Cognitive Psychology 106 (2018) 43–59

Contents lists available at ScienceDirect

Cognitive Psychology

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

How deep do we dig? Formal explanations as placeholders for inherent explanations

Susan A. Gelman[a,*], Andrei Cimpian[b,*], Steven O. Roberts[c]

[a] University of Michigan, USA
[b] New York University, USA
[c] Stanford University, USA

ABSTRACT

Formal explanations (e.g., "Mittens has whiskers *because she's a cat*") pose an intriguing puzzle in human cognition: they seem like little more than tautologies, yet they are surprisingly commonplace and natural-sounding. To resolve this puzzle, we hypothesized that formal explanations constitute an implicit appeal to a category's *inherent features* rather than simply to the category itself (as their explicit content would suggest); the latter is just a placeholder. We conducted a series of eight experiments with 951 participants that supported four predictions that followed from this hypothesis: First, formal explanations—though natural-sounding—were not particularly satisfying. Second, for natural kinds, formal explanations were less satisfying than inherent explanations (specifically, ones that appealed to a natural kind's causally powerful "essence"). Third, participants viewed essence-related inherent explanations as more specific versions of the ideas expressed by formal explanations, which were viewed as more general placeholders. Fourth, and finally, formal explanations tended to serve as placeholders for explanations that appealed to inherent features more so than for other types of explanations, such as ones that appealed to external, environmental factors. In addition to supporting our novel claim about the meaning of formal explanations, these data suggest a new way in which explanations do their psychological work: not via their literal content (as assumed by prior work on explanation), but rather via the additional inferences they encourage. We end by discussing the potential heuristic value of formal explanations for causal learning in childhood.

1. Introduction

People often explain what they observe (e.g., Fido has four legs) by simply appealing to a category, with statements such as "Because it's a dog," "Dogs are dogs," or "That's the way dogs are" (e.g., Prasada & Dillingham, 2006, 2009; Sánchez Tapia et al., 2016). These *formal explanations* are a unique mode of explanation.[1] Moreover, they are intuitively appealing (i.e., they sound natural); they extend across a wide range of domains (including natural kinds, artifacts, and social kinds); and they are systematically and distinctively linked to certain features (those with principled connections to kinds, such as dogs having four legs) but not others (those with statistical connections to kinds, such as dogs wearing collars; Prasada, 2017; Prasada & Dillingham, 2006, 2009). Formal explanations are also common in everyday discourse: preschool children as well as adults readily produce formal explanations to explain features with a principled connection to a kind (Coley & Vasilyeva, 2010; Haward, Wagner, Carey, & Prasada, 2017; Roberts, Gelman, & Ho, 2017; Taylor, Rhodes, & Gelman, 2009). Yet formal explanations are also puzzling. For instance, explaining why a dog has four legs by saying "Because it's a dog" doesn't tell us anything we don't already know, and "Dogs are dogs" is little more than a tautology. Given that formal explanations seemingly add little value to one's understanding of a property, why are they so common?

* Corresponding authors.
E-mail addresses: gelman@umich.edu (S.A. Gelman), andrei.cimpian@nyu.edu (A. Cimpian).

[1] The present paper focuses on a subtype of formal explanations—those that refer to categories, and that explain properties shared by category members. Formal explanations can also include non-category explanations, such as those referring to mathematical concepts or logical relations (e.g., Lombrozo & Vasilyeva, 2017), but these are beyond the scope of this investigation.

https://doi.org/10.1016/j.cogpsych.2018.08.002
Accepted 21 August 2018

Any account of explanation that is to take psychological data seriously must grapple with this intriguing phenomenon.

1.1. The proposal: Formal explanations as placeholders

The premise of this paper is that formal explanations, which on the surface are a direct appeal to kinds, may actually constitute placeholders for more-specific explanations that appeal to *features* of those kinds. To summarize, our argument is that people use formal explanations for reasons beyond the belief that category membership *itself* provides a satisfactory explanation for the facts at hand. Instead, we suggest that people rely on explanations that appeal to category membership (e.g., "because they're dogs") because they see these explanations as suggesting the existence of unstated or unknown category features that are deeper and more explanatory. The nature of these features will vary by domain, although we propose that people will frequently (but not obligatorily) settle on features that we term 'inherent.' This term refers to properties of an entity that are entirely about that entity rather than involving relations to other entities (e.g., Lewis, 1983; Weatherson & Marshall, 2017). For instance, in the case of natural kinds (e.g., dogs, diamonds), people may infer that the explanatory work is being done by a category-specific, inherent causal "essence" (Gelman, 2003; Medin, 1989; Rhodes & Mandalaywala, 2017). More generally, formal explanations may be viewed as stand-ins for more-specific (and often inherent) explanations that appeal to features of the category. In the words of Lombrozo and Vasilyeva (2017), "at least some formal explanations [may be] understood causally, as pointers to some category-associated essence or causal factor responsible for the properties being explained."

To elaborate on the centrality of inherent features in this proposal, we hypothesize that a formal explanation (e.g., "Because it's an X") will be generally perceived as under-informative, and so will prompt people (via pragmatic processes; Grice, 1975) to "look further" and seek more informative explanations. Due in part to the heuristic shortcuts that people use routinely in explanatory reasoning (e.g., Cimpian & Salomon, 2014), this process of "looking further" often leads people to arrive at more-specific interpretations in terms of inherent features (e.g., "it's some inherent feature of a dog, such as its DNA, that causes it to have four legs"). Inherent properties may be a frequent means of elaborating formal explanations in part because they are *plausible* as explanations in this context. To see why, consider that formal explanations are appropriate only for features that are thought to be aspects of their kinds (e.g., dogs have four legs; Prasada & Dillingham, 2009); in contrast, features that are merely statistically associated with their kinds (e.g., dogs wear a collar) do not receive formal explanations. If formal explanations simply signaled that some further explanation was available—without any constraints on the nature of this explanation—then they should be equally acceptable for all properties associated with a category. The fact that they are not is consistent with our hypothesis that formal explanations are primarily placeholders for a particular type of explanation—namely, one that appeals to inherent aspects of a category's members. Typically, extrinsic facts (e.g., facts pertaining to circumstances or past events) do not provide satisfying explanations for features that are seen as an integral aspect of a kind (e.g., dogs having four legs), but inherent facts do (e.g., Cimpian & Steinberg, 2014; Salomon & Cimpian, 2014). This will likely also guide reasoners' attention to inherent facts as they are trying to determine what formal explanations are standing in for.

1.2. Contributions to the literature on explanation

The present argument and the eight experiments that test it make several theoretical contributions to the literature on explanation. First, and most directly, the present proposal advances our understanding of formal explanations by identifying an implicit layer of meaning in these common, yet seemingly so empty, explanations. Formal explanations are acceptable and common in part because they are understood as placeholders for more-specific explanations involving inherent features of the relevant category. Second, our argument identifies a novel aspect of how explanations do their psychological work more generally. An unspoken assumption in the literature has been that the meaning of an explanation is exhausted by its literal content. If our proposal is correct, it would open the door to studying the additional inferences that explanations may promote and the additional benefits these implicit layers of meaning might have for those who receive them. Third, and related to the preceding point, treatments of explanation often underplay the psychological function or utility of explanations and view them instead as objectively correct answers to why-questions (see also Lombrozo & Carey, 2006). The present account strongly notes the psychological utility of formal explanations, in terms of their heuristic value. Fourth, the current contribution is novel in moving away from the typical view of explanation as a static *product* and toward a more psychologically apt view of explanation as a *process* that interacts with other cognitive processes (e.g., Cimpian & Keil, 2017; Lombrozo, 2012). Explanations are, in part, (1) communicative acts and (2) judgments, and thus a full understanding of how they work must integrate insights from (1) psycholinguistics (the "look further," pragmatic component of our argument) as well as (2) the literature on reasoning and judgment/decision-making (the inherence heuristic component of our argument). Thus, our argument situates explanation in the context of other cognitive processes with which it obviously must interact but which the prior literature has largely overlooked.

1.3. Predictions

We test four predictions that follow from our argument: If formal explanations are interpreted as a placeholder for inherent explanations, then we should find that (1) formal explanations are natural-sounding but not particularly satisfying/good (Experiment 1), (2) formal explanations are less satisfying than inherent explanations, particularly for properties with principled connections to their kinds (Experiments 2–5), and (3) inherent explanations are judged to be more specific versions of the placeholder-like formal explanations (Experiments 6–8). However, formal explanations should not be viewed as placeholders for just any other explanations: