A contrastive account of explanation generation

Seth Chin-Parker¹ · Alexandra Bradner²

Published online: 31 July 2017
© Psychonomic Society, Inc. 2017

Abstract In this article, we propose a contrastive account of explanation generation. Though researchers have long wrestled with the concepts of explanation and understanding, as well as with the procedures by which we might evaluate explanations, less attention has been paid to the initial generation stages of explanation. Before an explainer can answer a question, he or she must come to some understanding of the explanandum—what the question is asking—and of the explanatory form and content called for by the context. Here candidate explanations are constructed to respond to the particular interpretation of the question, which, according to the pragmatic approach to explanation, is constrained by a contrast class—a set of related but nonoccurring alternatives to the topic that emerge from the surrounding context and the explainer’s prior knowledge. In this article, we suggest that generating an explanation involves two operations: one that homes in on an interpretation of the question, and a second one that locates an answer. We review empirical work that supports this account, consider the implications of these contrastive processes, and identify areas for future study.

Keywords Explanation · Contrast class · Context

✉ Seth Chin-Parker
chinparkers@denison.edu

Alexandra Bradner
alexandrabradner@gmail.com

¹ Department of Psychology, Denison University, 100 West College Street, Granville, OH 43023, USA

² Department of Philosophy, Kenyon College, Ascension Hall, Gambier, OH 43022, USA

Philosophers have focused their studies of explanation on the necessary and sufficient criteria according to which a proposition counts as a good explanation, but psychologists have cast their net more broadly, examining, among other topics, why people engage in explanation, what implications explaining has for other cognitive activities, and what cognitive structures underlie explanation. This work has considered a number of interesting questions but has largely overlooked the process of generating explanations—that is, how explainers interpret a question and generate candidate answers. Researchers have recognized that explanation generation is a constructive process (Chi, 2000; Hale & Barsalou, 1995) that draws upon prior knowledge (Lombrozo, 2006) and relies on some form of structured mental representation (Chi, 2000; Keil, 2006). However, we lack a clear articulation of how these components are engaged during the generation of an explanation.

In this article, we propose a framework theory that draws heavily from prior pragmatic accounts of explanation (e.g., van Fraassen, 1980) and identifies two critical operations that play different functional roles in the generation of an explanation. According to this contrastive account, what makes an explanation explanatory is not that it has the correct content or type, but the fact that the explanation can distinguish the topic, what we want to explain, from a set of relevant but nonoccurring alternatives. In its current form, our account aims to highlight the critical junctures in processing that enable a cognitive agent to home in on a particular interpretation of a question and then produce a response. We hypothesize that two operations need to occur. The first involves gleaning the topic of the explanation—what the question is asking—and distinguishing that topic from other topic options noted explicitly in the question. The topic is then further specified in comparison to a class of relevant but nonoccurring alternatives—an implicit *contrast class*—indicated by the context in

which the question is asked. The second operation involves constructing the form and content of the explanation by considering how an answer might “grip” the topic at hand, but fail to engage the other members of the contrast class. Critical to this account is the notion that the “explanatoriness” of an explanation can be understood only in situ and only once a contrast class has been entertained.

Some notions of what is required to generate an explanation

Much of the current interest in explanation has stemmed from attempts to understand how explanation might be implicated in particular cognitive outcomes: How the explanation of a single instance can produce general knowledge, how explanation can increase learning in educational settings, and how explanation can affect our reasoning about the actions of another person. In exploring these topics, various lines of research¹ have illuminated particular aspects of, or processes involved in, generating an explanation.

Computer scientists have looked to explanation to motivate the development of formalized models of reasoning and judgment. One approach, dubbed *explanation-based learning* or *EBL*, features machine-learning programs that evaluate specific instances in terms of prior knowledge in order to produce useful generalizations (see Ellman, 1989, for a review). Although the processing varies depending on how a particular model is formalized, information about a target event is captured in a schema or other knowledge structure that can be generalized to new situations. Recent Bayesian models of explanation have taken a different approach. Within a particular causal system, these models generate a hypothesis, or explanation, for an outcome based upon the probabilities associated with unobserved but potentially explanatory variables (Pacer, Williams, Chen, Lombrozo, & Griffiths, 2013). In general, this work indicates that generating an explanation relies on access to prior knowledge and requires some way to structure that information in terms of what is to be explained.

Research examining the self-explanation effect (Chi, de Leeuw, Chiu, & LaVancher, 1994) has shown that positive learning outcomes are associated with generating explanations. In this line of research, the integration of new information with prior knowledge occurs during the construction of new, or the revision of existing, mental models (Chi, 2000). This focus on the role of structured knowledge representations in explanation reflects a longstanding claim in cognitive psychology that internal models underlie our abilities to

generate inferences, reason about complex systems to solve problems, and ultimately understand the world around us (Gentner & Stevens, 1983; Johnson-Laird, 1983).² The work examining self-explanation supports the ideas that prior knowledge plays a multifaceted role and that structured representations of that knowledge are referenced in the course of explanatory processing.

Research examining social reasoning also provides insight into explanatory processing. When someone speeds through a stop sign or holds the door open, we spontaneously generate an explanation in order to understand the cause of those behaviors. Theorists have hypothesized that complex causal structures underlie these attributions (e.g., Read, 1987), and that contrastive processing (Cheng & Novick, 1992; Hilton, 1990; McGill & Klein, 1993) plays a role in identifying relevant causal relationships. The possible cause is revealed by contrasting a particular case with alternate cases in which the effect does not occur: We assess whether X caused Y by comparing situations in which Y occurs to others in which Y is absent. If this contrast draws out X as the difference between the two situations, X is recognized as the cause for Y. Hilton (1990; Hilton & Erb, 1996) used this contrastive approach to develop an account of how people generate causal explanations. He suggests two processes are involved: *causal diagnosis*, a stage in which a probable causal connection is identified, and *causal explanation*, a stage in which the condition that does the explanatory work is selected by means of the contrastive processing. This work provides a rich set of processes that allow one to identify what causal information might be relevant when generating explanations.

In summary, prior research has posited the importance of constructive and comparative processes that rely heavily upon prior knowledge organized into useful mental structures. There is also evidence for distinct stages that address particular aspects of the emerging explanation. Although each of these approaches offers something to our account of how explanations are generated, we do not see a fully realized approach emerging from any one of them. In the formal modeling approach, little consideration is given to how the topic of the explanation is determined. In the self-explanation research, the focus is on abstract knowledge structures, so there is little grounding in the situations or contexts in which the explanations are being generated. The causal attribution literature does not make room for the plurality of explanatory forms that exist. We hypothesize that the lack of a fully realized approach may be the result of considering generation only retrospectively, that is, by hypothesizing about what might lead to particular explanatory outcomes. For instance, Colombo (2017) recognizes that multiple forms of explanation are possible and then hypothesizes that a “preliminary

¹ Note that this is not an exhaustive accounting of all research that has examined some aspect of the generation of explanation. Instead, it represents important lines of inquiry that illustrate how the field has approached the topic. Throughout the article, we will engage with other important studies of explanation as they relate to our proposal.

² Interestingly, the notion of a mental model was introduced within an account of explanation (Craik, 1967).

step” in generating an explanation involves identifying the kind of explanation the individual intends to offer. Although we agree that we must account for explanatory pluralism, we do not believe that explainers begin with a sense of the kind of explanation they would like to offer and then go searching for answers that satisfy that form of explanation. Instead, we believe a richer story can be told about how the processes involved in generating an explanation unfold. Our goal is to provide a useful framework that allows us to consider how these various pieces of the puzzle fit together.

The contrastive approach to the generation of explanations

According to our view, the process of responding to a prompt for explanation requires two contextually constrained operations that focus the explainer by steering her away from a dizzying array of options. These operations employ such cognitive processes as memory retrieval, selective attention, and reasoning, but are not fully reducible to these processes outside of the context. We can only understand these operations by attending to the way in which they read from and integrate the background knowledge of the explainer and the features of the situation at hand. Like van Fraassen (1980) and Hilton (1990), we recognize the pragmatic constraints on explanation, and like Hilton (1990, p. 78), we recognize that intrapersonal explanations—explanations initiated by oneself instead of by another person—will operate within the same framework.

Consider a parent making pizza with a child who asks: “Why does cutting onions make you cry?” The child recognizes that cutting an onion does, in fact, make the parent cry: That is the explanandum. The request asks the parent to provide some utterance that specifies why cutting an onion makes the cook cry. The parent has a rich web of knowledge about cooking and onions and the world that will have to be navigated in constructing an appropriate response to the inquiry in this situation. We hypothesize that two operations are involved in generating such an explanation.

Operation I: The context focuses the explainer on a particular explanandum

In the question posed by the child, there are at least four explananda on which an explainer might focus:

- (1) What is it about cutting *onions* such that you cry when working with them?
- (2) What is it about *cutting* onions such that you cry during that activity?
- (3) What is it about *you* such that you cry when cutting onions?

- (4) What is it about *crying* such that you do it when cutting onions?

To be explanatory, the explanation has to address the appropriate topic. We propose that selection of the topic is highly constrained, but not determined, by situational factors. As is developed in Sperber and Wilson’s (1986) relevance theory, an interaction between the individual’s context (including prior knowledge, goals, currently activated information) and the situation-specific input occurs in such a way that it privileges the information that will have the greatest effect on the current processing. If the parent and child are both focused on the onions after working with a number of other pizza ingredients—that is, if onions are temporally accessible and salient items—Topic 1 should stand out as the relevant topic of the question. Instead, if the parent has been recently helping the child to understand various bodily functions—for example, crying—the presence of that goal would alter the context and the focus might shift to Topic 4. For the purposes of our discussion, we assume that the first option, *onions*, is identified as the relevant topic. We propose that without contextual cues the generation of an explanation cannot get off the ground. Yet, contextual constraints receive only glancing consideration in existing accounts of explanation generation. For instance, Cimpian (2015) suggests that when asked “why so many children’s menus at restaurants have macaroni and cheese as an option,” we immediately zero in on what he identifies as the “main constituents” of the prompt: children and macaroni and cheese. However, this fails to address how menus and restaurants are removed from consideration as topics for the explanation. The account that he offers is insightful and draws out many interesting ideas about how access to particular information in memory plays a role in determining what goes into the explanation. However, we argue that it does not fully acknowledge the importance of context when identifying the topic of the explanandum. Without that initial accomplishment, explanatory processing cannot proceed.

In the pizza case, the topic is further specified in terms of possible but nonoccurring alternatives that are identified via contrastive processing. The points of emphasis in the questions above prompt different alternative cases: In Topic 1, “What is it about cutting *onions* (and not *w*) that makes the parent cry?” In Topic 2, what is it about *cutting* (and not *x*) such that performing that action on the onions makes the parent cry? In Topic 3, “What is it about *the parent* (and not *y*) such that she cries when cutting onions?” And in Topic 4, “What is it about *crying* (and not *z*) such the parent does it when cutting onions?” As was noted above, we focus on Topic 1. There are many ways to interpret this question, because there are an infinite number of ways for something not to be an onion. Generating an explanation is only possible if the explainer can limit—quite drastically—those options. This

quandary, of how to pinpoint a topic, has been directly addressed by philosophers through the introduction of contrast classes. For decades, epistemologists have debated the role that relevant alternatives play in the analysis of knowledge (Dretske, 1981; Goldman, 1976; Schaffer, 2008); Garfinkel (1990) and van Fraassen (1980) have recognized the role of contrast in social and scientific explanation; and Sinnott-Armstrong (2008) has defended contrastivism as a constitutive aspect of reasoning in any form. All of these accounts capitalize upon a basic insight: Situation-specific contrasts enable us to specify what is relevant about a topic. We hypothesize that a person who could *not* perform these culling (i.e., highlighting and eliminating) operations would find herself cognitively stalled, unable to understand a question and unable to build any traction regarding its answer.

Continuing on, we have to specify the topic of our explanation, *onions*. Beyond being simply an onion, we can think of that object in a number of ways: vegetable, projectile, savory ingredient, and so forth, each of which prompts a particular assemblage of prior knowledge (Ross & Murphy, 1999; Shafto, Kemp, Mansinghka, & Tenenbaum, 2011). In the context of making a pizza, the question asks not about onions as projectiles or layered things, but as pizza ingredients: “Why does cutting onions—and not other pizza ingredients—make cooks cry?” Access to knowledge related to a particular class membership is impacted by contextual factors (Macrae, Bodenhausen, & Milne, 1995), although the inherent qualities of the related knowledge can play a role (Cimpian, 2015; Patalano, Chin-Parker, & Ross, 2006). Since the organization of conceptual knowledge reflects potential contrasts (Davis & Love, 2010; Goldstone, 1996; Verheyen, De Deyne, Dry, & Storms, 2011; Voorspoels, Storms, & Vanpaemel, 2012), access to a particular sense of onion would activate appropriate candidates for the contrast class. In the pizza-making case, alternatives from the list of possible pizza ingredients (e.g., garlic, peppers, tomatoes, cheese, pineapple, and sausage) serve as candidates for the contrast class. All of the pizza ingredients regularly associated with pizza in that household are available, but the situation at hand will serve as a critical constraint on this list of possible candidates. For example, if the child were in kindergarten, the context might point to another vegetable that goes on pizza—for example, a tomato—as the relevant contrast to the onion. However, if the child were home from college, that comparison might be too coarse. Instead, the contrast might shift to something that shared with the onion not only the quality of being a pizza ingredient, but also the quality of being smelly—for example, garlic. Each of these potential contrasts would bring forth different arrays of knowledge, which would, in turn, impact the subsequent explanation. In the first case, the onion’s intense smell might become the focus of the explanation, because the intensity of the smell is what differentiates onions from tomatoes. In the second case, garlic also has an intense smell, so the explanation

might focus on some deeper difference between onions and garlic. This could require access to knowledge that associates smell with chemicals, some of which could be irritants (but we do not want to get ahead of ourselves in this account).

By focusing the explainer on a particular contrast class, the local context targets a specific difference between members of some class. This reveals to the explainer the topic of the question and the particular puzzle that the answer will need to address in order to serve as explanatory. At the end of the first operation, the explainer has identified *onions* as the topic of the explanation, and that topic has been specified further by its contrast class. The knowledge drawn out by the *onions (not tomatoes)* contrast is available at the start of the second operation.

Operation II: The explainer figures out how to relate an answer to the explanandum, but not to the other members of the contrast class

Despite the processing that has taken place in the first operation, the parent in our example still cannot respond. A particular point of inquiry has been identified—what is it about cutting onions, versus tomatoes, such that you cry when working with onions—but a potential answer must now be located. According to our account, the form and content of the explanation emerge as the result of two closely aligned processes. First, the conceptual knowledge activated within the first operation is captured within a structured knowledge representation—for instance, an internal model (see Chaigneau, Barsalou, & Sloman, 2004; Keil, 2006)—that contains both relevant concepts and their relations. Second, the comparison of this internal model to the explanandum highlights particular information that establishes the form and content of the explanation. In our example, knowledge associated with pizza, knives, cutting, onions, and tomatoes, among other relevant things, is organized and compared to the explanandum. Out of the comparison between the onion and the tomato, the smell is noted as an inconsistency—a point at which the cutting of the onion deviates from what would be expected, given what is known about cutting tomatoes. Whether it represents some unexpected element in the explanandum (Weiner, 1985), an abnormal state (Hilton & Slugoski, 1986), or failure of expectations or norms (Hitchcock & Knobe, 2009; Schank, 1982), the highlighted discrepancy reveals a space that becomes the focus of the explanatory processing. To explain, the answer must connect up with that point in the explanandum, but not do so with what has been revealed about the contrast class members. As we noted above, if tomato provides the contrast, an answer about the smell released by cutting the onion will respond to that space in the explanandum, but not connect up with what occurs when cutting tomatoes. However, that response is insufficient when garlic is the contrast, because an answer about smell connects up to *both* the explanandum and the contrast. Although it is beyond the scope of this account to

fully articulate the processes that act on the internal model during this operation, we hypothesize that this work reflects some form of alignment processing (e.g., Gentner & Markman, 1997) or other type of relational reasoning (Dumas, Alexander, & Grossnickle, 2013; Holyoak, 2012; Hummel, Licato, & Bringsjord, 2014). As this process unfolds, the content of the explanation—the knowledge associated with onion smell and smelling—is being assembled, but the issue of what form the explanation will take remains.

Although much of the work related to explanation generation has focused specifically on causal explanations, multiple forms of explanation are available (Colombo, 2017; Lombrozo, 2012; Saatsi & Pexton, 2013; Van Bouwel & Weber, 2008; although see Skow, 2014). The form provides the way in which the answer relates to explanandum, for example, as a mechanism that causes the crying, as an adaptation that contributes to the onion's survival, or as an essential feature of the onion. A causal relation might identify a chemical eye irritant in the onion that results in crying. A functional relation might view that same chemical as a defense mechanism that serves the purpose of deterring onion killers. And a principled relation (Prasada & Dillingham, 2009) might identify a quality thought to be essential to the very being of the onion. We propose that the explainer settles upon a form for the explanation by running through the internal model to locate (a) a relation that holds between the answer and the topic, but not the answer and the contrast class, and (b) a relation that is not discouraged by the situation at hand.

It is our view that though several relations might satisfy relation (a), the explainer's experience of the actual situation in which the question is asked serves to narrow that possibility space. In the onion case, on the basis of the parent's vast store of prior knowledge, the parent *could* respond to the college student's query by saying that onions have a defense mechanism that heads of garlic do not.³ Nothing in the explainer's prior knowledge alone eliminates the functional form of explanation as a possibility. But, in this context, cutting the onion is not viewed as a threatening behavior. On the contrary, cutting the onion moves the pizza making along. The context speaks against the appropriateness of the functional explanation: that style of explanation would confuse, instead of satisfy the college student, who is looking for the chemical source of the irritation. We could imagine a context in which the functional explanation would be appropriate. For instance, if the parent and the kindergartener were being silly, the parent might explain that the onion makes her cry, and the tomato doesn't, as a way to exact justice for the knife injury. In this context, the causal explanation that details a chemical mechanism would kill the mood. It would be inappropriate. In both the kindergartener and the college student cases, prior knowledge gives the explainer a possibility space, but the actual

facts of the local situation help the explainer to pick and choose among those possibilities. The possibility space is so permissive that the explainer *must* mine the context to identify the appropriate form.

Evidence for the contrastive account

We have all experienced situations in which we responded to a question with an inappropriate answer: I (Seth) once had a student ask why he hadn't received any points for an answer on an exam. Without missing a beat, I launched into a 5-min explanation of my exam grading, detailing which responses warranted full credit, partial credit, or no credit. But the student was simply pointing out that I had failed to write down the points he earned along with my other feedback. A more controlled illustration of contextually based shifts was provided by Hilton and Erb (1996). Participants were told that a watch face broke when hit with a hammer and then asked to rate how relevant the hammer hitting the watch was to this occurring. Without any specific context, the relevance of the hammer to the explanation was rated as quite high. However, when participants were informed that the event occurred during "a testing procedure in the factory," they rated the relevance of the hammer lower, likely because another feature—for example, the quality of the watch face—became relevant in that situation. Interestingly, the shift in context impacted the ratings of the relevance of the hammer in the explanation, but not the rating of whether it was true to say the hammer caused the break. This suggests the change in context affects the perceived topic of the explanation, whereas the causal relations at play remain stable. Other studies have shown that additional situational factors—for example, the goal of the explainer (Hale & Barsalou, 1995) or the wording of the request (Rips & Edwards, 2013)—can affect the identification of the topic.

A recent study of category-based explanation provides evidence for the role of the contrast class in explanation generation. Building off the work of Williams and Lombrozo (2010), Chin-Parker and Cantelon (2016) asked participants to explain the category membership of a target set of robots, Deegers (see Fig. 1). Those explanations were intermixed with prompts to explain the category membership of a second set of robots, either Lokads or Koozles. Although the prompt to explain the category membership focused on a single category at a time, "Why is this robot a Deeger type?," the presence of the alternate category led the participant to define the request as, "Why is this robot a Deeger (not Koozle/Lokad)?" All the participants focused on the variation of the features as the topic of their explanations, but the contrast available defined which features were drawn out as informative. The participants were sensitive to a

³ We thank our reviewer for the opportunity to address this potential objection.

Fig. 1 Robot categories from Chin-Parker and Cantelon (2016).

property of the features that was bound to the contrast class—that is, whether the features differentiated the categories—as opposed to something inherent in the category itself. The specification of the contrast class is also influenced by the prior knowledge activated in a given situation. Williams and Lombrozo (2013) asked participants to explain the category membership of items that had either knowledge-laden labels—for example, indoor robot or outdoor robot—or generic labels—for example, Glorp robot or Drent robot. When explaining the category membership of indoor/outdoor robots, the participants tended to focus on features that might be relevant to where the robot operated—that is, the foot shape. When the category labels were uninformative, the participants tended to focus on the antenna length, which was a simpler and less variable feature of the robots. In both conditions, the contrastive processing helped to draw out particular features for consideration, but the specification of the contrast class varied depending on the context.

The pattern of results found in Chin-Parker and Bradner (2010) supports the notion that context plays a role in both Operation I and Operation II. Participants were asked to generate an explanation of an animated clip (see Fig. 2 for a screenshot of the end of the animation). The clip showed a

Fig. 2 Final scene of the target clip from Chin-Parker and Bradner (2010). After watching this clip, participants were asked, “Please provide a possible explanation for what you just saw in the previous clip.”

form, the squiggle figure on the left, pressing on a lever that initiated a series of events that resulted in a liquid being deposited on the ground. The participants had watched an initial set of animated clips before explaining the target clip. Participants in the *nonsystematic condition* watched a set of animated clips in which a form pressed on a lever, initiating a series of events resulting in material being dumped on the ground. Participants in the *systematic condition* watched a set of animated clips in which the same events occurred, but the material released was returned to the form that had initiated the action. Figure 3a and b provide screenshots from an example of each type of clip. All participants received the same prompt after viewing the target clip, but the explanations they generated differed according to how the initial set of clips they had seen contrasted with the target clip.

The general experimental context was the same for all participants, so they all focused on the events in the clip as the topic for their explanations. However, as the initial clips offered different contrast, the specification of that topic differed between the groups. For a participant in the nonsystematic condition, the levers and connections in the initial clips differed from the target clip, so the contrast highlighted the mechanisms involved in releasing the material. The focus on these mechanisms highlighted causal relations, so participants in that condition tended to generate causal explanations. In the systematic condition, the food spilled onto the ground in the target clip instead of returning back to the form, so the contrast highlighted what could be considered a disruption of the system. In this situation, the focus on the purpose of the events highlighted functional relations, the role the mechanisms played in a system, and participants in that condition tended to generate functional explanations. This pattern of results fits well with the findings of Hale and Barsalou (1995). Participants focused on a comprehensive system tended to produce functional explanations. Participants focused on the events occurring within that system (and not into the broader system in which those events occurred) tended to produce causal, or mechanistic, explanations. In that study, a shift in the initial context, achieved by altering the goal of the explainer, accounted for this variation in the explanatory form. In Chin-Parker and Bradner (2010), the shift in the contrast class affected the explanatory form. In both cases, the form of explanation was tied to constraints on the processing that occurred as the topic of the explanation was being defined.

The virtue of our account of explanation generation is that it points to *how and why* explainers employ different forms of explanation in different circumstances. We consider there to be strong evidence for the role that the context plays in the generation of explanations. We also think that a growing body of evidence suggests that contrastive processes underlie the specification of the

Fig. 3 Examples of each type of initial clips from Chin-Parker and Bradner (2010). Panel a illustrates a nonsystematic clip, whereas panel b illustrates a systematic clip. The events that occurred within each clip

were identical, except for where the material fell—either on the ground or back into the shape that had initiated the actions.

explanandum. However, we also recognize that further study will be required so we can better understand the specific role that these constraints are playing in each of the operations we propose.

Concluding comments on the contrastive account

We argue that elements of the context—for instance, immediate goals and objects at hand work progressively to home in on some appropriate sense of the explanandum. Once the question has been specified, the processing turns to the selection of both an answer and an explanatory form that satisfies the questioner. During this stage, the explainer uses contextual information to narrow down the multiple possible explanatory relations that might be available through the existing internal model of the situation.

Several criteria for what guides the evaluation of explanations have emerged, such as simplicity (Lombrozo, 2007), coherence (Thagard, 2006), and elaboration (Rottman & Keil, 2011). These proposals identify qualities that an explanation should have in order to count as a good explanation (or the best one), but do not address the process through which generation unfolds. We have offered an account of explanation generation, not evaluation, so it is important to note that there is no guarantee that the explanations generated according to this process will be the best ones. Despite the naturalness of explanation, our explanations often fall short of being ideal (Keil, 2006; Wilson & Keil, 2000). There are several ways in which an explanation might be compromised: Situational factors can distract or point to irrelevant topics, inadequate or false prior knowledge can negatively impact the ability to establish an appropriate contrast class, and mistakes about the needs of the questioner can block access to the relevant features of the internal model. One virtue of our account is that it can explain how and why such problems arise.

Discussion

Implications of the contrastive account

Constraint Pragmatic theorists of explanation are sometimes derided by philosophers for offering only hand-waving accounts of explanatory relevance. For a pragmatist, it's more important that an answer be relevant than true. But when asked to specify a general account of the concept of relevance, the pragmatist fails to respond, saying only that it will become clear in context which answers are relevant. As we described in the prior sections, empirical study has identified some of the constraints that play a role in establishing relevance during the generation of an explanation, but we also recognize that much work is still to be done in this regard. Our own work examining explanation (Chin-Parker & Bradner, 2010; Chin-Parker & Cantelon, 2016) has only started to identify some of the constraints in place at particular points in the processing. When determining the topic of the explanation, the context activates some of the explainer's prior knowledge to impose particular constraints on the question's interpretation via the contrast class. It is less clear, though, how the context and an internal model might operate in concert to determine the form of explanation. We think explanation can be considered something like similarity, a comparably slippery construct that is constrained by relevant processing. Instead of searching for *a priori* criteria to define similarity, the local process of comparison allows a sense of similarity to emerge (Medin, Goldstone, & Gentner, 1993). The contrastive view proposed here is similar in this regard—sensitivity to the context and contrastive processing during the generation of an explanation allows the explainer to identify candidate answers. So, instead of pointing to criteria that will establish explanatory relevance *a priori*, our account emphasizes the need to better understand the constituent processes and their roles within this explanatory framework.

The constraints that we have identified in this account will be useful to consider as computational approaches to explanation are developed. The plausibility

of a computational model of explanation has increased in the past decade as theorists have developed new ways to incorporate prior knowledge into models of complex cognition. For instance, recent work exploring causal induction has made use of hierarchical Bayesian modeling (Tenenbaum, Griffiths, & Kemp, 2006) to account for the way that prior knowledge limits the hypothesis space over which the probability of a particular causal relationship is computed. Prior knowledge is used to indicate which kinds of properties and relations are relevant for a given situation, populating the model at a higher level with a “general causal theory” (Griffiths & Tenenbaum, 2009) that limits the possible causal relationships considered in terms of the data. Similarly, there has been progress in accounting for how relational systems—for instance, kinship and categorical relationships—might be discovered by providing information about the general sets of concepts and relations that exist, and then computing which of the more explicit “theories,” or specific set of concepts and relationships, drawn from that initial space fit a particular data set (Kemp, Tenenbaum, Niyogi, & Griffiths, 2010). In both cases, without a means to limit the possibility space that these models consider, the models are unable to proceed (Jones & Love, 2011). But with appropriate constraints in place, the models are able to successfully learn relatively complex sets of concepts and their relationships. Using a very different formal system, Hummel, Licato, and Bringsjord (2014) provided a compelling account of the kind of relational processing that could underlie explanation, but again with no specific consideration of how the topic of the explanation is determined, since the constituents were specified *a priori*. We believe that the constraints we propose in our account—for instance, situational relevance and contrastive processing—could serve to inform components of more formal accounts of explanation generation.

Explanatory pluralism An important contribution of this approach is that it makes room for multiple styles of explanation within a single framework. Although much of the present work has focused on the functional and causal styles of explanation (e.g., Chin-Parker & Bradner, 2010; Lombrozo, 2009), other styles of explanation have been considered (Keil, 2006; Lombrozo, 2012; Prasada & Dillingham, 2009). We do not think it is in question as to whether people can generate different styles of explanations, but debate continues with regard to which styles should be considered explanatory. The contrastive account we have outlined here sets aside that debate and instead provides a processing framework from which multiple styles of explanation might be realized. It is important to recognize that this does not mean that, in situ, anything goes. As we detailed above, there are local constraints on the

generation of explanations. The flexibility of the contrastive account in this regard is an advantage, in that we can use the same framework to explore the variety of explanatory styles. However, there is also a cost to this flexibility. Hilton’s (1990) conversational model of explanation generation, which shares some features with the contrastive account described here, is able to go into greater detail about what occurs during the contrastive stage, because it focuses only on causal attribution. Although our account forgoes some of that specificity, we believe it is useful to provide an account that considers a fuller range of explanatory styles in order to develop a better understanding of the explanatory processes that are shared across those various styles.

Learning and generalization Explanation has been tied closely to learning (Chi, 2000; Lombrozo, 2006; Richey & Nokes-Malach, 2015; Williams & Lombrozo, 2010), so it is important to consider how the contrastive account informs that topic. As we argued previously, the operations that cooperate to specify the explanandum are necessary in order to generate an explanation. The properties of the event or object that are not involved in these operations are not fully implicated in the processing and subsequently less available for future cognitive tasks. For instance, Legare and Lombrozo (2014) showed that when children were asked to explain a toy, they had a harder time remembering the properties of the toy that did not correspond to the explanations they had generated. Contrastive processing can be specifically implicated in this outcome because information not highlighted during the contrastive processing may be ignored. As we mentioned before, Chin-Parker and Cantelon (2016) found that participants did not consider the properties shared across categories to be important when explaining membership in a target category, even though every member of the target category had that property. Those features were rarely mentioned in the explanations and did not affect the subsequent classification task performance. Furthermore, Lombrozo and Gwynne (2014) showed that the style of explanation used by the participant impacts the later generalization of information: After a participant had provided a mechanistic explanation, there was a willingness to generalize to entities that shared a particular mechanism, but after a functional explanation, there was a willingness to generalize to an entity that shared the same function. These examples illustrate that the operations that occur during explanation generation can limit access to information, which can, in turn, affect downstream learning and generalization.

There is evidence that explanation licenses knowledge of special relationships (e.g., causal schemas) or other generalizable patterns that can then be carried forward, applied to new cases, and used to bridge inductive gaps (Brem & Rips, 2000; Lombrozo, 2011; Lombrozo & Carey, 2006). Although we admit that some data support the idea that this kind of abstract

and transferable knowledge can emerge via explanation, we urge caution here. As we have developed in this account, the cognitive process of generating explanations is thoroughly site-specific. Answers are direct artifacts of the contextually situated possibility spaces that they (the answers) are constructed to satisfy, so there is reason to be cautious with regard to how “exportable” such answers would be to new cases. Ultimately, we agree that the generation of explanations is important to subsequent cognitive processes, but we are concerned that the transfer to new situations might not be as robust as some accounts suggest. Given that most studies of generalizability have not varied the context much between the explanation and transfer phases, it has been difficult to assess transfer. Shemwell, Chase, and Schwartz (2015) have provided a useful illustration of these tensions, showing that transfer to later problems can occur, especially when generalization is specifically prompted, but even then transfer is not guaranteed. There is clear evidence that explanation can play a role in learning, but we ask that future studies carefully distinguish (a) the role played by the explanation, which requires the use of an internal model but is site-specific, and (b) the activation and use of general knowledge in constructing the internal model, which might be more generalizable.

In thinking about how generating explanations might affect later cognitive activity, we think it is useful to consider accounts of transfer that specifically address the situated nature of knowledge (Yeh & Barsalou, 2006). A basic premise of this approach is that transfer from one situation to another is fairly conservative and relies on contextualized, as opposed to more abstract, information (e.g., Medin & Ross, 1989). We hypothesize that only in a situation with a similar explanandum possibility space can an individual reengage with the answers generated in a previous episode; only when the same sense of the explanandum is activated can we attribute the ability to abstract and transfer knowledge to some prior act of explaining. We recognize the speculative nature of these comments. However, they aim to clarify how this account provides for a different way of thinking about the role of explanation generation.

Concluding thoughts

The account outlined in this article is intended to identify and provide a structure for the various processes that occur during the generation of an explanation. We have not specified all of the cognitive processes involved. For instance, there is no indication yet of how the different components of the initial context (e.g., the physical environment, the internal goals of the explainer, and the needs of the questioner) might be weighted when identifying the topic of an explanation request, and our proposal does not detail the type of memory search conducted as the contrast class is populated. More study will be needed here.

As we mentioned previously, this account is geared toward understanding the generation of explanations. Many people investigate explanation in hopes of understanding other cognitive tasks, such a justification. For instance, some have explored the evaluation of explanations with an eye toward securing the epistemic value of a claim on the basis of its explanatory power (Colombo, 2017). Some of the contextual constraints we identify in this account may not apply to these other tasks. Chin-Parker and Bradner (2010) found that although the style of the explanations generated by participants shifted depending on which contrasting clips they viewed, the ratings of different styles of explanations did not. This suggests that the evaluation of explanations may rely on a different set of processes than on which the generation of explanations relies. We urge caution in accepting only those theories of explanation that can serve one’s preferred theory of justification, for that strategy can lead theorists to discount the pragmatic theory of explanation generation out of hand.

Our approach is an alternative to normative accounts of *good* explanation that attempt to identify such constraints *a priori*, often by prioritizing particular styles of explanation or particular qualities of the explanation. Accounts that specify how we should evaluate explanations are useful, because we can use the winning explanations to ground belief and reason well, but they say little about how explanations are in fact generated. Our account suggests that contextual and contrastive constraints in place during the generation of explanations specify the explanandum and the form and content of the answers that can satisfy the explanation request. Such constraints make it possible for an explainer to proceed in the face of an unbounded possibility space.

References

- Brem, S. K., & Rips, L. J. (2000). Explanation and evidence in informal argument. *Cognitive Science*, 24, 573–604.
- Chaigneau, S. E., Barsalou, L. W., & Sloman, S. A. (2004). Assessing the causal structure of function. *Journal of Experimental Psychology: General*, 133, 601–625. doi:10.1037/0096-3445.133.4.601
- Cheng, P. W., & Novick, L. R. (1992). Covariation in natural causal induction. *Psychological Review*, 99, 365–382.
- Chi, M. T. (2000). Self-explaining expository texts: The dual processes of generating inferences and repairing mental models. In R. Glaser (Ed.), *Advances in instructional psychology* (pp. 161–238). Mahwah, NJ: Erlbaum.
- Chi, M. T. H., de Leeuw, N., Chiu, M. H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, 18, 439–477.
- Chin-Parker, S., & Bradner, A. (2010). Background shifts affect explanatory style: How a pragmatic theory of explanation accounts for background effects in the generation of explanations. *Cognitive Processing*, 11, 227–249.
- Chin-Parker, S., & Cantelon, J. (2016). Contrastive constraints guide explanation-based category learning. *Cognitive Science*. Advance online publication. doi:10.1111/cogs.12405

- Cimpian, A. (2015). The inference heuristic: Generating everyday explanations. In R. Scott & S. Kosslyn (Eds.), *Emerging Trends in the Social and Behavioral Sciences* (pp. 1–15). Hoboken, NJ: John Wiley and Sons.
- Colombo, M. (2017). Experimental philosophy of explanation rising: The case for a plurality of concepts of explanation. *Cognitive Science*, 41, 503–517.
- Craik, K. J. W. (1967). *The nature of explanation* (2nd ed.). Cambridge, UK: Cambridge University Press.
- Davis, T., & Love, B. C. (2010). Memory for category information is idealized through contrast with competing options. *Psychological Science*, 21, 234–242.
- Dretske, F. (1981). The pragmatic dimension of knowledge. *Philosophical Studies*, 40, 363–378.
- Dumas, D., Alexander, P. A., & Grossnickle, E. M. (2013). Relational reasoning and its manifestations in the educational context: A systematic review of the literature. *Educational Psychology Review*, 25, 391–427.
- Ellman, T. (1989). Explanation-based learning: A survey of programs and perspectives. *ACM Computing Surveys*, 21, 163–221.
- Garfinkel, A. (1990). *Forms of explanation: Rethinking the questions in social theory*. New Haven, CT: Yale University Press.
- Gentner, D., & Markman, A. B. (1997). Structure mapping in analogy and similarity. *American Psychologist*, 52, 45–56. doi:10.1037/0003-066X.52.1.45
- Gentner, D., & Stevens, A. L. (Eds.). (1983). *Mental models*. Hillsdale, NJ: Erlbaum.
- Goldman, A. I. (1976). Discrimination and perceptual knowledge. *Journal of Philosophy*, 73, 771–791.
- Goldstone, R. L. (1996). Isolated and interrelated concepts. *Memory & Cognition*, 24, 608–628. doi:10.3758/BF03201087
- Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological Review*, 116, 661–716. doi:10.1037/a0017201
- Hale, C. R., & Barsalou, L. W. (1995). Explanation content and construction during system learning and troubleshooting. *Journal of the Learning Sciences*, 4, 385–436.
- Hilton, D. J. (1990). Conversational processes and causal explanation. *Psychological Bulletin*, 107, 65–81. doi:10.1037/0033-2909.107.1.65
- Hilton, D. J., & Erb, H. (1996). Mental models and causal explanation: Judgments of probable cause and explanatory relevance. *Thinking & Reasoning*, 2, 273–308.
- Hilton, D. J., & Slugoski, B. R. (1986). Knowledge-based causal attribution: The abnormal conditions focus model. *Psychological Review*, 93, 75–88. doi:10.1037/0033-295X.93.1.75
- Hitchcock, C., & Knobe, J. (2009). Cause and norm. *Journal of Philosophy*, 106, 587–612.
- Holyoak, K. J. (2012). Analogy and relational reasoning. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 234–259). Oxford, UK: Oxford University Press.
- Hummel, J. E., Licato, J., & Bringsjord, S. (2014). Analogy, explanation, and proof. *Frontiers in Human Neuroscience*, 8, 867. doi:10.3389/fnhum.2014.00867
- Johnson-Laird, P. N. (1983). *Mental models*. Cambridge, MA: Harvard University Press.
- Jones, M., & Love, B. C. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences*, 34, 169–188.
- Keil, F. C. (2006). Explanation and understanding. *Annual Review of Psychology*, 57, 227–254.
- Kemp, C., Tenenbaum, J. B., Niyogi, S., & Griffiths, T. L. (2010). A probabilistic model of theory formation. *Cognition*, 114, 165–196.
- Legare, C. H., & Lombrozo, T. L. (2014). Selective effects of explanation on learning during early childhood. *Journal of Experimental Child Psychology*, 126, 198–212.
- Lombrozo, T. (2006). The structure and function of explanations. *Trends in Cognitive Sciences*, 10, 464–470.
- Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive Psychology*, 55, 232–257.
- Lombrozo, T. (2009). Explanation and categorization: How “why?” informs “what?”. *Cognition*, 110, 248–253.
- Lombrozo, T. (2011). The instrumental value of explanation. *Philosophy Compass*, 6(8), 539–551.
- Lombrozo, T. (2012). Explanation and abductive inference. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford handbook of thinking and reasoning* (pp. 260–276). Oxford, UK: Oxford University Press.
- Lombrozo, T., & Carey, S. (2006). Functional explanation and the function of explanation. *Cognition*, 99, 167–204.
- Lombrozo, T., & Gwynne, N. Z. (2014). Explanation and inference: Mechanistic and functional explanations guide property generalization. *Frontiers in Human Neuroscience*, 8, 700. doi:10.3389/fnhum.2014.00700
- Macrae, C. N., Bodenhausen, G. V., & Milne, A. B. (1995). The dissection of selection in person perception: Inhibitory processes in social stereotyping. *Journal of Personality and Social Psychology*, 69, 397–407.
- McGill, A. L., & Klein, J. G. (1993). Contrastive and counterfactual reasoning in causal judgment. *Journal of Personality and Social Psychology*, 64, 897–905.
- Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review*, 100, 254–278. doi:10.1037/0033-295X.100.2.254
- Medin, D. L., & Ross, B. H. (1989). The specific character of abstract thought: Categorization, problem solving, and induction. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence* (Vol. 5, pp. 189–223). Hillsdale, NJ: Erlbaum.
- Pacer, M., Williams, J., Chen, X., Lombrozo, T., & Griffiths, T. (2013). Evaluating computational models of explanation using human judgments. In A. Nicholson & P. Smyth (Eds.), *Uncertainty in Artificial Intelligence: Proceedings of the Twenty-Ninth Conference (2013)* (pp. 498–507). Corvallis, OR: AUA Press. arXiv:1309.6855
- Patalano, A. L., Chin-Parker, S., & Ross, B. H. (2006). The importance of being coherent: Category coherence, cross-classification, and reasoning. *Journal of Memory and Language*, 54, 407–424.
- Prasada, S., & Dillingham, E. M. (2009). Representation of principled connections: A window onto the formal aspect of common sense conception. *Cognitive Science*, 33, 401–448.
- Read, S. J. (1987). Constructing causal scenarios: A knowledge structure approach to causal reasoning. *Journal of Personality and Social Psychology*, 52, 288–302.
- Richey, J. E., & Nokes-Malach, T. J. (2015). Comparing four instructional techniques for promoting robust knowledge. *Educational Psychology Review*, 27, 181–218.
- Rips, L. J., & Edwards, B. J. (2013). Inference and explanation in counterfactual reasoning. *Cognitive Science*, 37, 1107–1135.
- Ross, B. H., & Murphy, G. L. (1999). Food for thought: Cross-classification and category organization in a complex real-world domain. *Cognitive Psychology*, 38, 495–553.
- Rottman, B. M., & Keil, F. C. (2011). What matters in scientific explanation: Effects of elaboration and content. *Cognition*, 121, 324–337. doi:10.1016/j.cognition.2011.08.009
- Saatsi, J., & Pexton, M. (2013). Reassessing Woodward’s account of explanation: Regularities, counterfactuals, and non-causal explanations. *Philosophy of Science*, 80, 613–624.
- Schaffer, J. (2008). The contrast-sensitivity of knowledge ascriptions. *Social Epistemology*, 22, 235–245.
- Schank, R. C. (1982). *Dynamic memory: A theory of learning in people and computers*. New York, NY: Cambridge University Press.
- Shafto, P., Kemp, C., Mansinghka, V., & Tenenbaum, J. B. (2011). A probabilistic model of cross-categorization. *Cognition*, 120, 1–25.

- Shemwell, J. T., Chase, C. C., & Schwartz, D. L. (2015). Seeking the general explanation: A test of inductive activities for learning and transfer. *Journal of Research in Science Teaching*, 52, 58–83.
- Sinnott-Armstrong, W. (2008). A contrastivist manifesto. *Social Epistemology*, 22, 257–270.
- Skow, B. (2014). Are there non-causal explanations (of particular events)? *British Journal for the Philosophy of Science*, 65, 445–467.
- Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition*. Cambridge, MA: Harvard University Press.
- Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, 10, 309–318.
- Thagard, P. (2006). Evaluating explanations in law, science, and everyday life. *Current Directions in Psychological Science*, 15, 141–145.
- Van Bouwel, J., & Weber, E. (2008). A pragmatist defense of non-relativistic explanatory pluralism in history and social science. *History and Theory*, 47, 168–182.
- van Fraassen, B. (1980). *The scientific image*. Oxford, UK: Oxford University Press.
- Verheyen, S., De Deyne, S., Dry, M. J., & Storms, G. (2011). Uncovering contrast categories in categorization with a probabilistic threshold model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37, 1515–1531.
- Voorspoels, W., Storms, G., & Vanpaemel, W. (2012). Contrast effects in typicality judgments: A hierarchical Bayesian approach. *Quarterly Journal of Experimental Psychology*, 65, 1721–1739.
- Weiner, B. (1985). An attributional theory of achievement motivation and emotion. *Psychological Review*, 92, 548–573. doi:[10.1037/0033-295X.92.4.548](https://doi.org/10.1037/0033-295X.92.4.548)
- Williams, J. J., & Lombrozo, T. (2010). The role of explanation in discovery and generalization: Evidence from category learning. *Cognitive Science*, 34, 776–806.
- Williams, J. J., & Lombrozo, T. (2013). Explanation and prior knowledge interact to guide learning. *Cognitive Psychology*, 66, 55–84.
- Wilson, R. A., & Keil, F. C. (2000). The shadows and shallows of explanation. In F. C., Keil, & R. A. Wilson (Eds.) *Explanation and cognition* (pp. 87–114). Cambridge, MA: MIT Press.
- Yeh, W., & Barsalou, L. W. (2006). The situated nature of concepts. *American Journal of Psychology*, 119, 349–384.