = .08$, $p = .74$.

Discussion

These findings suggest that the explanatory criteria used to evaluate everyday explanations may differ from those previously identified. The biggest departure from existing theories is the finding that people prefer complex explanations—specifically, a preference for explanations that appeal to multiple causal mechanisms (though see Ahn & Bailenson, 1996). One limitation of the study, however, is its reliance on correlational analyses. In addition, by using naturalistic explanations that were not modified extensively, the explanations vary in many respects other than complexity. To address these concerns, we conducted an additional experiment that manipulated the number of mechanisms present in each explanation.

Experiment 2

We conducted a follow-up study using controlled stimuli to examine whether explanations with multiple independent causal pathways are preferred. We expected that people would prefer explanations that appeal to multiple causal mechanisms, even when a single mechanism is sufficient.

Method

Participants. 90 participants located in the United States participated in the experiment via Amazon's Mechanical Turk.

Materials. For each of the six explananda listed in Table 6, we created two explanations (denoted A and B) that appeal to entirely distinct mechanisms. Explanations were constructed from those found online using multiple online sources, including Reddit, Wikipedia, and HowThingsWork.com. For instance, one explanation for why China's population is rising despite their one-child policy is because ethnic minorities and rural populations are exempt from the rule; another explanation suggests that Chinese are living longer on average and that wealthy couples can afford to pay fines associated with violating the policy. Explanations were designed to be roughly equal in length, amount of detail, and number of mechanisms. We also created a third explanation that was simply a concatenation of the other two (denoted AB). This explanation encompasses the other two as it appeals to all of the causal mechanisms in A and B and does not vary in any other way. See the Supplementary Material for the full text of each explanation used in the study.

Procedure. Participants read an explanandum and were asked to make two ratings on a 7-point Likert scale: "How many reasons or mechanisms should a good answer to this question include?" and "How detailed should a good answer to this question be?" The response scale for both questions was ordinal, ranging from "1 – A good answer would offer only one reason or mechanism" to "7- A good answer would appeal to many reasons or mechanisms" for the former question, and from "1 – Very little detail is needed" to "7 – A lot of details are needed" for the latter question.

On the following page, participants read the full explanation and rated the number of mechanisms in the explanation (from "1 – Only a single mechanism" to "7 – A lot of mechanisms"), the amount of detail in the explanation (from "1 – Not detailed at all" to "7 – Very detailed"), the overall quality of the explanation (whether it was a "good" explanation to the question being asked, from "1 – Strongly disagree" to "7 – Strongly agree"), and whether the participant learned a lot from the explanation (from "1 – I didn't learn anything at all" to "7 – I learned a great deal").

This procedure was repeated for all six explananda. Each participant was shown only one of three explanations (A, B, or AB) for each explanandum. The explanations were counterbalanced so that each explanation was presented to exactly 30 participants. The order of the explanations was also counterbalanced. Prior to beginning the experiment, participants were shown an example explanation and informed of the types of questions they would be answering.

Table 6. List of explananda used in Experiment 2.

Explananda
• Why isn't China's population declining if they have had a one-child policy for 35 years?
• Why are cancer rates increasing?
• How did the Black Death in the 14th century come to an end?
• How do vaccines work?
• How do "speed limits enforced by aircraft" signs work?
• Before the invention of radio communication, how did a country at war communicate with their navy while they were out at sea?

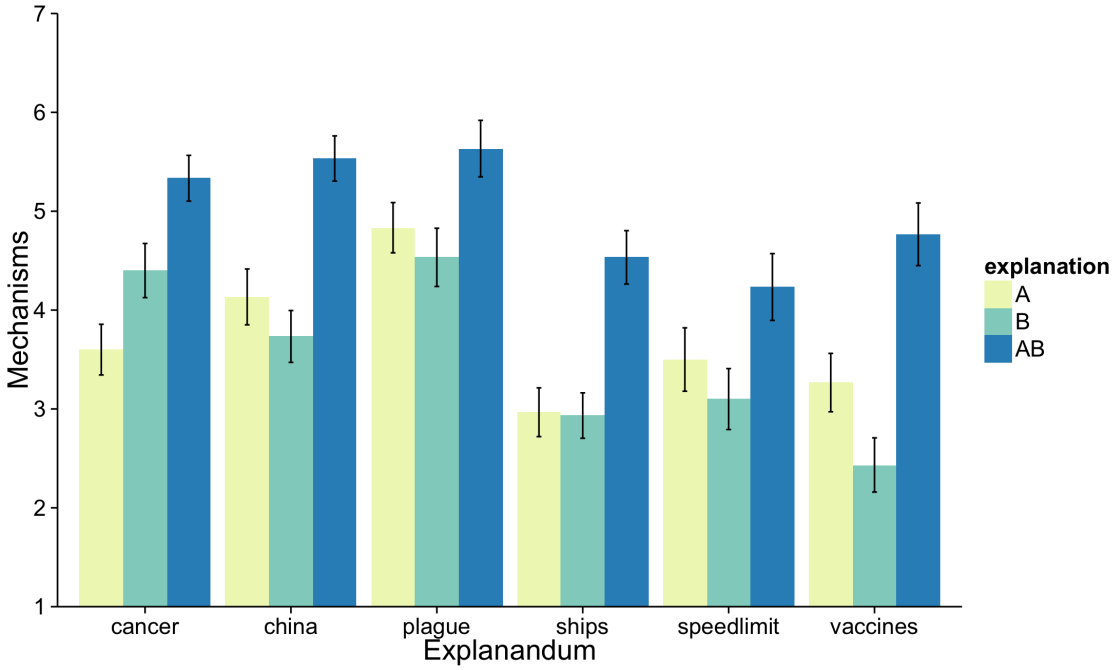

Figure 5. Participants were shown one explanation (A, B, or AB) for each of the six explananda. The AB explanations were rated as appealing to more mechanisms than the A and B explanations. The response scale ranged from "1 – Only a single mechanism" to "7 – A lot of mechanisms." Error bars denote standard error of the mean.

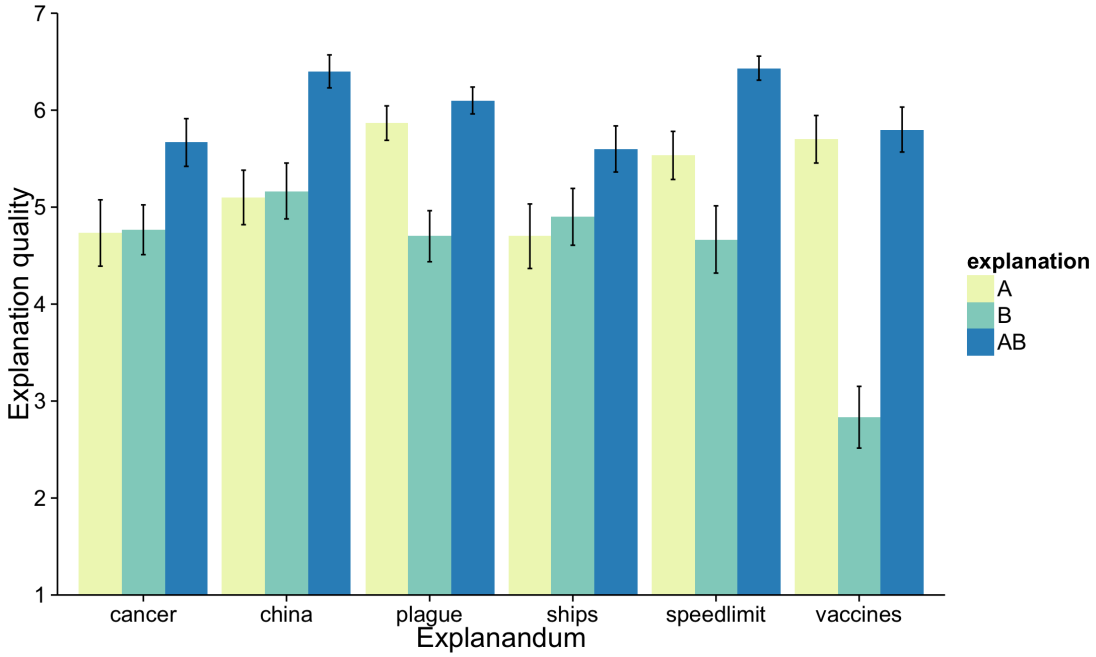

Figure 6. Participants were shown one explanation (A, B, or AB) for each of the six explananda. The AB explanations were rated as better explanations than the A and B explanations. Error bars denote standard error of the mean.

Results

For all six explananda, the concatenated AB explanation was rated as having more mechanisms than the A and B explanations (pooled), all $p < .05$. In five cases, the AB explanation was rated as having significantly more mechanisms than its nearest competitor (A or B; all $p < .05$ except $p_{ships} = .12$). See Figure 5. This validates our manipulation—explanations with more mechanisms were rated as such. Additionally, AB explanations were rated as having more details than their corresponding A and B explanations for all six explananda (pooled), as well as having more details than their nearest competitor (A or B), all $p < .001$.

In all six explananda, the quality of the AB explanation was rated significantly higher than the individual A and B explanations (pooled), all $p < .05$. See Figure 6. The concatenated explanation (AB) typically performed better than the second-most preferred explanation (A or B); significant in three explanations ($p < .05$), nearly significant in one ($p_{ships} = .07$) and not significant in two ($p_{plague} = .3$, $p_{vaccines} = .77$). All of the component explanations (A or B) except one were rated as above average in quality (above the midpoint). Despite this, participants showed a preference for the concatenated explanation (AB) in each case. In addition, 75 of 90 participants rated AB explanations higher on average compared to A or B explanations (across all six explananda), binomial test $p < .001$.

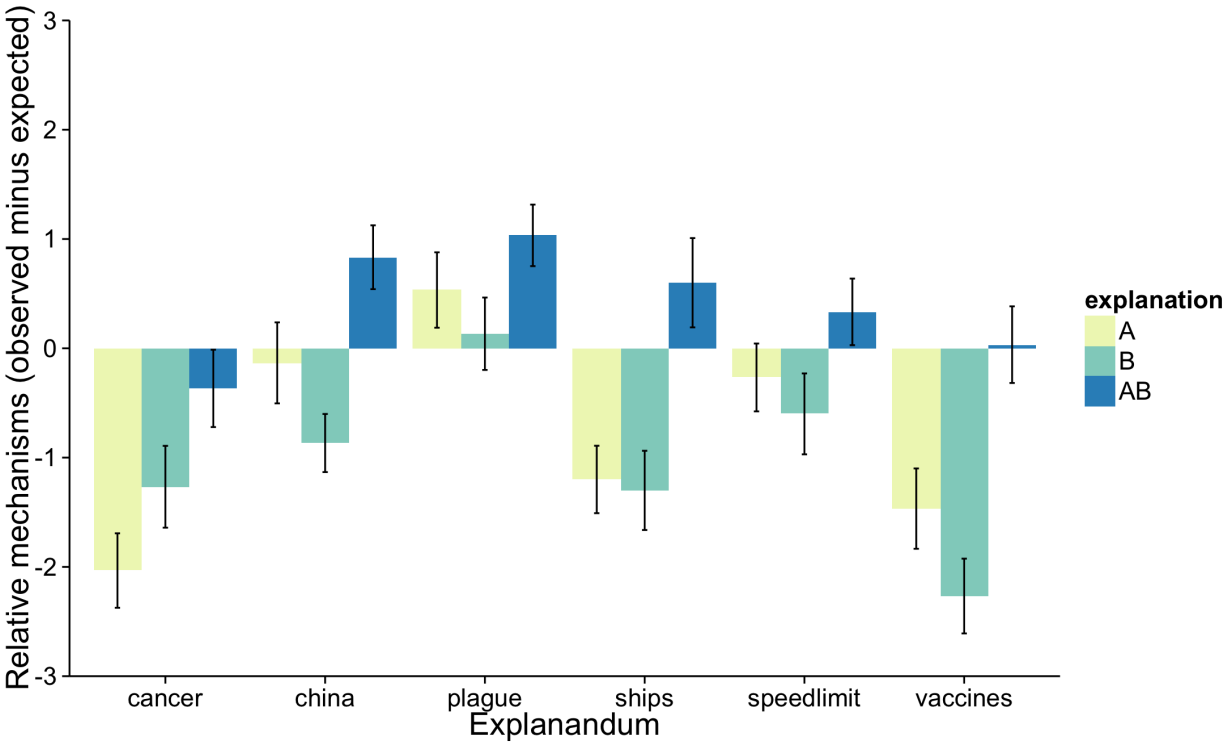

Figure 7. Participants were shown one explanation (A, B, or AB) for each of the six explananda. Prior to seeing an explanation, they rated how many mechanisms "a good explanation to the question should include," and after reading the explanation rated "how many mechanisms does this explanation appeal to". The y-axis denotes the latter minus the former. Error bars denote standard error of the mean.

Perhaps participants judged AB explanations as better because the phenomena to be explained were complicated, and thus benefited from an appeal to numerous mechanisms. Prior to reading each explanation, participants indicated that a good explanation *should* appeal to multiple mechanisms (mean ratings for the six explananda range from 3.8 to 5.7). We explored this possibility further by examining the relative complexity of each explanation: whether an explanation that appealed to more mechanisms than a "good explanation to the question should include" would still be preferred. We calculated a measure of relative complexity for each explanation by subtracting the mean rating of the number of mechanisms a good explanation *should* appeal to from the mean rating of the number of mechanisms the explanation *did* appeal to. As shown in Figure 7, the majority of

the AB explanations appealed to more mechanisms than initially expected from a good explanation.

Furthermore, this relative complexity measure predicted quality ratings as well or better than the ordinal rating of number of mechanisms alone. Multiple regression controlling for explanandum found a near-significant effect of relative complexity (partial R = .45, p = .07) but no significant effect of the number of mechanisms (partial R = .26, p = .31). However, we found no evidence that an explanation could suffer from being "too complex"; in no case was the AB explanation rated worse than either of its component explanations.

Discussion

Using controlled stimuli that differed only in the number of mechanisms, we replicated one of the key findings of Experiment 1, showing that people prefer explanations that appeal to multiple causal mechanisms. We found that the relative complexity (observed minus expected complexity) was a strong predictor of explanation quality, however even those explanations that were "too complex" were rated highly. These results contrast with previous findings that have shown people prefer explanations that appeal to the fewest number of causes (Lombrozo, 2007; Read & Marcus-Newhall, 1993).

One possible reason for the discrepancy is that our participants are driven by the desire to know that the explanandum is fully accounted for. In probabilistic terms, this is best represented as the likelihood of the explanandum given a set of causes. This interpretation is consistent with some previous findings (Pacer, Williams, Lombrozo, Xi, & Griffiths, 2013; Vasilyeva & Lombrozo, 2015). Providing additional causes typically increases the likelihood of the explanandum, making it seem inevitable. This hypothesis

leads to an interesting prediction that complex explanations should not be preferred after controlling for the likelihood of the explanandum. In contrast, experiments that use causal inference paradigms require participants to select the causes that best explain the observed data. For instance, Lombrozo (2007) uses a causal inference paradigm to provide support for the simplicity principle, but instructs participants that the effect *always* follows from a cause—thus, regardless of which explanation a participant prefers, the explanandum is inevitable. It is possible that causal inference paradigms overemphasize a particular explanatory goal: selecting the most likely cause.

Our data are inconsistent with an alternative hypothesis that participants should prefer explanations that increase the likelihood of a set of causes given the explanandum (Pearl, 1988). This hypothesis predicts a preference for simplicity, as the likelihood of a set of causes necessarily decreases (or remains the same) as additional causes are proposed. Moreover, an emphasis on predictive, as opposed to diagnostic reasoning, may lead people to endorse causes that are not true. For instance, one explanation suggested that China's population is growing in part because of a decline in the death rate, despite the fact that the death rate has actually increased slightly since the one-child policy took effect (The World Bank, 2015). Our results suggest that when participants are not judging whether a cause is true (as in causal inference paradigms), they may not evaluate the likelihood of the causes at all and instead assume them to be true (Fricker, 2002).

Another possibility is that people sometimes prefer simple explanations because natural phenomena are often the result of few causal mechanisms. Thomas Aquinas (1948) advocates for simplicity in his claim that "nature does not employ two instruments when one suffices." In this sense, a preference for simplicity could be a rational adaptation to the

environment. However, many phenomena have complex antecedents. Carruthers (2006) counters that biological explanations may not be simple because "one should expect biological systems to be messy and complicated, full of exaptations and smart kludges" (p. 151). Similarly, Salmon (2001) argues that the "desirability of simplicity seems to be an empirical question. In the social sciences, for example, it appears that simple hypotheses may be considered implausible because they are apt to be oversimplifications" (p. 129). Our results echo Carruthers and Salmon: even prior to reading an explanation, participants expected each explanandum to be explained through multiple causal mechanisms, and so single-cause explanations may have appeared too simplistic.

General Discussion

Everyday explanations differ from scientific explanations in their structure and in their goals. In Experiment 1, we evaluated potential explanatory criteria and found that some explanatory virtues, such as internal coherence, are good predictors of explanation quality even for everyday explanations. However other criteria such as articulation and perceived expertise are also highly predictive, even though they do not reflect the intrinsic quality of an explanation.

In two experiments, we find evidence that when evaluating explanations, people prefer explanations that are subjectively complex and appeal to multiple causal mechanisms. While this finding is at odds with claims that simpler explanations are usually better, it is consistent with prominent theories in philosophy that suggest the role of an explanation is to identify the causal network of events leading up to an event (Salmon, 1984; Strevens, 2008). The explanations used in the present experiments differ from those often used in the psychological literature because they attempt to be complete and, for the

most part, describe the causal mechanisms that led to an event. Rather than asking participants to ascribe causes to events or adjudicate between simple causes, participants were asked to evaluate the aptness of an entire causal system.

The ubiquity of explanations in our lives leads us to constantly evaluate potential causal mechanisms affecting the world. Deepening our understanding of what leads us to accept some explanations and reject others has implications for scientific communication, pubic policy, legal precedents, and beyond. Our current findings suggest that everyday explanatory practices are more complex and nuanced than previously thought.

Acknowledgements

This project/publication was made possible through the support of a grant from The Varieties of Understanding Project at Fordham University and The John Templeton Foundation. The opinions expressed in this publication are those of the author(s) and do not necessarily reflect the views of The Varieties of Understanding Project, Fordham University, or The John Templeton Foundation. The authors would like to thank Kethural Manokaran and Kenneth Peluso for their excellent research assistance, Uriel Cohen Priva for linguistic analysis suggestions, and members of the Sloman-Austerweil lab for helpful discussions. In addition, we thank Tania Lombrozo and Sunny Khemlani for their feedback on an earlier draft of this manuscript.

Supplementary Material

The full text of the explanations used in Experiments 1 and 2 are provided in Appendices A and B, respectively. The causal models used in analysis for Experiment 1 are provided in

Appendix C. Additional analyses referenced in the main text are provided in Appendix D. The raw data, experiment code, and reproducible analysis code are available online at https://osf.io/54gxq/. An interactive tool that can be used to explore the data from Experiment 1 is accessible at http://slomanlab.shinyapps.io/reddit/.

References

Ahn, W. K., & Bailenson, J. (1996). Causal attribution as a search for underlying mechanisms: An explanation of the conjunction fallacy and the discounting principle. *Cognitive Psychology*, *31*(1), 82-123.

Aquinas, T. (1945) *Basic Writings of St. Thomas Aquinas*, trans. A.C. Pegis, New York: Random House.

Bechlivanidis, C., Lagnado, D. A., Zemla, J. C., & Sloman, S. (under review). Concreteness and Abstraction in Everyday Explanation.

Bovens, L., & Olsson, E. J. (2000). Coherentism, reliability and Bayesian networks. *Mind*, *109*(436), 685-719.

Byrne, M. D. (1995). The convergence of explanatory coherence and the story model: A case study in juror decision. In J. D. Moore & J. F. Lehman (Eds.), *Proceedings of the Seventeenth Annual Meeting of the Cognitive Science Society,* (pp. 539-543). Mahwah, NJ: Lawrence Erlbaum.

Campos, D. G. (2011). On the distinction between Peirce's abduction and Lipton's Inference to the best explanation. *Synthese*, *180*(3), 419-442.

Carruthers, P. (2006). *The architecture of the mind*. New York, NY: Oxford University Press.

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*(05), 461-480.

Dray, W. H. (2000). Explanation in history. *Science, Explanation, and Rationality: Aspects of the Philosophy of Carl G. Hempel*, 217-42.

Eriksson, K. (2012). The nonsense math effect. *Judgment and Decision Making*, *7*(6), 746-749.

Fernbach, P. M., Darlow, A., & Sloman, S. A. (2010). Neglect of alternative causes in predictive but not diagnostic reasoning. *Psychological Science*, *21*(3), 329-336.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, *32*(3), 221-233.

Fricker, E. (2002). Trusting others in the sciences: *a priori* or empirical warrant? *Studies in History and Philosophy of Science Part A*, *33*(2), 373-383.

Glymour, C. (2014). Probability and the Explanatory Virtues. *The British Journal for the Philosophy of Science*, axt051, 1-14.

Harman, G. H. (1965). The inference to the best explanation. *The Philosophical Review*, 88-95.

Hempel, C. G. (1965). Inductive-statistical explanation. In *Aspects of scientific explanation* (pp. 381-403). New York, NY: Free Press.

Hirt, E. R., & Markman, K. D. (1995). Multiple explanation: A consider-an-alternative strategy for debiasing judgments. *Journal of Personality and Social Psychology*, *69*(6), 1069-1086.

Johnson-Laird, P. N., Girotto, V., & Legrenzi, P. (2004). Reasoning from inconsistency to consistency. *Psychological Review*, *111*(3), 640-661.

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

Keil, F. C., Stein, C., Webb, L., Billings, V. D., & Rozenblit, L. (2008). Discerning the division of cognitive labor: An emerging understanding of how knowledge is clustered in other minds. *Cognitive Science*, *32*(2), 259-300.

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

Khemlani, S. S., & Johnson-Laird, P. N. (2011). The need to explain. *The Quarterly Journal of Experimental Psychology*, *64*(11), 2276-2288.

Khemlani, S. S., Sussman, A. B., & Oppenheimer, D. M. (2011). Harry Potter and the sorcerer's scope: latent scope biases in explanatory reasoning. *Memory & Cognition*, *39*(3), 527-535.

Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas (automated readability index, fog count and Flesch reading ease formula) for Navy enlisted personnel* (No. RBR-8-75). Naval Technical Training Command Millington TN Research Branch.

Kintsch, W. & van Dijk, T. A. (1978). Towards a model of text comprehension and production. *Psychological Review*, *85*, 363-394.

Kitcher, P. (1989). Explanatory unification and the causal structure of the world. *Scientific Explanation*, *13*, 410-505.

Kuhn, D. (1991). *The skills of argument*. New York, NY: Cambridge University Press.

Lagnado, D. A., & Sloman, S. A. (2004). Inside and outside probability judgment. *Blackwell Handbook of Judgment and Decision Making*, 157-176.

Lipton, P. (2004). *Inference to the best explanation.* Oxford: Oxford University Press. Second edition.

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

Lombrozo, T. (2011). The instrumental value of explanations. *Philosophy Compass*, *6*(8), 539-551.

Lombrozo, T. & Vasilyeva. N. (in press). Causal explanation. In M. Waldmann (Ed.), Oxford Handbook of Causal Reasoning. Oxford, UK: Oxford University Press.

Mackonis, A. (2013). Inference to the best explanation, coherence and other explanatory virtues. *Synthese*, *190*(6), 975-995.

Minnameier, G. (2004). Peirce-suit of truth–why inference to the best explanation and abduction ought not to be confused. *Erkenntnis*, *60*(1), 75-105.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, *92*(3), 289-316.

Oppenheimer, D. M. (2006). Consequences of erudite vernacular utilized irrespective of necessity: Problems with using long words needlessly. *Applied Cognitive Psychology*, *20*(2), 139-156.

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, *45*(4), 867-872.

Pacer, M., Williams, J., Xi, C., Lombrozo, T., & Griffiths, T. L. (2013). Evaluating computational models of explanation using human judgments. *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*. arXiv:1309.6855 [cs.AI]

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, *5*(5), 411-419.

Patterson, R., Operskalski, J. T., & Barbey, A. K. (2015). Motivated explanation. *Frontiers in Human Neuroscience*, *9*, 1-15.

Pearl, J. (1988). Probabilistic reasoning in intelligent systems. San Francisco, CA: Morgan Kaufmann.

Pennington, N., & Hastie, R. (1986). Evidence evaluation in complex decision making. *Journal of Personality and Social Psychology*, *51*(2), 242-258.

Pennington, N., & Hastie, R. (1988). Explanation-based decision making: Effects of memory structure on judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(3), 521-533.

Read, S. J., & Marcus-Newhall, A. (1993). Explanatory coherence in social explanations: A parallel distributed processing account. *Journal of Personality and Social Psychology*, *65*(3), 429-447.

Salmon, W. C. (1984). Scientific explanation and the causal structure of the world. Princeton, NJ: Princeton University Press.

Salmon, W. C. (2001). Reflections of a bashful Bayesian: A reply to Peter Lipton. In *Explanation* (pp. 121-136). Springer Netherlands.

Schank, P., & Ranney, M. (1992). Assessing explanatory coherence: A new method for

integrating verbal data with models of on-line belief revision. In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society* (pp. 599-604). Hillsdale, NJ: Erlbaum.

Sloman, S. (2005). *Causal models: How people think about the world and its alternatives.* New York, NY: Oxford University Press.

Strevens, M. (2007). *Why explanations lie: Idealization in explanation*. *Unpublished Manuscript*. Retrieved from http://www.strevens.org/research/expln/Idealization.pdf

Strevens, M. (2008). *Depth: An account of scientific explanation*. Harvard University Press.

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

Thagard, P. (1989). Explanatory coherence. *Behavioral and Brain Sciences*, *12*, 435-502.

Vasilyeva, N., & Lombrozo, T. (2015) Explanations and causal judgments are differentially sensitive to covariation and mechanism information . In Noelle, D. C., Dale, R., Warlaumont, A. S., Yoshimi, J., Matlock, T., Jennings, C. D., & Maglio, P. P. (Eds.), *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*. (p. 2475-2480). Austin, TX: Cognitive Science Society.

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.

Weisberg, D. S., Taylor, J. C., & Hopkins, E. J. (2015). Deconstructing the seductive allure of neuroscience explanations. *Judgment and Decision Making*, *10*(5), 429-441.

Woodward, J. & Hitchcock, C. (2003). Explanatory generalizations, part I: A counterfactual account. *Noûs*, *37*(1), 1-24.

The World Bank, World Development Indicators (2015). *Death rate, crude (per 1,000 people)* [Data file]. Retrieved from http://data.worldbank.org/indicator/SP.DYN.CDRT.IN