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:

(4) We predict that participants will more often assume a hierarchical relationship between formal explanations and explanations that appeal to inherent as compared to extrinsic facts (Experiment 8).

We will test these predictions largely in the context of natural kinds. However, in principle our argument extends beyond this domain. For instance, Knobe, Prasada, and Newman (2013) suggest that dual-character concepts such as *scientist* or *artist* are held together by certain ideals or values (e.g., scientists share a commitment to arriving at the truth using empirical observations; see also Gelman & Rhodes, 2012). In these cases as well, formal explanations of a category member's features and behavior (e.g., "she is in the lab because she is a scientist") may just be placeholders for explanations that appeal to the inherent values associated with the category (e.g., finding out the truth via empirical observation). Though the scope of our argument is broad, here we took a first step toward testing it by focusing mostly (with the exception of Study 1) on formal explanations of the features of natural kinds.

We also note the boundaries of our argument. Most importantly, formal explanations will "bottom out" in cases where the members of a category do not actually share many inherent features to which one can appeal (e.g., nominal kinds; Schwartz, 1980). For example, a formal explanation such as "that is round because it's a circle" does not stand in for a more detailed analysis, since no such analysis is possible here.

1.4. Overview of the present experiments

We conducted eight experiments to test the four predictions above. First, we asked participants to rate how natural and satisfying they found a range of formal explanations (Experiment 1), with the prediction that formal explanations would be perceived as natural but not very satisfying. Second, we compared formal explanations for the features of natural kinds with the explanations they are hypothesized to stand in for—namely, more-specific explanations that appeal to inherent properties. We asked participants to rate how satisfying formal and inherent explanations are and predicted that inherent explanations would be rated as more satisfying than formal explanations (Experiments 2–5). To test our third prediction that formal explanations serve as an under-informative placeholder filled in by explanations appealing to inherent features, we asked participants to judge the logical relation between formal and inherent explanations (Experiments 6–8). We predicted that people would judge this relation to be one of inclusion: inherent explanations are more specific restatements of formal explanations. To test our fourth prediction that this inclusion relation is unique to formal and inherent explanations, in Experiment 8 we also asked participants to judge the logical relation between formal and extrinsic (specifically, environmental) explanations. Our prediction was that formal explanations would not be seen as placeholders for these other explanations. Across these experiments, the data support the notion of formal explanations as a pointer to inherent explanations that are more informative and, as a result, more satisfying.

2. Experiment 1: How natural and satisfying are formal explanations?

Prior research has found that formal explanations are judged as sounding natural—it sounds perfectly fine to say that Fido has four legs because he is a dog. However, the goodness of an explanation is more appropriately judged by how satisfying it is (e.g., Lombrozo, 2007; Lombrozo & Carey, 2006). We hypothesized that formal explanations, though natural, would be judged as relatively unsatisfying (because, we argue, they merely serve as placeholders). We tested this in a simple and direct way: participants were asked to provide ratings of how natural and satisfying they found a variety of formal explanations. If formal explanations are placeholders for inherent explanations, then participants should find formal explanations natural, but not satisfying.

Our goal was to obtain an accurate assessment of whether formal explanations are seen as satisfying. However, the typical way of assessing formal explanations in the literature might overestimate their explanatory value. In prior research examining formal explanations, people were asked to judge question-answer pairs of this form (Prasada & Dillingham, 2006): *Why does that [pointing to a dog] have four legs? / Because it is a dog.* Because the question did not include the category label (i.e., the item was referred to as "that" rather than "that dog"), the value of a formal explanation may have been inflated. Without the label, the question left unclear whether the questioner has appropriately identified the item, so the response—which mentions the item's category—could be seen as informative in part because it ensures that the questioner has properly recognized the item. In other words, whether the question includes or excludes the category label could affect people's judgments of how satisfying they find a formal explanation to be. In Experiment 1, we therefore systematically varied whether or not the questions included the category label in order to determine how much these wording differences affect people's judgments.

2.1. Method

2.1.1. Participants

Seventy-eight Mechanical Turk (MTurk) workers participated: 35 men, 43 women; M_{age} = 33.7 years, range 18–71; 53 White/Caucasian, 8 Black/African-American, 8 Asian/Asian-American, 4 Latino/Hispanic, 1 Native American, 3 Multiracial, 1 Non-reported. Each was randomly assigned to one of four conditions: Natural/Label (n = 20), Natural/No Label (n = 19), Satisfying/Label (n = 18), and Satisfying/No Label (n = 21).

2.1.2. Design

The study had a 2 (Judgment: Natural vs. Satisfying) × 2 (Question Wording: Label vs. No Label) × 3 (Domain: Living Kind, Artifact, Social) design, with Judgment and Wording as between-subjects factors, and Domain as a within-subject factor.

Table 1
Items used in Experiment 1.

	Category/property	Prasada and Dillingham (2006) rating
Living kind	Bananas/yellow	5.28
	Birds/fly	5.61
	Carrots/crunchy	5.39
	Cheetahs/run fast	5.56
	Cherries/red	5.50
	Dogs/four legs	5.06
Artifact	Cars/four wheels	5.39
	Diapers/absorbent	5.72
	Fire trucks/hoses	5.61
	Needles/sharp	5.33
	Raincoats/waterproof	5.39
	Tables/flat	5.39
Social	Architects/design buildings	5.44
	Artists/creative	5.28
	Cheerleaders/spirited	5.33
	Christians/read the Bible	5.28
	Doctors/diagnose ailments	5.61
	Journalists/report news	5.61

Table 2
Sample item in each of the four wording conditions in Experiment 1.

Condition	Sample Item	Judgment
Natural/Label	Joe was asked, "Why do dogs have four legs?" Joe replied, "Because they are dogs"	How natural is this response? (1 = not at all natural; 7 = extremely natural)
Natural/No Label	Joe was asked, "Why do those [*pointing to some dogs*] have four legs?" Joe replied, "Because they are dogs"	How natural is this response? (1 = not at all natural; 7 = extremely natural)
Satisfying/Label	Joe was asked, "Why do dogs have four legs?" Joe replied, "Because they are dogs"	How satisfying is this explanation? (1 = not at all satisfying; 7 = extremely satisfying)
Satisfying/No Label	Joe was asked, "Why do those [*pointing to some dogs*] have four legs?" Joe replied, "Because they are dogs"	How satisfying is this explanation? (1 = not at all satisfying; 7 = extremely satisfying)

2.1.3. Materials and procedure

Each participant received 18 items, six from each of three domains (living kind, artifact, social), each involving a category paired with a typical property, such as dogs/four legs (see Table 1 for full listing). The items were a subset of the principled connections used by Prasada and Dillingham (2006; Experiment 2A); we selected items for which formal explanations had the highest naturalness ratings within that domain.[2] For each item, participants read a question/response pair and judged the response for either how natural it was or how satisfying it was, on a 7-point scale (1 = not at all natural/satisfying; 7 = extremely natural/satisfying).

The instructions for naturalness ratings were identical to those from Prasada and Dillingham (2006): "We are simply interested in your gut feeling about how natural or good a given response is to the question. We are not interested in whether it is a scientifically or socially good or acceptable answer, only if it feels natural." The instructions for satisfyingness ratings were as follows: "We are simply interested in your gut feeling about how satisfying a given explanation is. We are not interested in whether it is a scientifically or socially good or acceptable explanation, only if it feels satisfying."[3]

For half the participants, the question included the category label (e.g., "Why do dogs have four legs?"); for half, the question did not include the category label (e.g., "Why do those [*pointing to some dogs*] have four legs?"). The response was always a formal explanation (e.g., "Because they are dogs"). Each participant was randomly assigned to one of four conditions (Natural/Label, Natural/No-Label, Satisfying/Label, Satisfying/No-Label), which crossed judgment type with question-wording type. Items were blocked by domain; the order of the blocks and the order of items within each block were randomized separately for each participant. Table 2 includes a sample item from each of the conditions.

[2] There were only two exceptions: for the Natural items, we did not include "grass/green" (rating of 5.10), in order to avoid mass nouns, and for the Social items, "dancers/move gracefully" (rating of 5.28) was tied with two other items (artists, Christians).

[3] Because one key goal of this experiment is to determine how the naturalness measure used in prior research relates to explanation 'goodness' (as assessed by satisfyingness ratings), the analyses include direct comparisons across the two scales (see Results and Discussion, below). Although in general one must be cautious in assuming that different scales have equivalent meaning (e.g., that '2' indicates the same level of support across different scales), the current scales were designed to minimize this concern. Specifically, the naturalness and satisfyingness scales were carefully designed to be as equivalent to one another as possible (identical range, scale, and intensifiers; closely matched instructions).

2.1.4. Open data and syntax

The raw data and analytic syntax for this and all subsequent studies are available on Open Science Framework: https://osf.io/bk3h5/?view_only=7da41aea840b42398acf8141c56a2956

2.2. Results and discussion

We analyzed participants' ratings with a multilevel mixed-effects linear model with cross-classified random intercepts for participants and items. This analytic strategy permits us to test if the results generalize along both dimensions of participants and items (e.g., Baayen, Davidson, & Bates, 2008). Question-wording (Label, No Label) and Judgment (Natural, Satisfying) were between-subjects (level-2), dichotomous variables, and Domain (Living Kind, Artifact, Social) was a within-subject (level-1), three-level categorical variable. All two- and three-way interactions of these predictors were also included in the model.

Our primary hypothesis was supported—that is, formal explanations were judged to be considerably more natural-sounding ($M = 4.25$; see Prasada & Dillingham, 2006) than satisfying ($M = 2.85$; both judgments on a 1–7 scale), Wald $\chi^2 = 22.41$, $p < .001$. Simple effects tests confirmed that this difference was significant in each of the domains tested separately, $ps < .001$ (see Table 3).

Table 3
Mean naturalness and satisfyingness ratings in Experiment 1 (on a scale of 1 = not at all natural/satisfying to 7 = extremely natural/satisfying), as a function of domain and wording. SDs are in parentheses.

	Living Kind	Artifact	Social
No label ("these")			
Natural?	4.15 (1.99)	4.29 (1.95)	4.78 (1.76)
Satisfying?	2.93 (1.94)	3.34 (2.01)	3.87 (2.07)
Label ("dogs")			
Natural?	3.74 (1.39)	3.95 (1.62)	4.60 (1.61)
Satisfying?	1.84 (1.13)	2.47 (1.67)	2.65 (1.76)

Our secondary hypothesis was supported as well—namely, ratings for formal explanations were higher when no label was provided in the question ($M = 3.89$) than when a label was provided in the question ($M = 3.21$), Wald $\chi^2 = 5.42$, $p = .020$. This boost was significant for the ratings of how satisfying an explanation is ($M_{\text{label}} = 2.32$ vs. $M_{\text{no-label}} = 3.38$, Wald $\chi^2 = 6.48$, $p = .011$), but not for ratings of how natural it is ($M_{\text{label}} = 4.10$ vs. $M_{\text{no-label}} = 4.41$, Wald $\chi^2 = 0.55$, $p = .46$). The Question-wording × Judgment interaction was not significant, however, Wald $\chi^2 = 1.63$, $p = .20$. These results suggest that merely providing a label when one was absent in the question accounts for some of the value of formal explanations. Participants may have assumed that when the questioner failed to provide a label, he or she was ignorant of the category membership of the entity in question, in which case providing the label is deemed helpful. Conversely, when the questioner indicates knowledge of the category by including the label in the question, the formal explanation loses some of its appeal. Thus, it seems likely that pragmatic factors are at play in either boosting or depressing participants' ratings (which is broadly consistent with our argument that pragmatics are also at play in how participants ultimately interpret these explanations, by "looking further" for more-informative explanations).

Finally, we also found that ratings differed by domain, Wald $\chi^2 = 19.08$, $p < .001$. Items in the Social domain received higher ratings than those in either the Living Kind or Artifact domain, Wald $\chi^2 = 18.93$ and 6.31, respectively, $ps \leq .012$, but there was no significant difference between Living Kind and Artifact items, Wald $\chi^2 = 3.38$, $p = .066$. The high ratings in the social domain may reflect the quasi-definitional status of some of these features (e.g., a journalist would not be a journalist if he or she didn't report news). None of the interactions in the main model were significant, all $ps > .084$.

In sum, participants did not judge formal explanations to be very satisfying. This finding provides important support for the first prediction of our proposal (that formal explanations are more natural-sounding than satisfying) and is at least consistent with the broader proposal that formal explanations constitute an implicit appeal to a category's inherent features rather than simply to the category itself. However, these data also raise the question of precisely how satisfying formal explanations are, which we explore in the next experiment.

3. Experiment 2: Formal vs. inherent explanations

Experiment 1 suggests that formal explanations sound natural, yet are relatively unsatisfying. Experiment 2 is designed to address our second prediction, that formal explanations are less satisfying than inherent explanations, particularly for properties with principled connections to their kinds. Past research found that formal explanations were judged as more natural than one type of inherent explanation—that is, essentialist explanations (Prasada & Dillingham, 2006). However, those data did not speak to the issue of how *satisfying* people find the two types of explanations. Furthermore, as the authors themselves noted, the relatively low naturalness ratings for essentialist explanations may have reflected the somewhat stilted wording used in the experiments (e.g., "Because it has the essence of a dog which causes it to have four legs"). We therefore designed Experiment 2 to focus on ratings of how satisfying formal and inherent (specifically, essentialist) explanations are, using more colloquial wording for the latter. Here and in all

subsequent studies, we carried out our comparison of formal and inherent explanations using the domain of natural kinds as a test case.

3.1. Method

3.1.1. Participants

Seventy-four MTurk workers participated: 35 men, 39 women; M_{age} = 33.0 years, range 19–68; 53 White/Caucasian, 4 Black/African-American, 6 Asian/Asian-American, 7 Latino/Hispanic, 4 Multiracial.[4] Additionally, 36 MTurk workers participated in a separate norming study (described below, in Materials and Procedure).

3.1.2. Design

The study had a 2 (Explanation: Formal, Inherent) × 3 (Category Type: Animals, Plants, Substances) design, with Explanation and Category Type as within-subject factors.

3.1.3. Materials and procedure

In this and subsequent studies, our stimuli were selected from a single domain: natural kind categories. Natural kinds were chosen in part because prior research provides a detailed portrait of what inherent features people generally find explanatory in this domain (i.e., physical, internal, structural inherent features, often termed "essences"; e.g., Ahn et al., 2001; Gelman, 2003). Our stimuli included animals, plants (including plant products, such as fruit), and natural substances. All questions included the category label to ensure that judgments focused on the explanatory value of the different explanations (vs. simply identifying a potentially ambiguous item; see Experiment 1). Given the interest in explanatory force, and given the results of Experiment 1, participants only judged how satisfying the explanations were.

Each participant received 12 items, four from each of three types of natural kinds (animals, plants, substances). Each item involved a category paired with a typical property, as in Experiment 1. All of the natural kinds from Experiment 1 were included, to which we added one animal (zebras/stripes), one plant (roses/thorns), and four substances (diamonds/hard, gold/shiny, iron/magnetic, water/transparent).

We chose these kind/property pairs so that they all embody principled connections. This assumption was validated by a separate norming study that we conducted on MTurk (N = 36). There were 36 items in the norming study: the 12 items included in the main study (see above), as well as 24 items selected to have statistical connections with their kinds (two from each of the same 12 categories; e.g., "Dogs wear collars"), half of which were selected for use in Experiment 5. For each item in this norming study, participants read a sentence (e.g., "Dogs have four legs"), followed by a sentence expressing a principled connection between the category and the property (e.g., "Dogs, by virtue of being the kinds of things they are, have four legs") (wording based on Prasada & Dillingham, 2006). They were then asked to judge whether the second sentence could be understood to be a good paraphrase of the first sentence, on a scale of 1 = not at all good to 7 = very good. The average rating for the 12 items included in this study was 4.61, as compared to 3.44 for the statistical connections, a statistically significant difference, $t(35) = 5.40$, $p < .001$. The ratings for these items can be found in Table 4.

The explanations included both formal explanations (as in Experiment 1) and inherent explanations. The inherent explanations in this and subsequent studies made reference to internal, structural properties that approximate what people believe to be the "essences" of natural kinds (e.g., Ahn et al., 2001; Gelman, 2003; Rhodes & Mandalaywala, 2017). They were tailored to their respective ontological type (e.g., it was assumed that an inherent feature of a dog would differ from one of gold). For the animals and plants, the inherent explanations were: "because of deeper biological features" and "because of their DNA." For the substances, the inherent explanations were: "because of its/their chemical structure" and "because of its/their underlying molecular properties." Which inherent explanation a participant received for a given item was randomized.

For each item, participants read a question (e.g., "Abi asked, 'Why do dogs have four legs?'") followed by two different explanations, each provided by a different person. Each item paired a formal explanation (e.g., "Person 1 replied, 'Because they are dogs.'") with an inherent explanation (e.g., "Person 2 replied, 'Because of deeper biological features.'"). The order in which these explanations appeared for a given item was randomized within participants.

For each item, after reading both explanations, participants judged each of the two explanations for how satisfying it was, on a 7-point scale (1 = not at all satisfying; 7 = extremely satisfying). The instructions for satisfyingness ratings were the same as in Experiment 1.

3.2. Results and discussion

A mixed-effects linear model revealed a main effect of Explanation (M_{formal} = 2.28; $M_{essentialist}$ = 4.70), Wald χ^2 = 1470.77, $p < .001$, and an Explanation × Category Type interaction, Wald χ^2 = 73.38, $p < .001$ (see Table 5 for means). The significant interaction revealed that the formal-vs.-inherent difference was larger for substances than either animals or plants, Wald χ^2 = 32.39

[4] Experiments 2 and 3 and the experiment reported in the online supplement also included attention checks (trials on which participants were instructed to answer a certain way, e.g., "Select '2'") to weed out people who were not paying attention. Over 90% of participants in each of these studies passed the attention checks, and the effects were identical when such participants were included or excluded.

Table 4
Ratings of principled connections (for Experiments 2–5) and statistical connections (for Experiment 5), on a scale of 1–7, where higher numbers indicated more principled connections.

	Category	Principled (Expts. 2–5)		Statistical (Expt. 5)	
Animals	Birds	Fly	4.58	Sit on statues	3.33
	Cheetahs	Run fast	4.94	Found in a zoo	3.03
	Dogs	Four legs	4.42	Wear a collar	2.97
	Zebras	Stripes	4.11	Have dirty hooves	3.83
Plants	Bananas	Yellow	4.44	Cheap	2.89
	Carrots	Crunchy	4.44	Packaged in a plastic bag	2.78
	Cherries	Red	4.25	Sold by weight	2.64
	Roses	Thorns	4.42	Come in a dozen	2.58
Substances	Diamonds	Hard	5.00	Mined in a war zone	2.61
	Gold	Shiny	4.83	Stored as bricks	3.42
	Iron	Magnetic	4.89	Found at an ancient archeological site	3.53
	Water	Transparent	5.00	Absent on the moon	3.25

Table 5
Mean satisfyingness ratings of formal and inherent explanations (on a scale of 1–7) in Experiment 2, as a function of category type. SDs in parentheses.

	Formal	Inherent
Animals	2.29 (1.67)	4.57 (1.70)
Plants	2.43 (1.77)	4.25 (1.85)
Substances	2.12 (1.51)	5.28 (1.48)

and 76.05, $ps < .001$, and larger for animals than plants, Wald $\chi^2 = 9.23$, $p = .002$. Importantly, however, inherent explanations were judged to be more satisfying than formal explanations within each of the three types of categories, all $ps < .001$.[5]

We also found a main effect of Category Type, Wald $\chi^2 = 23.36$, $p < .001$. Substances received higher ratings than Animals or Plants, Wald $\chi^2 = 11.93$ and 21.67, respectively, $ps < .001$, but there was no significant difference between Animals and Plants, Wald $\chi^2 = 1.45$, $p = .229$.

These data suggest that formal explanations are judged to be not very satisfying (as in Experiment 1; see also the Supplementary Experiment), and also demonstrate that formal explanations are less satisfying than inherent explanations. The differences held up consistently across the three different superordinate categories of natural kinds (animals, plants, and substances). These findings are consistent with our hypothesis that formal explanations point the way toward (more informative and satisfying) inherent explanations.

However, these results may be due to differences in the wording of the formal and inherent explanations in this study, beyond explanatory value per se. For example, the inherent explanations included words that likely sounded more scientific or sophisticated (e.g., "biological," "chemical," "molecular," "DNA"), and may have been rated more highly for this reason. Additionally, across the course of the experiment, participants repeatedly heard just a single formal explanation, in contrast to four different inherent explanations. It is possible that the repetition of the formal explanation rendered it relatively less satisfying. We therefore designed the next experiment (Experiment 3) to control for both these factors.

4. Experiment 3: Formal vs. inherent explanations, controlling for language register and number of distinct explanations

Experiments 3 and 4 were designed to control for (1) the register of the terminology used across formal vs. inherent explanations (specifically, how scientific and sophisticated the language of the explanations was) and (2) the number of distinct explanations offered for each explanation type across the course of the task. To control for register, we obtained ratings of how sophisticated and scientific the terminology of each explanation was, and used these ratings as a covariate in the analyses. To control for the number of distinct explanations offered for each explanation type, we provided three distinct formal explanations and three distinct inherent explanations. Experiments 3 and 4 are nearly identical; the small differences between them are detailed below.

4.1. Method

4.1.1. Participants

Sixty-two MTurk workers participated: 27 men, 35 women; $M_{age} = 35.4$ years, range 19–69; 45 White/Caucasian, 6 Black/

[5] We also conducted post-hoc *t* tests to test, within each category type, how each inherent explanation (e.g., for animals and plants: "because of deeper biological features", "because of their DNA") compared to formal explanations. For all six comparisons, the inherent explanations received higher ratings than the corresponding formal explanations, $ts(73)$ ranging from 6.88 to 25.98, $ps < .001$.

Table 6
Mean sophistication and scientific-ness ratings for the formal and inherent explanations in Experiments 3 and 4.

	Explanation	Sophist.	Scient.
Formal	(F1) Essentially, they are that way because they are Xs.	2.24	2.14
	(F2) They are that way simply because they are Xs.	1.96	1.90
	(F3) They are that way because they are in fact Xs.	1.92	2.13
Inherent	(I1) They are that way because of deeper internal features.	2.51	2.56
	(I2) They are that way because of deeper physical features.	2.53	2.74
	(I3) They are that way because of deeper structural features.	2.63	2.71

African-American, 2 Asian/Asian-American, 6 Latino/Hispanic, 3 Multiracial. An additional person was dropped for failing an attention check. An additional 36 MTurk participants provided scientific/sophistication ratings (described below, in Materials and Procedure).

4.1.2. Design

The study had a 2 (Explanation: Formal, Inherent) × 3 (Category Type: Animals, Plants, Substances) design, with Explanation and Category Type as within-subject factors.

4.1.3. Materials and procedure

The task was modeled on Experiment 2, using the same categories/properties and procedure. The only difference was the wording of the explanations (see Table 6). On each trial, participants read one of three new formal explanations and one of three new inherent explanations, each of which could apply to any of the animals, plants, or substances in the experiment (see Table 4 for a full list). To simplify presentation and analyses, formal and inherent explanations were paired (F1-I1, F2-I2, F3-I3; see Table 6), and which pair appeared on each trial was randomized.

To obtain ratings of how scientific and sophisticated the language was in the explanations, we first generated a list that included all the words of the actual explanations[6] as well as filler words (54 words total). We asked a separate group of 36 participants to rate each word in isolation, presented in random order. This process allowed us to obtain an objective measure of the components of each explanation, uncontaminated by participants' evaluations of each explanation as a whole. Each participant rated all 54 words on two dimensions: "How scientific do the words below sound to you?", and "How sophisticated to the words below sound to you?" (from 1 = not at all scientific/sophisticated to 7 = extremely scientific/sophisticated). Whether participants first judged words as scientific or sophisticated was counterbalanced. The ratings in Table 6 are averaged across each word token in the explanation.

4.2. Results and discussion

The Scientific and Sophistication ratings were highly correlated, $r = 0.93$, $p < .001$, and thus were averaged to create a composite Science/Sophistication variable. We conducted a multilevel mixed-effects linear model with cross-classified random intercepts for participants, items, and explanation wordings. The fixed effects included Explanation (a within-subject [level-1], dichotomous variable); Category Type (a within-subject [level-1], three-level categorical variable); the interaction of these two variables; and Science/Sophistication (a continuous, within-subject [level-1] covariate).

Results indicated a main effect for Explanation ($M_{\text{formal}} = 2.31$; $M_{\text{inherent}} = 3.77$), Wald $\chi^2 = 32.36$, $p < .001$, and an Explanation × Category Type interaction, Wald $\chi^2 = 6.57$, $p = .037$ (see Table 7 for means). The interaction was driven by the fact that the inherent advantage was significantly larger for Substances than for Animals, Wald $\chi^2 = 6.57$, $p = .010$. Importantly, however, inherent explanations were judged to be more satisfying than formal explanations within each of the three types of categories, all $ps < .001$, even when controlling for the register of the language in the two types of explanations. The Science/Sophistication covariate was itself positively related to satisfyingness ratings, $b = 0.60$, $z = 1.77$, $p = .077$, but this relationship did not reach significance.

Again, we find that inherent explanations are rated as more satisfying than formal explanations even when adjusting for any differences between these explanation types in how scientific and sophisticated they sound. Thus, the register of the language employed cannot explain the differences obtained in satisfyingness ratings.

5. Experiment 4: Replication of experiment 3

Experiment 4 provided a replication of Experiment 3, with one methodological change: we provided just one explanation per trial, rather than two. This design provides a more conservative test of the prediction that inherent explanations are more satisfying than formal explanations. If participants are not invited to compare formal and inherent explanations side by side, then they have the opportunity to report that they find both options equally satisfying.

[6] The ratings focused exclusively on words in the explanations that were constant across items. Thus, we did not ask raters to assess the category labels themselves (e.g., 'birds,' 'bananas,' 'iron'), although these words appeared in the formal explanations (i.e., 'Xs' in Table 6).

Table 7
Satisfyingness ratings (on a scale of 1–7) as a function of explanation type and category type in Experiment 3. SDs in parentheses.

	Formal	Inherent
Animals	2.41 (1.69)	3.68 (1.83)
Plants	2.23 (1.53)	3.69 (1.76)
Substances	2.29 (1.58)	3.94 (1.75)

Table 8
Satisfyingness ratings (on a scale of 1–7) as a function of explanation type and category type in Experiment 4. SDs in parentheses.

	Formal	Inherent
Animals	1.75 (1.34)	2.79 (1.65)
Plants	1.77 (1.30)	2.80 (1.60)
Substances	1.81 (1.37)	2.94 (1.64)

5.1. Method

5.1.1. Participants

Two-hundred-two MTurk workers participated: 104 men, 97 women, 1 other; M_{age} = 35.0 years, range 19–68; 151 White/Caucasian, 12 Black/African-American, 11 Asian/Asian-American, 14 Latino/Hispanic, 11 Multiracial, 3 Other or Not Reported. The sample size was larger here than in Study 3 to maintain sufficient statistical power given that we had fewer observations per participant and also expected a smaller effect.

5.1.2. Design

The study had a 2 (Explanation: Formal, Inherent) × 3 (Category Type: Animals, Plants, Substances) design, with Explanation and Category Type as within-subject factors.

5.1.3. Materials and procedure

The items were identical to those of Experiment 3. The task was very similar, except that (1) participants rated just one explanation at a time, rather than viewing explanations in pairs, and (2) participants rated just one explanation type per item (e.g., if a given participant rated a formal explanation for the *bird* item, they did not rate an inherent explanation for this item, and vice versa).

5.2. Results and discussion

As in Experiment 3, we conducted a multilevel mixed-effects linear model with cross-classified random intercepts for participants, items, and explanation wordings. The fixed effects included Explanation (a within-subject [level-1], dichotomous variable); Category Type (a within-subject [level-1], three-level categorical variable); the interaction of these two variables; and Science/Sophistication (a continuous, within-subject [level-1] covariate; the same as in Experiment 3).

We obtained a main effect for Explanation (M_{formal} = 1.77; $M_{inherent}$ = 2.84), Wald χ^2 = 35.58, $p < .001$. Neither the main effect for Category Type, Wald χ^2 = 3.89, p = .14, nor the Explanation × Category Type interaction, Wald χ^2 = 0.88, p = .64, were statistically significant. Furthermore, inherent explanations were judged to be more satisfying than formal explanations within each of the three types of categories (see Table 8 for means), all $ps < .001$. Finally, the Science/Sophistication covariate did not explain significant variance in ratings beyond the other predictors, b = 0.31, z = 1.21, p = .23.

Again, we find that inherent explanations are rated as more satisfying than formal explanations, even when controlling for ratings of the language used and even when explanations are judged individually. This result supports our argument that formal explanations are generally perceived as under-informative placeholders for explanations involving (inherent) features. It is also notable that overall, the ratings are lower in this study than in Experiment 3, consistent with the idea that the more absolute judgment engendered by presenting just one explanation at a time provides a more conservative test of participants' evaluations.

6. Experiment 5: Formal vs. inherent explanations for principled and statistical features

To this point, we have found that formal explanations, though natural-sounding, are not judged to be very satisfying (Experiment 1), and indeed are less satisfying than inherent explanations (Experiments 2–4). Experiment 5 examines the explanatory value of formal and essentialist explanations in a new way, by varying the type of feature that is being explained. Specifically, we contrast principled connections (involving features that are aspects of that category; e.g., "Dogs are four-legged") with statistical connections (involving features that are tied to the category simply by virtue of being frequent; e.g., "Dogs wear collars"; see Prasada &

Dillingham, 2006, 2009, for more discussion).[7]

On our account, formal explanations are likely to be elaborated via inherent properties particularly in cases where formal explanations are sensible—namely, for properties with principled connections to the kind. In contrast, we may not see an advantage for inherent explanations when the properties have a statistical connection to the kind (for which formal explanations are less sensible in the first place). We thus predict a two-way interaction between property type (principled vs. statistical) and explanation type (formal vs. inherent).

Our account makes a further prediction as well—namely, judgments of whether an explanation is *natural* will pattern differently from judgments of whether it is *satisfying*. Both formal and inherent explanations are likely to sound less natural for statistical than principled properties. Thus, the Property Type × Explanation interaction predicted above should be weaker or absent for naturalness judgments. Because Experiments 1–4 focused exclusively on principled features, we have not yet tested these predictions.

6.1. Method

6.1.1. Participants

Two-hundred sixty-one MTurk workers participated: 124 men, 137 women; M_{age} = 36.2 years, range 18–83; 200 White/Caucasian, 20 Black/African-American, 25 Asian/Asian-American, 9 Latino/Hispanic, 5 Multiracial, 1 Native American, 1 Other. The sample size was larger here than in previous studies to maintain sufficient statistical power given that our main manipulations were between subjects and that our main prediction was of a three-way interaction.

6.1.2. Design

The study had a 2 (Feature: Principled, Statistical) × 2 (Judgment: Natural, Satisfying) × 2 (Explanation: Formal, Inherent) × 3 (Category Type: Animals, Plants, Substances) design, with Feature and Judgment as between-subjects factors, and Explanation and Category Type as within-subject factors.

6.1.3. Materials and procedure

There were 24 feature/kind pairs (12 principled, 12 statistical), selected from the norming study described in Experiment 2 (see Experiment 2 for details about the method of this norming study, and Table 4 for the full set of ratings). On a scale of 1 to 7, where higher numbers indicated more principled connections, the 12 principled properties had a mean rating of 4.61 (SD = 0.31), and the 12 statistical properties had a mean rating of 3.07 (SD = 0.40).

Each item included a question/explanation pair, followed by a 1–7 scale on which the participant judged how satisfying or natural the explanation was, from 1 (not at all satisfying/natural) to 7 (extremely satisfying/natural). For example, "Q: Why does that [*pointing to a bird*] fly?" "A: Essentially, because it is a bird." To align our methods with those used by Prasada and Dillingham (2006), items focused on an individual member of the category rather than the category as a whole (e.g., a single bird vs. birds), and the questions did not include the category label. Note that this latter aspect of the design provides a particularly stringent test of our hypotheses, because it means that formal explanations provide a label that otherwise may have been uncertain or unknown.

The explanations consisted of the three formal and three inherent explanations from Experiments 3 and 4. The explanations were rated one at a time, as in Experiment 4; however, participants rated two explanations for each item: one formal and one inherent, randomly chosen from the three variants. The two explanations for a given item appeared simultaneously, on the same screen as the question. The order of the item/explanation pairs was counterbalanced across participants.

Participants were randomly assigned to one of four experimental conditions formed by the crossing of Feature Type (principled vs. statistical) and Judgment (natural vs. satisfying).

6.2. Results and discussion

We conducted a multilevel mixed-effects linear model with cross-classified random intercepts for participants, items, and explanation wordings. The fixed effects included Feature and Judgment (both between-subjects [level-2], dichotomous variables); Explanation (a within-subject [level-1], dichotomous variable); Category Type (a within-subject [level-1], three-level categorical variable); all possible interactions of these four variables; and Science/Sophistication (a continuous, within-subject [level-1] covariate; again using the composite variable from Experiment 3).

All main effects and interactions involving Judgment, Explanation, and Feature were significant, $ps < .01$, including the predicted three-way interaction involving Judgment, Explanation, and Feature, Wald $\chi^2 = 14.95$, $p < .001$.[8] This three-way interaction is best understood by comparing the two-way Explanation (formal vs. inherent) × Feature (principled vs. statistical) interaction

[7] Note that these data do not speak to the larger question of how statistical features relate to explanatory power. Evidence indicates that statistical information is one relevant factor in determining explanatory power (e.g., Colombo, Bucher, & Sprenger, 2017; Schupbach & Sprenger, 2011). In this experiment, we use statistical features merely as a methodological tool to test the selectivity of formal explanations, by examining how they contrast with principled features.

[8] Additionally, there were several effects involving category type. However, because we had no a priori predictions regarding category type, and because this factor did not interact with the key three-way interaction (Judgment × Explanation × Feature), we do not report the category type effects here.

Table 9
Mean naturalness and satisfyingness ratings in Experiment 5, as a function of feature type, judgment, and explanation. SDs are in parentheses. The cells that are the focus of the predicted interaction are highlighted with italics.

	Principled features		Statistical features	
Explanation:	Natural	*Satisfying*	Natural	*Satisfying*
Formal	4.51 (1.73)	*2.63 (1.77)*	3.52 (1.84)	*2.67 (1.73)*
Inherent	3.53 (1.79)	*2.97 (1.85)*	2.38 (1.65)	*2.32 (1.71)*

within each level of the Judgment variable (satisfying vs. natural).

We first consider participants' ratings of how *satisfying* the explanations were. Here, we predicted that inherent explanations would be more satisfying than formal explanations, but primarily for principled features; the advantage of inherent (over formal) explanations should disappear for statistical features (to which formal explanations should in principle not apply), leading to an Explanation × Feature interaction. This interaction was indeed significant, Wald $\chi^2 = 52.70$, $p < .001$, and driven by the expected pattern of means: Inherent explanations were marginally more satisfying than formal ones only for principled features ($p = .067$); for statistical features, the inherent explanations were actually *less* satisfying ($p < .001$), which we did not predict a priori but does not contradict our account (see Table 9).

In contrast, participants' ratings of how *natural* the explanations were revealed only a marginal Explanation × Feature interaction, Wald $\chi^2 = 2.90$, $p = .089$ (hence the three-way Explanation × Feature × Judgment interaction). Formal explanations sounded more natural than inherent ones for both principled and statistical properties ($ps < .001$; see Table 9).

Also note that when collapsing over feature type, both inherent and formal explanations were judged to be natural at higher rates than they were judged to be satisfying, $ps \leq .029$. This is consistent with the results of Experiment 1, and further makes the point that how natural an explanation sounds cannot be used as a proxy for how satisfying that explanation is.

Altogether, these findings support the hypothesis that formal explanations signal that some inherent property of the category explains the (principled) feature in question (e.g., the capacity of a bird to fly), a placeholder that in the case of natural kinds can then be filled in with a mechanism tied to the causal essence of the category.

7. Experiment 6: Hierarchical relation between formal and inherent explanations

The final three experiments assess the hypothesized relation between formal and inherent explanations—namely, that the formal explanations serve as a placeholder, to be filled in by some sort of inherent feature. That is, we hypothesized that people would construe the relation between formal and inherent explanations as *hierarchical*, with formal explanations as more general and inherent explanations as more specific and informative.

7.1. Method

7.1.1. Participants

Fifty-three MTurk workers participated: 28 men, 25 women; $M_{age} = 35.9$ years, range 20–69; 43 White/Caucasian, 6 Black/African-American, 2 Asian/Asian-American, 1 Latino/Hispanic, 1 Multiracial.

7.1.2. Materials and procedure

There were 12 items, corresponding to the 12 principled connections employed in Experiment 5. Each item included a question (e.g., "Why does that [*pointing to a bird*] fly?") followed by a two-part answer consisting of a formal explanation and an inherent explanation, in counterbalanced order within participants. The formal explanation was always "because it is a(n) X"; the inherent explanation was always "because of deep internal features." Between the two explanations was a blank, and participants' job was to choose which phrase belongs in the blank: either "More specifically" or "More generally." Below is an example of how a question was formatted:

Q: Why does that [*pointing to a bird*] fly?.
A: It flies because of deep internal features.
_____, it flies because it is a bird.

Choose which phrase belongs in the blank above:
◯ More specifically
◯ More generally

7.1.2.1. Post-test. At the end of the experiment we included a set of four trials to determine if participants understood "More specifically" and "More generally" literally—that is, as they applied to pairs of hierarchically organized categories (i.e., cats/animals, wolves/animals, daisies/plants, tulips/plants). For example, on one item, participants might read, "Daisies need sunlight. _____, many plants on earth need sunlight." We counterbalanced the order of the sentences within each post-test item, such that each participant received two specific-to-general (one animal, one plant) and two general-to-specific (one animal, one plant) items.

7.1.3. Design

To summarize, the study had a 2 (Order of Explanations: Formal-first vs. Inherent-first) × 3 (Category Type: Animals, Plants, Substances) repeated-measures design.

7.2. Results and discussion

The post-test trials indicated that participants had no difficulty understanding the literal meanings of "more specifically" and "more generally". When the blank preceded the more specific category (e.g., cats), participants selected "more specifically" 80% of the time, and when the blank preceded the more general category (e.g., animals), participants selected "more specifically" only 16% of the time. A mixed-effects multilevel logistic regression with cross-classified random intercepts for participants and items indicated this difference was statistically significant, Wald $\chi^2 = 43.00$, $p < .001$.

For the primary data, we conducted a similar mixed-effects multilevel logistic regression with the following predictors: Order (Formal-first vs. Inherent-first), which was a within-subject (level-1), dichotomous variable; Category Type (animal vs. plant vs. substance), which was a within-subject (level-1), three-level categorical variable; and their interaction. We found a very strong tendency for people to report that inherent explanations were more specific versions of formal explanations, Wald $\chi^2 = 102.83$, $p < .001$. That is, when the blank preceded the inherent explanation, people selected "more specifically" 78% of the time, whereas when the blank preceded the formal explanation, people selected "more specifically" only 36% of the time. There was no effect or interaction involving Category Type, $ps > .80$. Thus, people judged the relation between the two types of explanations as involving a consistent hierarchical relation, with formal explanations broader than and encompassing inherent explanations.

We also conducted a supplementary analysis on those participants who performed well on the post-test trials (answering correctly on at least 3 of the 4 trials; $n = 40$ out of the original 53). These data replicate those with the overall sample, with a main effect of Order, Wald $\chi^2 = 82.61$, $p < .001$, and no effects involving Category Type, $ps > .90$. When the blank preceded the inherent explanation, these participants selected "more specifically" 78% of the time, whereas when the blank preceded the formal explanation, they selected "more specifically" 34% of the time.

At the end of the experiment, participants were invited to explain why they answered as they did. One participant spelled out the logic of his responses as follows: "I felt that the 'deep internal features' represented a more specific, or precise, answer to the question of 'why' and [so] thought that should be associated with 'more specifically.' Simply saying something is some way because that's what it is a more general response—less specific and precise—and therefore was a 'more generally.'" This response articulates the logic of our "placeholder" account of formal explanations.

8. Experiment 7: Hierarchical relation between formal and inherent explanations

Experiment 7 provided a replication of Experiment 6, with two methodological changes: we provided each participant with just one linking phrase ("more generally" or "more specifically") rather than two, and we had them rate (on a 1–7 scale) how well the phrase belonged in the blank that appeared in each item. This design provides a more conservative test of the hypothesis that people would construe inherent explanations as more specific instantiations of formal explanations. By rating just one connector instead of having a binary choice, participants can report that the connector is equally satisfying regardless of the order of the explanations.

8.1. Method

8.1.1. Participants

One-hundred-eight MTurk workers participated: 49 men, 59 women; $M_{age} = 32.8$ years, range 20–71; 81 White/Caucasian, 8 Black/African-American, 11 Asian/Asian-American, 4 Latino/Hispanic, 2 Multiracial, 2 Native American.

8.1.2. Design

The study had a 2 (Connecting Phrase: "More Specifically" vs. "More Generally") × 2 (Order of Explanations: Formal-first vs. Inherent-first) × 3 (Category Type: Animals, Plants, Substances) design. Connecting Phrase was a between-subjects variable; Order and Category Type were within-subjects variables.

8.1.3. Materials and procedure

There were 12 items, corresponding to the 12 principled connections employed in Experiment 6. Each item included a question (e.g., "Why does that [*pointing to a bird*] fly?") followed by a two-part answer consisting of a formal explanation and an inherent explanation, in counterbalanced order within participants. As in Experiment 6, the formal explanation was "because it is a(n) X"; the inherent explanation was "because of deep internal features." Between the two explanations was a blank, and participants' job was to rate a phrase (either "More specifically" or "More generally", as a between-participants factor) for how well it fits in the blank, from 1 (not at all) to 7 (very much).

The post-test was identical to that of Experiment 6, except that again each participant rated a single linking phrase on a 1–7 scale.

8.2. Results and discussion

The post-test trials indicated that participants had no difficulty understanding the literal meanings of "more specifically" and "more

generally." We conducted a mixed-effects multilevel linear regression with cross-classified random intercepts for participants and items, and the following predictors: Connecting Phrase ("More Generally" vs. "More Specifically"), which was a between-subject (level-2), dichotomous variable; Order (Specific-first vs. General-first), which was a within-subject (level-1), dichotomous variable; and their interaction. This analysis indicated a main effect of Connecting Phrase, Wald $\chi^2 = 9.07$, $p = .003$, as well as a significant Connecting Phrase × Order interaction, Wald $\chi^2 = 60.83$, $p < .001$. As expected, participants gave a higher rating to "more specifically" when it preceded the more specific category (e.g., cats; $M = 4.77$) than when it preceded the more general category (e.g., animals; $M = 3.61$), Wald $\chi^2 = 11.94$, $p < .001$. Conversely, and also as expected, participants gave a higher rating to "more generally" when it preceded the more general category than when it preceded the more specific category (*M*s = 5.36 vs. 3.96), Wald $\chi^2 = 57.96$, $p < .001$.

For the primary data, we conducted a similar mixed-effects multilevel linear regression with the following predictors: Connecting Phrase (as above); Order (as above); Category Type (animal vs. plant vs. substance), which was a within-subject (level-1), three-level categorical variable; and their interactions. We found, once again, that people reported that inherent explanations were more specific than formal explanations, as seen by a Connecting Phrase × Order interaction, Wald $\chi^2 = 26.96$, $p < .001$. Planned contrasts revealed that people rated "more specifically" higher when it preceded inherent explanations than when it preceded formal explanations (*M*s = 4.71 vs. 3.92), Wald $\chi^2 = 39.00$, $p < .001$. Participants' ratings of "more generally" were in the predicted direction but did not reach significance (*M*s = 4.04 and 3.91, when "more generally" preceded formal and inherent explanations, respectively), Wald $\chi^2 = 1.10$, $p = .29$. The model also revealed a main effect of Connecting Phrase, Wald $\chi^2 = 13.14$, $p < .001$, and a main effect of Category Type, Wald $\chi^2 = 12.64$, $p < .002$. None of the other effects were significant.

Finally, we conducted a supplementary analysis on just those participants who performed well on the post-test trials (i.e., provided a higher mean rating for the connecting phrase in the "correct" context than in the "incorrect" context). For example, to be included, a participant rating the "more specifically" phrase had to give a higher mean rating when it preceded the more specific category (e.g., cats) than when it preceded the more general category (e.g., animals). Note that this was a difficult task, and only 56 out of 108 participants could be included in this supplementary analysis ($N = 26$ who rated "more specifically", and $N = 30$ who rated "more generally"). These data replicate those with the overall sample, with a significant Connecting Phrase × Order interaction, Wald $\chi^2 = 15.91$, $p < .001$, as well as a main effect of Connecting Phrase, Wald $\chi^2 = 17.08$, $p < .001$. As when all participants were included, the effect is significant only for participants who received the "more specifically" wording.

It is not entirely clear why the results differed for the two kinds of connectors. One possibility is that this asymmetry reflects the order in which the two explanations typically appear in people's thought processes. That is, it may be easier to think about the relation between the two kinds of explanations when they appear in the canonical order, with the placeholder (formal explanation) preceding the elaboration (inherent explanation), than when they appear in the reverse order (inherent first, and then formal). This would be analogous to how people find it easier to reason with information (e.g., a sequence of events) whose structure matches (vs. mismatches) how that information is represented in semantic memory (e.g., conventional scripts; Kintsch, Mandel, & Kozminsky, 1977). Relatedly, this may be why participants provided higher ratings when they saw a "match" (formal-first, with the appropriate "more specifically" linking the two explanations) than in any other combination. In contrast, no asymmetry would be expected for the literal post-test trials, because the relation between basic and superordinate categories (e.g., cats, animals) is an unordered, logical one.

To summarize across Experiments 6 and 7, the data indicate that the relation between formal and inherent explanations is one of inclusion: inherent explanations are judged to be more specific instantiations or elaborations of formal explanations. This result obtained both when participants were asked to select which of two connecting phrases ("more specifically" vs. "more generally") is a better fit between the two sorts of explanations (Experiment 6), and when participants were asked to rate just a single connecting phrase (Experiment 7).

9. Experiment 8: Hierarchical relation between formal explanations and inherent vs. extrinsic explanations

Experiment 8 provided a test of our fourth prediction: namely, that formal explanations are judged to be placeholders specifically for inherent explanations—not for non-inherent, extrinsic explanations. Here, we included environmental (extrinsic) explanations as a contrast case to inherent explanations. Thus, half the items asked participants to rate the hierarchical link between formal and environmental explanations, rather than between formal and inherent explanations. These items allowed us to test whether formal explanations indiscriminately suggest any sort of more-informative explanation, or whether they typically point to inherent properties.

Another change introduced in Experiment 8 was that the wording explicitly asked whether a formal explanation was *a more specific or more general version* of the relevant non-formal explanation (inherent or environmental). That is, the hierarchical link between the two explanations was directly queried. This question format provides a more stringent test of the theorized hierarchical link between formal and inherent explanations.

9.1. Method

9.1.1. Participants

One-hundred-thirteen MTurk workers participated: 56 men, 56 women, and 1 who identified as genderless; $M_{age} = 34.8$ years, range 20–68; 80 White/Caucasian, 11 Black/African-American, 9 Asian/Asian-American, 9 Latino/Hispanic, 2 Multiracial, and 1 Native American.

9.1.2. Design

The study had a 2 (Trial Type: Formal/Inherent vs. Formal/Environmental) × 3 (Category Type: Animals, Plants, Substances) design. Trial Type and Category Type were within-subject variables.

9.1.3. Materials and procedure

There were 12 items, corresponding to the 12 principled connections employed in Experiments 6 and 7. Each item included a question (e.g., "Why does that [*pointing to a bird*] fly?") followed by two explanations (labeled Answer A and Answer B): a formal explanation and a non-formal explanation, in counterbalanced order within participants. For half the items presented to a participant, the non-formal explanation was inherent, and for half the items, the non-formal explanation was environmental. We counterbalanced which items were inherent vs. environmental across participants (e.g., for the item about the bird flying, approximately half of the participants received an inherent explanation, and the others received an environmental explanation). As in Experiments 6 and 7, the formal explanation was "because it is a(n) X," and the inherent explanation was "because of deep internal features." The environmental explanation was "because of its environment." For each item, participants were asked, "Is Answer A a more **specific** or a more **general** version of Answer B?" This question was accompanied by a 7-point scale, with the end-points labeled "Answer A is a more **specific** version of Answer B" and "Answer A is a more **general** version of Answer B."

The post-test was similar to that of Experiment 6, except that the format was revised to match that of the test trials. That is, the target sentences were listed as Sentence A (e.g., "Daisies need sunlight") and Sentence B (e.g., "Many plants on earth need sunlight"), and the question was, "Is Sentence A a more **specific** or a more **general** version of Sentence B?" (1–7 scale).

9.2. Results and discussion

Scores were first transformed such that higher values on the 1–7 scale always corresponded to rating the non-formal—inherent or environmental—explanations (for the test trials) or the basic-level sentences (for the post-test trials) as more specific. As in Experiments 6 and 7, we analyzed these data with mixed-effects multilevel linear regressions with cross-classified random intercepts for participants and items.

On the post-test trials, participants correctly judged the sentences including the more specific categories (e.g., daisies) to be specific versions of the sentences including the more general categories (e.g., plants), $M = 5.44$ (vs. the midpoint of 4), Wald $\chi^2 = 63.27$, $p < .001$.

For the primary data, our analysis included Trial Type (Formal/Inherent vs. Formal/Environmental; within-subject), Category Type (animal vs. plant vs. substance; within-subject), and their interaction as predictors. We found that participants judged inherent explanations to be specific versions of formal explanations ($M = 4.82$) significantly more often than they judged environmental explanations to be specific versions of formal explanations ($M = 4.22$), Wald $\chi^2 = 42.12$, $p < .001$. Ratings were significantly above the scale midpoint (4) for the Formal/Inherent trials, Wald $\chi^2 = 27.90$, $p < .001$, indicating that participants reliably judged inherent explanations to be more specific versions of formal explanations. This was not the case for the Formal/Environmental trials, Wald $\chi^2 = 1.89$, $p = .17$. No other effects reached significance in this analysis.

Finally, we conducted a supplementary analysis on just those participants who "passed" the post-test trials (i.e., provided an average rating above the midpoint of 4). For example, to be included, a participant had to (on average) rate the sentences with the more specific categories (e.g., daisies) as more specific compared to the sentences with the more general categories (e.g., plants). Eighty out of 113 participants (71%) scored well enough to be included in this supplementary analysis. The results on this subset of participants replicated those with the full sample, with a significant difference between Formal/Inherent trials ($M = 4.95$) and Formal/Environmental trials ($M = 4.33$), Wald $\chi^2 = 32.17$, $p < .001$. Also as in the full sample, only the Formal/Inherent scores were significantly above the midpoint of 4, Wald $\chi^2 = 27.71$, $p < .001$, again indicating that participants consistently judged inherent explanations to be specific versions of the formal ones. The Formal/Environmental scores did not differ significantly from the midpoint, Wald $\chi^2 = 3.23$, $p = .072$.

To summarize across Experiments 6, 7, and 8, the data indicate that the relation between formal and inherent explanations is one of inclusion: inherent explanations are judged to be more specific instantiations of formal explanations. This result obtained when participants were asked to select which of two connecting phrases ("more specifically" vs. "more generally") is a better fit between the two sorts of explanations, when participants were asked to rate just a single connecting phrase, and when participants were asked to explicitly judge whether an inherent explanation is a more specific version of the formal explanation. Importantly, Experiment 8 also demonstrates the specificity of this effect: Formal explanations can more easily be elaborated as inherent explanations than environmental explanations.

10. General discussion

Formal explanations, which appeal to a category to explain an observation (e.g., "Mittens has whiskers *because she's a cat*"), pose an interesting paradox: they are commonplace and natural-sounding, but they nonetheless seem tautological (we already know that Mittens is a cat, so how does this tell us anything new?). Why, then, are they produced at all? We proposed that the solution to this puzzle is that formal explanations promote the search for more-informative explanations that appeal to inherent features of the relevant categories (e.g., "because of some to-be-determined, inherent feature of cats"). In other words, when people explain something by appealing to a category (e.g., Coley & Vasilyeva, 2010; Prasada, 2017; Prasada & Dillingham, 2006, 2009; Roberts et al., 2017; Taylor et al., 2009), they aren't merely invoking the category per se as an explanation; rather, this link to the category

suggests a mediating link between the observation and the category—a link to be filled in by some "deeper," inherent feature. Formal explanations are judged to be natural because they point to this (unspecified) inherent account. Thus, the puzzle that formal explanations are so natural while also apparently so empty reflects their dual nature as heuristically useful, but explanatorily dissatisfying (at least as literally stated, without further inferences). If one were to ask whether formal explanations are good or bad, the answer is that they are both. They are good in the sense of signaling, "There's more here to uncover", but they are bad in the sense of failing to explicate what that 'more' is.

We conducted a series of eight experiments that test four key predictions that follow from the central hypothesis articulated above: First, formal explanations may be natural-sounding but they were predicted to be relatively unsatisfying; second, formal explanations were predicted to be less satisfying than inherent explanations, particularly for properties with principled connections; third, we predicted that people would judge inherent explanations to constitute more specific versions of the ideas expressed by formal explanations; and fourth, we predicted that this relation would be specific to inherent explanations, and would not hold for non-inherent explanations.

All four predictions were supported. In their intuitive ratings, people judged formal explanations to be natural, but not particularly satisfying (Experiment 1). The remaining studies focused on the domain of natural kinds and compared formal explanations with inherent explanations that appealed to internal, structural properties—the sorts of properties that people typically view as being tied to the kind's essence (Gelman, 2003; Rhodes & Mandalaywala, 2017). We found that these inherent explanations were judged to be more satisfying than formal explanations (Experiment 2), even controlling for superficial aspects of the language of the explanations that were provided (i.e., how scientific and sophisticated it sounded; Experiments 3 and 4). Importantly, however, these patterns of results obtained only for stimuli involving principled connections to a kind, which admit formal explanations (Experiment 5). In contrast, there was no advantage for inherent explanations when the stimuli described statistical connections to a kind. Finally, Experiments 6, 7, and 8 provided critical tests of the hypothesized relation between formal and inherent explanations. In all three experiments, people viewed formal and inherent explanations as related to one another in an inclusion relation: inherent explanations are more specific instantiations of the placeholder-like formal explanations. In contrast (Experiment 8), environmental explanations were not viewed as having a hierarchical link to formal explanations. Altogether, these findings suggest a novel theoretical account in which explanation contributes to understanding not strictly in terms of the informational content that is conveyed, but also as a heuristically useful prompt to search for an inherent link between feature and category. Often this will comport with a causal reductionist account (Reutlinger, 2017), in which a seemingly non-causal explanation (in this case, formal) can ultimately be understood as a causal explanation.

In addition to what they tell us about formal explanations, these findings have implications for debates regarding the role of essentialism in conceptual structure. In an influential set of papers, Strevens (2000, 2001) has proposed that essences are not a necessary component of how people conceive of natural kinds and their features. Instead, Strevens argued, people explain the link between a kind and its features simply by appealing to the kind itself as a causal force, for example, they "might think that it is just a brute fact about the world that being a tiger causes an animal to grow stripes" (Strevens, 2000, p. 154). This so-called "minimalist" hypothesis has been the source of some debate (e.g., Ahn et al., 2001; Cimpian & Salomon, 2014; Meyer, Leslie, Gelman, & Stilwell, 2013). Adding to this debate, the present findings reveal a potential problem with this hypothesis: appeals to the kind as an explanation often are a just a placeholder for some more-detailed causal mechanism that involves the kind's inherent essence. Thus, the minimalist position may not accurately describe how people (or at least educated adults) reason about the features of natural kinds—people don't seem to actually attribute the presence of a kind's features to the kind itself. To the extent that our studies provide evidence against Strevens's minimalist hypothesis, they also reinforce the psychological reality of essences and their central place in human conceptual structure.

Although we have been emphasizing that formal explanations are placeholders for inherent explanations, we do not wish to suggest that formal explanations are unimportant. To the contrary, we believe they play a key role in understanding human reasoning. First, that formal explanations are so common is an important reminder that kinds are central to how people reason about the world. This conclusion is also supported by a wealth of data on the centrality of kinds to everyday language, in the form of generic nouns (e.g., Brandone, Cimpian, Leslie, & Gelman, 2012; Gelman, Coley, Rosengren, Hartman, & Pappas, 1998; Leslie, 2008), and on the centrality of kinds in early development (e.g., Cimpian, 2016; Ferry, Hespos, & Waxman, 2013; Gelman, 2003; Markman, 1989). Second, if the present position is correct, then formal explanations provide a pathway toward inferring inherent causes, including essences, when making sense of the world. Having this means of suggesting essences is important because people's knowledge of essences is often so vague (leading to the notion of an "essence placeholder"; Gelman, 2003; Medin, 1989; Medin & Ortony, 1989). The covert way in which formal explanations prompt the listener to "look further," in tandem with the explanatory processes that this prompt sets into motion, may well contribute to the pervasive tendency to essentialize—to attribute surface features to the workings of an inner category essence. Third, and related to the second point, formal explanations might also provide an important means of promoting the process of learning from others across development. That is, formal explanations might be another linguistic device (alongside generics; Gelman et al., 2010) by which essentialist ideas are conveyed from adults to young children. An interesting direction for future research would be to examine the inferences that children make upon hearing formal explanations regarding novel properties, and the extent to which such explanations license essentialist implications.

We have focused on formal and inherent explanations in the domain of natural kinds, but we suspect that a comparable relation between formal and inherent explanations can be found for other sorts of entities (e.g., social categories). Extending the present analyses to other domains is an exciting direction for future research (see Supplementary Experiment in the online materials for preliminary evidence). At the same time, it is also clear that formal explanations do not always serve as placeholders; their value

depends on the domain under consideration. For example, even if formal explanations are incomplete when reasoning about why dogs bark or why lions are aggressive, they may be fully satisfying when reasoning about other sorts of concepts, such as why circles don't have angles, which is stipulated by definition. Accordingly, it is important not to throw out the baby with the bathwater, so to speak; although we claim that formal explanations are often promissory notes, these explanations nevertheless provide important insights into human cognition.

Acknowledgements

We thank Antonio Malkoun, Pragya Mathur, and Abigail Tzau for their assistance with data coding, and Joseph Cimpian for help with statistical analyses.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.cogpsych.2018.08.002.

References

Ahn, W. K., Kalish, C., Gelman, S. A., Medin, D. L., Luhmann, C., Atran, S., et al. (2001). Why essences are essential in the psychology of concepts. *Cognition, 82*(1), 59–69.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*(4), 390–412.

Brandone, A. C., Cimpian, A., Leslie, S. J., & Gelman, S. A. (2012). Do lions have manes? For children, generics are about kinds rather than quantities. *Child Development, 83*(2), 423–433.

Cimpian, A. (2016). The privileged status of category representations in early development. *Child Development Perspectives, 10*(2), 99–104.

Cimpian, A., & Keil, F. (2017). Preface for the special issue on the process of explanation. *Psychonomic Bulletin & Review, 24*(5), 1361–1363.

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*(5), 461–480.

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

Coley, J. D., & Vasilyeva, N. Y. (2010). Generating inductive inferences: Premise relations and property effects. *Psychology of Learning and Motivation, 53*, 183–226.

Colombo, M., Bucher, L., & Sprenger, J. (2017). Determinants of judgments of explanatory power: Credibility, generalizability, and causal framing. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. J. Davelaar (Eds.). *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 1806–1811). Austin, TX: Cognitive Science Society.

Ferry, A., Hespos, S., & Waxman, S. (2013). Non-human primate vocalizations support categorization in very young human infants. *PNAS, 110*(38), 15231–15235.

Gelman, S. A. (2003). *The essential child: Origins of essentialism in everyday thought.* New York: Oxford University Press.

Gelman, S. A., Coley, J. D., Rosengren, K., Hartman, E., & Pappas, A. (1998). Beyond labeling: The role of maternal input in the acquisition of richly-structured categories. In Monographs of the Society for Research in Child Development. Serial No. 253, Vol. 63, No. 1.

Gelman, S. A., & Rhodes, M. (2012). "Two-thousand years of stasis": How psychological essentialism impedes evolutionary understanding. In K. S. Rosengren, S. Brem, E. M. Evans, & G. Sinatra (Eds.). *Evolution Challenges: Integrating research and practice in teaching and learning about evolution.* Cambridge: Oxford University Press.

Grice, H. P. (1975). Logic and conversation. In P. Cole, & J. Morgan (Vol. Eds.), *Syntax and Semantics: vol. 3*Academic Press.

Haward, P., Wagner, L., Carey, S., & Prasada, S. (2017). *The development of principled connections and kind representations.* Ms. under review: Harvard University.

Kintsch, W., Mandel, T. S., & Kozminsky, E. (1977). Summarizing scrambled stories. *Memory & Cognition, 5*(5), 547–552.

Knobe, J., Prasada, S., & Newman, G. E. (2013). Dual character concepts and the normative dimension of conceptual representation. *Cognition, 127*(2), 242–257.

Leslie, S. J. (2008). Generics: Cognition and acquisition. *Philosophical Review, 117*(1), 1–47.

Lewis, D. (1983). Extrinsic properties. *Philosophical Studies, 44*(2), 197–200.

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

Lombrozo, T. (2012). Explanation and abductive inference. In K. J. Holyoak, & R. G. Morrison (Eds.). *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*(2), 167–204.

Lombrozo, T., & Vasilyeva, N. (2017). Causal explanation. In M. Waldmann (Ed.). *Oxford handbook of causal reasoning* (pp. 415–432). Oxford, UK: Oxford University Press.

Markman, E. M. (1989). *Categorization and naming in children: Problems of induction.* Cambridge, MA: MIT Press.

Medin, D. L. (1989). Concepts and conceptual structure. *American Psychologist, 44*(12), 1469.

Medin, D. L., & Ortony, A. (1989). Psychological essentialism. In S. Vosniadou, & A. Ortony (Eds.). *Similarity and analogical reasoning* (pp. 179–195). New York, NY: Cambridge University Press.

Meyer, M., Leslie, S. J., Gelman, S. A., & Stilwell, S. M. (2013). Essentialist beliefs about bodily transplants in the United States and India. *Cognitive Science, 37*(4), 668–710.

Prasada, S. (2017). The scope of formal explanation. *Psychonomic Bulletin and Review*. https://doi.org/10.3758/s13423-017-1276-x.

Prasada, S., & Dillingham, E. M. (2006). Principled and statistical connections in common sense conception. *Cognition, 99*(1), 73–112.

Prasada, S., & Dillingham, E. M. (2009). Representation of principled connections: A window onto the formal aspect of common sense conception. *Cognitive Science, 33*(3), 401–448.

Reutlinger, A. (2017). Explanation beyond causation? New directions in the philosophy of scientific explanation. *Philosophy Compass, 12*(2), e12395.

Rhodes, M., & Mandalaywala, T. M. (2017). The development and developmental consequences of social essentialism. *Wiley Interdisciplinary Reviews: Cognitive Science, 8*(4).

Roberts, S. O., Gelman, S. A., & Ho, A. K. (2017). So it is, so it shall be: Group regularities license children's prescriptive judgments. *Cognitive Science, 41*(S3), 576–600.

Salomon, E., & Cimpian, A. (2014). The inherence heuristic as a source of essentialist thought. *Personality and Social Psychology Bulletin, 40*(10), 1297–1315.

Sánchez Tapia, I., Gelman, S. A., Hollander, M. A., Manczak, E. M., Mannheim, B., & Escalante, C. (2016). Development of teleological explanations in Peruvian Quechua-speaking and U.S. English-speaking preschoolers and adults. *Child Development, 87*(3), 747–758.

Schupbach, J. N., & Sprenger, J. (2011). The logic of explanatory power. *Philosophy of Science, 78*(1), 105–127.

Schwartz, S. P. (1980). Natural kinds and nominal kinds. *Mind, 89*(354), 182–195.

Strevens, M. (2000). The essentialist aspect of naive theories. *Cognition, 74*(2), 149–175.

Strevens, M. (2001). Only causation matters: reply to Ahn et al. *Cognition, 82*(1), 71–76.
Taylor, M. G., Rhodes, M., & Gelman, S. A. (2009). Boys will be boys; cows will be cows: Children's essentialist reasoning about gender categories and animal species. *Child Development, 80*(2), 461–481.
Weatherson, B., & Marshall, D. (2017). Intrinsic vs. extrinsic properties. In E. Zalta (Ed.), The Stanford Encyclopedia of Philosophy. Available at: http://plato.stanford.edu/entries/intrinsic-extrinsic/.