Psychon Bull Rev (2017) 24:1478–1487
DOI 10.3758/s13423-017-1276-x

BRIEF REPORT

The scope of formal explanation

Sandeep Prasada[1]

Published online: 3 April 2017

Abstract The existence of multiple modes of explanation means that a crucial step in the process of generating explanations has to be selecting a particular mode. The present article identifies the key conceptual, as well as some pragmatic and epistemological, considerations that license the use of the *formal* mode of explanation, and thus that enter into the process of selecting and generating a formal explanation. Formal explanations explain the presence of certain properties in an instance of a kind by reference to the kind of thing it is (e.g. *That has four legs because it is a dog*). As such, this mode of explanation is intrinsically tied to kind representations and is applicable domain-generally. Although it is possible for formal explanation to apply domain-generally, for any given kind it is selective in its application, in that it can explain some, but not all, properties of the instances of a kind. It also appears that different types of properties can receive formal explanations across different domains. This article provides a sketch of a theory of the selectivity of formal explanation that results from the manner in which kinds of different types are distinguished. The present discussion also suggests how the mechanisms underlying formal explanations may contribute to the illusion of explanatory depth Keil (*Trends in Cognitive Sciences*, *7*, 368–373,2003), the operation of the inherence heuristic Cimpian & Salomon (*Behavioral and Brain Sciences*, *37*, 461–480, 2014a; *Behavioral and Brain Sciences*, *37*, 506–527,2014b), and psychological essentialism (Gelman, 2003).

✉ Sandeep Prasada
sprasada@hunter.cuny.edu

[1] Department of Psychology, Hunter College, CUNY, 695 Park Avenue, New York, NY 10065, USA

Keywords Explanation · Concepts · Kind representation

We are naturally interested in finding out why things are the way they are (Keil & Wilson, 2000). Moravcsik (1990, 1998) has argued that humans are essentially explanation-seeking and explanation-forming animals, rather than information-processing creatures. This idea aligns well with the theory-based view of concepts (Carey, 1985; Gopnik & Meltzoff, 1997; Keil, 1989; Murphy & Medin, 1985). According to this view, concepts possess explanatory structure and are embedded within theories with explanatory structure. As such, explanations and explanatory structure are not esoteric products or frills of advanced conceptual systems, but are at the core of human thought. In fact, our conceptual systems naturally make use of multiple modes of explanation (e.g., Keil, 1994; Kelemen, 1999; Kelemen & Rosset, 2009; Lombrozo, 2009; Lombrozo & Gwynne, 2014; Prasada & Dillingham, 2006). The existence of multiple modes of explanation has important implications for the process of explanation, because it means that a crucial step in the process of generating explanations has to be selecting a particular mode. The present article identifies the key conceptual considerations that license the use of the *formal* mode of explanation—a mode in which the properties of an instance of a kind are explained by reference to the kind of thing it is (e.g., *That has four legs because it is a dog*)—as well as some pragmatic and epistemological considerations that license this mode's use. Finally, the discussion suggests how the mechanisms underlying formal explanations may contribute to the illusion of explanatory depth (Keil, 2003), the operation of the inherence heuristic (Cimpian & Salomon, 2014a, 2014b), and psychological essentialism (Gelman, 2003).

Modes of explanation in common-sense conception

Aristotle distinguished four modes of explanation, each of which identifies a distinct aspect of the world as grounding the phenomena to be explained. For example, to explain why a chair burns, we can cite what the chair is constituted of, identifying the material factor (e.g., *That burns because it is made of wood*). On the other hand, to explain why a chair has the shape it does, we cannot cite what it is constituted of, but we can cite the function of chairs, identifying the teleological factor (e.g., *That has the shape it does so that a person could sit on it*). To explain why a chair has the shape it does, we can also cite the carpenter and his actions—the efficient factor (e.g., *That has the shape it does because the carpenter nailed the boards together that way*). Finally, we can also cite what the thing is to explain why it has the shape it does, identifying the formal factor (e.g., *That has the shape it does because it is a chair*).

A wealth of research has documented the early emergence and importance of a causal mode of explanation (e.g., Baillargeon, 2002; Carey, 1985; Gelman, 1990, Gelman 2003; Gopnik & Meltzoff, 1997; Keil, 1989), which, in Aristotelian terms, draws upon a combination of the material and efficient modes of explanation. Also, a large amount of data now suggest the early emergence and importance of a teleological mode of explanation (e.g., Keil, 1994; Kelemen, 1999; Kelemen & Rosset, 2009; Lombrozo, 2009; Lombrozo & Gwynne, 2014; Opfer & Gelman, 2001).

Recent research has shown that our conceptual systems also make use of a formal mode of explanation (Haward, Wagner, Carey, & Prasada, 2017; Prasada & Dillingham, 2006, 2009; Prasada, Khemlani, Leslie, & Glucksberg, 2013). Prasada and Dillingham (2006) proposed that our conceptual systems distinguish the properties of an instance of a kind that are determined by the kind of thing it is (e.g., having four legs for a dog) from properties that are not determined by the kind of thing it is (e.g., wearing a collar, for a dog). The former properties were dubbed *k-properties* and were said to have a *principled connection* to the kind. Prasada and Dillingham (2006) found that formal explanations that were given to explain the presence of k-properties in an instance of a kind were rated as being much better than formal explanations given to explain the presence of properties that do not have a principled connection to the kind, even if the properties are highly prevalent in the kind. For example, the explanation in Item 1 below was rated as being much better when it was used to explain the presence of a k-property in an instance (Item 2) than when it was used to explain the presence of a property that does not have a principled connection to the kind (Item 3).

(1) Because it is a dog.
(2) Why does that [pointing to a dog] have four legs?
(3) Why is that [pointing to a dog] wearing a collar?

This pattern of results was found for items involving natural, artifact, and social kinds, suggesting that formal explanations are available across content domains. Furthermore, analyses using items matched on the prevalence of the property to be explained within the kind in question displayed the same pattern of results, showing that the acceptability of formal explanations does not depend on the prevalence of the property in kind members, but on whether or not the property has a principled connection to the kind. In addition to supporting formal explanations, properties that have a principled connection to a kind were found to support normative expectations concerning the presence of the property (e.g., we think dogs *should* have four legs) and were also expected to generally be present in instances of the kind (e.g., we expect dogs to generally have four legs).

Prasada, Khemlani, Leslie, and Glucksberg (2013) extended our understanding of the scope of formal explanations by showing (i) that formal explanations are possible when there is a principled connection between a property and the kind, even if the property is true of only a minority of instances of the kind (e.g., laying eggs, for ducks), and (ii) that formal explanations are not possible for properties that have a *causal* but not a *principled* connection to the kind (e.g., carrying malaria, for mosquitoes). It thus appears that having a principled connection to the kind is both necessary and sufficient for licensing a formal explanation of a property.

Haward, Wagner, Carey, and Prasada (2015; Haward et al. 2017 The development of principled connections and kind representations, submitted) found that children as young as 4 years of age naturally produce formal explanations to explain why two instances of a kind share a property. In fact, in Haward et al.'s experiments, formal explanations were the most prevalent form of explanation produced, with children producing formal explanations more than twice as often as teleological and causal explanations, which were produced about equally frequently. Furthermore, children produced more formal explanations if the properties to be explained had a principled connection to a kind than if the property had a merely statistical connection to the kind. It thus appears that formal explanations are a natural and early-emerging part of our conceptual repertoire.

Nature and scope of formal explanation

Formal explanations explain the presence of certain properties in an instance of a kind by reference to the kind of thing it is. In so doing, formal explanations unify two deep characteristics of human cognition: (i) the pervasive tendency to think and talk about things as instances of kinds, and (ii) the drive to understand and explain. In identifying something as an instance of a kind, we understand some of its properties as being due to its being the kind of thing it is. Identifying something as an instance of a kind and explaining some of its properties in terms of its being the kind of thing it is are not two distinct

activities, but a single cognitive activity. In identifying something as an instance of a kind, we render some of its properties intelligible, and identifying the same thing in a different way renders different properties intelligible. For example, if we identify something as an object of some kind (e.g., a bowl), its shape is rendered intelligible. We understand it to have the shape it does because it is a bowl, and that this shape is non-arbitrary (Prasada, Ferenz, & Haskell, 2002). If, on the other hand, that same entity is identified as a hunk of wood, its shape is not rendered intelligible—identifying it as wood does not explain why it has the shape it does, but other properties are rendered intelligible (e.g., its burnability).

One important consequence of the intrinsic connection between formal explanation and kind representations is that formal explanation is domain-general in its applicability: For any and all kinds of entities, some of the properties of instances of the relevant kind are explained by the instance being the kind of thing it is. For example, we can use formal explanation for properties of animals (*That has four legs because it is a dog*), plants (*That is yellow because it is a dandelion*), artifacts (*You can sit on that because it is a chair*), imaginary beings (*That has a horn because it is a unicor*n), immaterial imaginary beings (*That can go through walls because it is a ghost*), immaterial geometric things (*That has three sides because it is a triangle*), and any other kind of thing we may choose to consider. This is not the case for other modes of explanation.

Causal explanation is limited to those kinds of things that are material, and thus does not apply to immaterial kinds of things, such as triangles or ghosts (Prasada & Dillingham, 2006). It makes no sense, for example, to try to give a causal explanation for the three-sidedness of triangles; in such cases, causal explanation does not get a foothold. Teleological explanation is limited to kinds of things that are viewed as having functions or ends, such as artifacts and the parts of living things (Keil, 1994). Adults do not think entities such as mountains and triangles are for anything, and thus we cannot explain their existence or their properties via teleological explanations. Formal explanations, in contrast, are available for any and all kinds of things, because they explain by reference to the kind of thing something is. The domain generality of formal explanation means that during the process of generating an explanation, one cannot rule out, on the basis of domain information, the possibility of formulating a formal explanation, even though domain information could be used to rule out the applicability of causal and teleological modes of explanation.

Though formal explanation can apply to any and all kinds, for any given kind it can only explain some of the properties of instances of that kind. As such, formal explanation is selective in its application. Previous research has shown that formal explanations are possible for properties that have a principled connection to a kind. Furthermore, properties that have a principled connection to a kind are represented as aspects of being the kind of thing in question (Prasada & Dillingham, 2009). However, previous research does not specify how we come to represent some properties, but not others, as having a principled connection to a kind and thus may receive formal explanations. The next section provides a sketch of a theory of the selectivity of formal explanations.

Explaining the selectivity of formal explanation

Which properties of an instance of a kind can receive a formal explanation? Logically, one must consider the possibility that there is no general account of the selectivity of formal explanation, because which properties of a kind may receive formal explanation are determined arbitrarily. If this were the case, one would have to learn which properties of a kind could receive formal explanations on a case-by-case basis, and no general account of the selectivity of formal explanation would be possible. It is clear, however, that the selectivity of formal explanation is not determined arbitrarily. Being four-legged is a k-property of dogs, cats, cows, horses, zebras, and lions, to name just a few animals for which this is the case; being eight-legged is a k-property of spiders; and having two wings is a k-property of birds. All of this hardly seems to be an accident or arbitrary. Furthermore, Prasada and Dillingham (2006) observed that the numbers and types of properties that have a principled connection to a kind appear to vary across domains. They noted, for example, that it is generally easier to find properties that have a principled connection to a natural than to an artifact kind. Furthermore, whereas color often has a principled connection for natural kinds, it rarely has one for artifact kinds. We would not expect to find such regularities if the properties of a kind that can receive formal explanation were arbitrarily determined. It thus appears that an account of the selectivity of formal explanation will require identifying the types of properties that kinds of different types license for formal explanation.

Because formal explanations are licensed by principled connections, an account of the selectivity of formal explanations must identify which types of properties have principled connections to kinds of different types. In principle, types of kinds may be identified independently of the types of properties they have principled connections to. Alternatively, types of kinds may have no independent characterization, but simply be identified by the types of properties they have principled connections to. The theory developed here is of the latter sort. The problem, then, is to identify the types of properties.

Given the intrinsic connection between kind representations and explanation, kinds of different types should be distinguished by principled connections to types of properties that are explanatorily significant. A natural suggestion is that kinds of different types are distinguished by having principled connections to one or more of the types of properties that

Aristotle identified as providing the bases for explanations. In addition to the Aristotelian explanatory factors' intuitive appeal, psychological and linguistic evidence has suggested their importance in accounting for a range of conceptual and linguistic phenomena (Moravcsik, 1998; Prasada, 1999; Prasada & Dillingham, 2006; Pustejovsky, 1995).

The selectivity of formal explanations is thus proposed to be due to kind representations intrinsically representing principled connections to one or more of the types of properties that play an explanatory role in the material, efficient, or teleological modes of explanation (i.e., what something is constituted of, what produced the thing, and the thing's functional capacities). It is important to note that hypothesizing principled connections to the types of properties that play an explanatory role in other modes of explanation does not depend on these other modes of explanation being available. For example, representing a principled connection between a kind and the parts or stuff of which it is constituted does not require that the material mode of explanation be available (i.e., that one can explain any properties of a thing by reference to what it is constituted of).

The possibilities for kind representations that arise under this proposal are schematized in Fig. 1a. Specific kind representations have a principled connection to one or more of the types of properties. In this account, no kind representations exist that do not have principled connections to one or more of these types of properties. In short, there are no bare kinds (i.e., kinds that aren't represented as some type of kind—as having a principled connection to one or more of these types of properties): Kind representations are intrinsically typed. Note also that kinds of different types are distinguished formally (by the numbers and types of properties they have principled connections to), rather than via explicit, hierarchically related kind representations.

Finally, each type of property to which a kind has principled connections may be realized in a number of ways (see Fig. 1b). For example, the property that plays a role in the material mode of explanation (m) may be realized as either the parts of which a given kind of thing is constituted or the stuff of which it is constituted. The property that plays a role in the efficient mode of explanation (e) may be realized by something of the same kind as the thing produced, or by something of a different kind from the thing produced. The property that plays a role in the teleological mode of explanation (t) may be an intrinsically monadic functional capacity that relates a thing to itself (e.g., the capacity for growth), an intrinsically dyadic functional capacity that relates two things (e.g., the capacity for perception), or an intrinsically triadic functional capacity, such as understanding or reasoning. Consequently, kinds of different types are distinguished on the basis of the types of properties they have principled connections to, as well as by the manner in which those types of properties are realized. This results in different numbers and types of properties being licensed for formal explanations in kinds of different types. Furthermore, the types of kinds that are generated in this manner map onto ontologically significant categories (e.g., living things, artifacts, plants, animals, people, etc.; see Prasada, 2000, and the next section). As such, the theory provides an account for Prasada and Dillingham's (2006) observation that the numbers and types of properties that have a principled connection to a kind appear to vary across content domains.

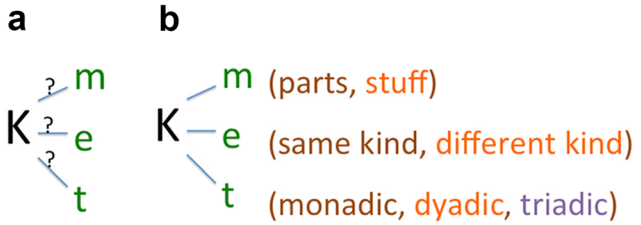

Fig. 1 (**a**) Schema for possible kind representations. Specific kind representations have a principled connection to one or more of the types of properties that have an explanatory role in the material, efficient, and teleological modes of explanation. (**b**) Properties that play an explanatory role in the material mode (**m**) may be the parts and/or the stuff of which a kind of thing is constituted. Properties that play an explanatory role in the efficient mode (**e**) may be something of the same kind as, or something of a different kind from, the thing produced. Properties that play an explanatory role in the teleological mode (**t**) may be intrinsically monadic functional capacities that relate a thing to itself (e.g., the capacity for growth), intrinsically dyadic functional capacities that relate two things (e.g., the capacity for perception), or intrinsically triadic functional capacities, such as understanding or reasoning

Illustrating the selectivity of formal explanation for different types of kinds

We can illustrate how the proposed account of selectivity captures differences in the types and numbers of properties that receive formal explanations by considering how kinds from ontologically distinct domains are represented. For example, for kinds of animals we represent principled connections to the properties that play an explanatory role in the material, efficient, and teleological modes of explanation. As a consequence, we can provide formal explanations for both the parts and stuff of which kinds of animals are constituted—for example, *That has a tail because it is a dog* or *That contains dog blood because it is a dog*.[1] We can also explain why an instance of the kind is produced by another instance of the kind via a formal explanation. Finally, we can explain an animal's dyadic functional capacities (e.g., capacities for perception and goal-directed movement) via a formal explanation (e.g., *That flies/swims/runs/hops because it is a bird/fish/cheetah/rabbit*; *That can see/echolocate because it is a dog/owl*).

[1] Though the dog blood example may feel like a tautology, it is not one. To see this, consider the following scenario: *Why does that have dog blood in it (pointing to a test tube)*?

We can contrast this with the manner in which artifact kinds are represented and the formal explanations they support. We represent principled connections to the properties that play an explanatory role in the material, efficient, and teleological modes of explanation for both animal and artifact kinds; however, these connections are realized differently. Consider the material component (that of which each kind is constituted). Though we represent principled connections to both the parts and stuff of which animal kinds are constituted, we do not represent a principled connection between artifact kinds and the stuff of which they are constituted (see Fig. 2).

This means that whereas we can provide formal explanations for both the parts and stuff of which animals are constituted, we can provide formal explanations only for the parts of which artifacts are constituted (e.g., *That has a seat because it is a chair*), but not for the stuff of which they are constituted (e.g., #*That is made of wood because it is a chair*). This contrast holds generally for animal versus artifact kinds.

It is important to note that when there is a principled connection between the kind and the stuff of which instances are constituted, the kind *determines* the stuff of which instances are constituted, rather than *merely constraining* the stuff of which instances are constituted. For example, being a dog determines the kind of flesh and blood it is constituted of; dogs are constituted of dog flesh and dog blood. On the other hand, being a radiation blanket does not determine the stuff that must constitute the blanket, though its kind information does constrain what stuff this may be. Radiation blankets are typically made of lead, but they do not have to be; they could be made of any other substance with the right causal properties. For similar reasons, generalizations such as *Microchips are made of silicon*, *Skyscrapers are made of steel*, *Windows are made of glass*, *Winter hats are made of wool*, and *Tires are made of rubber* do not provide counterexamples to the proposal that we do not maintain principled connections for the stuff that constitutes artifacts.[2] In all of these cases, it is not part of our understanding that instances of the kinds must be constituted of the materials mentioned. The kind only constrains the causal properties of the materials that may constitute any instances, and we are perfectly happy thinking of instances of these kinds being made of different kinds of materials (e.g., Plexiglas windows, tires made of some synthetic polymer, winter hats made of fur or fleece, etc.). In each of these cases, there is a causal but not a principled connection between the kind and the stuff that constitutes it. For example, it is unlikely that we think being made of wool is an aspect of being a winter hat, and unlikely we will think anything is wrong with a winter hat that is not made of wool, but of a causally appropriate material such as fur or fleece. Examples of this sort do not seem to involve principled connections, and thus are not expected to support formal explanations. Prasada et al. (2013) have provided evidence that formal explanations are not supported when there is a causal but not a principled connection to the kind.

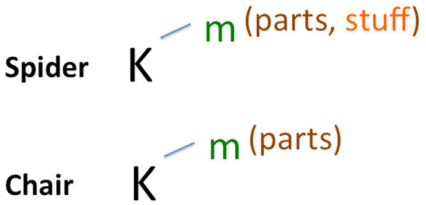

Fig. 2 Principled connections within the material component for the kind "spider" and the kind "chair"

The determination of stuff by animal kinds means that a number of stuff-related properties (e.g., texture, color, or smell) have a principled connection to the animal kind but do not have principled connections to artifact kinds. Thus, there should not be principled connections between artifact kinds and properties that are dependent on a *specific kind of stuff*—for example, the smell of leather or the feel of silk. Consequently, formal explanations such as *That smells like leather because it is a shoe* sound odd. Qualities that are related to stuff, but not to a specific kind of stuff (e.g., an unspecified kind of softness), however, can have a principled connection to an artifact kind if the qualities are required for the kind's function. For example, there is a principled connection between pillows and softness. Note, however, that the softness of the pillow does not have to be the softness of down, or feathers, or foam. Any of these will do. The softness of silk, on the other hand, is a specific kind of softness that is not found in other kinds of stuff. Such a quality cannot receive a formal explanation for an artifact kind, because artifact kinds do not have a principled connection to the kind of stuff of which they are constituted. Artifact kinds constrain, but do not determine, the stuff that constitutes them. Consequently, properties that depend on a specific kind of substance cannot be explained by reference to the artifact kind.

Because there is a principled connection between artifact kinds and the teleological component, we can provide formal explanations for the functional capacities of artifact kinds. We can also provide a formal explanation of why an instance of an artifact kind is made by an instance of another kind (human being) because of the principled connection to the efficient component.

These possibilities are not present, however, for kinds of natural objects such as mountains, because we do not represent a principled connection between mountains and the efficient or teleological components. We do not think that a specific kind of thing produces mountains: Some mountains may be produced by volcanoes, others by tectonic plates colliding, still others by rifting and erosion, or any combination of these forces. Consequently, formal explanations sound odd (e.g., #*That was produced by a volcano because it is a mountain*). Similarly, though canyons are usually

[2] I thank an anonymous reviewer and the editor for these examples.

understood to be produced by rivers, being produced by a river is not understood to be an aspect of being a canyon, and we can easily imagine that a canyon was produced by a glacier, an earthquake, or some other force or combination of forces. What matters to being a canyon is the nature of the formation that results, rather than what produced it. Similarly, a rainbow may be produced by light and a glass prism, or by raindrops or a plastic window. It does not much matter to us which of these produced the rainbow. Furthermore, in many cases we may be ignorant of what kinds of things produce a natural object. For example, I may not have any idea of what produces a rainbow, and yet still have the concept "rainbow." Furthermore, I do not need to have the expectation that one specific kind of thing produces rainbows. In this, natural objects differ from both artifact kinds and animal kinds, both of which involve principled connections to the efficient factor, so that we expect a specific kind of thing to produce them.[3] Natural objects also differ from artifacts and animals in that they do not have principled connections to a property that plays an explanatory role in the teleological mode of explanation.

These examples illustrate that ontologically meaningful distinctions correspond to types of kinds that are distinguished by the principled connections they have to the types of properties that are explanatorily relevant in different modes of explanation. Furthermore, the examples show how the selectivity of formal explanation across domains may be accounted for. Kinds in different ontological domains have principled connections to different numbers and types of explanatorily significant properties, and thus support different numbers and types of formal explanations. Further differences in the numbers and types of formal explanations arise from the different ways in which principled connections to a given type of property may be realized across domains. For example, a principled connection to the material component supports formal explanations of both the parts and stuff that constitute animal kinds, but only the parts of artifact kinds.

The discussion above is limited to examples of broad ontological categories, because these are the types of categories for which researchers have noted differences in the numbers and types of properties that have principled connections to their kinds. The theory proposed here is also capable of making finer-grained divisions between types of kinds. For example, the theory naturally distinguishes standard artifacts, such as chairs and ladders, from "informational objects," such as books and maps. The two types of artifact kinds can be distinguished with respect to the characteristics of the properties that realize their material and teleological components. Standard artifacts are understood to be constituted of concrete, perceptible matter, whereas informational objects are understood to be constituted of both concrete, perceptible matter and abstract, intelligible matter. As a consequence, standard artifact kinds are understood to have principled connections to intrinsically dyadic functional capacities (e.g., they are objects of sitting, climbing, and other actions). On the other hand, books, maps, and other "informational objects" also have principled connections to intrinsically triadic functional capacities (e.g., they are objects of understanding). The full range of kinds of different types that may be distinguished by the theory proposed above awaits a full specification of the ways in which each of the types of properties that have principled connections to a kind may be realized.

[3] That we understand rain to be produced by clouds appears to be a counterexample, because it seems to be part of our understanding of what rain is that it is produced by clouds. However, it is likely that rain is not conceived of as a kind of "natural object," but as a kind of natural occurrence or event. If so, this does not count as a counterexample. The present article is limited to a discussion of the selectivity of formal explanations for the properties of different kinds of things/objects. An important question for future research will be to investigate how kinds of events and occurrences are conceived of, and their relation to our conceptions of kinds of things.

Licensing formal explanations and over-hypotheses

The account of the selectivity of formal explanations presented above relies on the intrinsic typing of kinds by principled connections to different types of explanatorily significant properties. Kinds of different types license formal explanations for different numbers and types of properties, because kinds of different types have principled connections to different numbers of explanatorily significant properties and/or to how those types of properties may be realized.

This account shares similarities with, as well as potentially important differences from, theories of induction formulated in terms of over-hypotheses (Goodman 1983; Shipley, 1993). *Over-hypotheses* are projected hypotheses about the types of properties of kinds of things—for example, *Each kind of animal possesses a means of locomotion*. Goodman proposed that projections about individuals of a kind (e.g., dogs have four legs, birds have two wings) could gain entrenchment via relevant over-hypotheses. Shipley (1993) adapted Goodman's theory to show that over-hypotheses can allow inductive inferences about kinds of thing to be made on the basis of limited evidence. For example, given the over-hypothesis *Each kind of animal possesses a means of locomotion*, a learner could see just a few fish and make the inductive inference *Fish have fins* (Macario, Shipley, & Billman, 1990). One can imagine adapting over-hypotheses to develop an account of how principled connections between kinds and properties are acquired (Prasada & Dillingham, 2006). Like the account of the selectivity of formal explanations developed here, over-hypotheses identify kinds of different types as being systematically related to properties of different types. To provide an account of the selectivity of formal explanation, that systematic relation between types of kinds and types of properties would have to be limited to principled connections (rather

than statistical or causal connections—see Prasada et al., 2013). Existing theories of over-hypothesis formation and use are not constrained in this manner, and so can generate and make use of over-hypotheses, such as *Bagfuls of marbles are uniform in color* (Dewar & Xu, 2010; Kemp, Perfors, & Tenenbaum, 2007), that clearly do not involve principled connections. An important question for future research will be whether and how over-hypotheses and the mechanisms that generate them may be modified in this manner. The present account also differs from accounts based on over-hypotheses in that it proposes that kind representations are intrinsically typed (by principled connections to explanatorily significant types of properties), whereas accounts based on over-hypotheses do not propose that kind representations are intrinsically typed. On the other hand, they do assume explicit hierarchical relations between kind representations, whereas the present account does not. Though it is clear that the theory of the selectivity of formal explanations developed in this article bears important similarities to accounts of induction that make use of over-hypotheses, it remains for future research to determine whether an account of the selectivity of formal explanation can be formulated within the over-hypothesis framework. If such an account can be formulated, it will be important to determine the extent to which the similarities and differences between it and the present account are substantive and how they accord with relevant psychological data.

Some factors that enter into choosing to generate a formal explanation

Given the availability of multiple modes of explanation, any time we seek to generate an explanation, a mode of explanation must be selected. The choice is likely to be influenced by a variety of factors, including pragmatic, conceptual, and epistemological considerations. The theory of the selectivity of formal explanations presented above specifies conceptual conditions for licensing formal explanations. We turn now to briefly consider other factors that may enter into choosing to generate a formal explanation.

Pragmatic considerations such as the knowledge state of the hearer or the goals of an explainer can lead to one or another mode of explanation being preferred. If someone asks why a given thing has the shape it does, providing a formal explanation makes sense if you suspect that the person does not know what kind of object the thing is. If, on the other hand, you think the person does know that a thing is a chair, but may not know the function of chairs, it would make sense to avoid a formal explanation, but to choose a teleological explanation instead. Though no research has directly tested this claim, it follows from general pragmatic considerations that even preschoolers are sensitive to (Diesendruck, 2005).

Vasilyeva, Wilkenfeld, and Lombrozo (2015) provided evidence that people's current goals affect their evaluations of different types of explanations, such that explanations receive a relative boost when they provide support for goal-relevant inferences. For example, Vasilyeva et al. found that formal explanations were evaluated as being better explanations when participants had the goal of determining the category membership of items than when participants had no goal.

Perhaps the most important pragmatic consideration in choosing a mode of explanation is the type of question that the explanation is meant to address. Formal explanations can potentially be used to explain why one or why multiple instances of a kind have certain properties. Formal explanations cannot be used to explain why instance(s) of a kind or why the kind itself exists. They also cannot be used to explain *how* instance(s) of a kind come to have the properties they do.

Turning to the conceptual considerations that enter into choosing the formal mode of explanation, the key condition is that the property to be explained must have a principled connection to the kind in question, and thus be understood as an aspect of the relevant kind. As we saw above, whether or not a property is understood to have a principled connection to a kind depends on the type of kind under consideration.

Epistemological considerations concerning qualities such as the simplicity, depth, scope, and insightfulness of explanations may also be used to select one mode of explanation over another. The originators of the scientific revolution considered the formal and teleological modes of explanation to be inappropriate for conducting science (Henry, 1997; Kuhn, 1977). The formal mode of explanation presupposes a kind perspective, in which things are thought of as instances of kinds that have certain properties because they are the kinds of things they are. Instances that lack these properties are considered defective. Furthermore, variation among instances of a kind is understood to be accidental and a source of noise that is explanatorily insignificant. This perspective is at odds with certain modern scientific theories, including the theory of evolution (Gelman & Rhodes, 2012; Hull, 1965), but is intrinsic to both common-sense conception and language. As such, formal explanations may be chosen as part of everyday discourse, but they are not appropriate for conducting and discussing most modern scientific research.

Formal explanation and common-sense conception

Formal explanation is intrinsic to our capacity to think of things as instances of kinds. As such, it has a number of qualities: It is a *domain-general* mode of explanation that applies to any and all kinds of things. It is also *selective*, in that not all properties of an instance of a kind may be

accounted for via formal explanation. The types of properties that can receive formal explanation depend, in part, on the type of kind something is. Finally, formal explanation is *limited in scope*. It cannot provide explanations of the existence of particular instances of a kind, nor can it provide an explanation of *how* the properties of an instance of a kind come about. As such, the formal mode of explanation is used in conjunction with other explanatory modes within a nonmechanistic framework of understanding and explanation, in which the goal is to identify the aspect of the world that explains the phenomena of interest. This framework is relevant to the understandings expressed via our linguistic system.

In addition to explaining the characteristics of instances of kinds, formal explanations identify the phenomena that a number of alternative modes of understanding or explanation may seek to explain. For example, it is likely that *psychological essentialism* presupposes formal explanation of the properties that are explained causally by the underlying essence (in the domains where we posit essences). This is because psychological essentialism embodies a *kind* assumption as well as an *essence* assumption (Gelman, 2004). The kind assumption is "that people treat certain categories as richly structured 'kinds' with clusters of correlated properties" (Gelman, 2004, p. 408). This brief characterization of the kind assumption is problematic, however, because within the cluster of correlated properties, some properties are likely not to be good candidates for being understood as caused by an underlying essence. For example, the property of wearing collars is likely to be in the cluster of correlated properties for the kind "dog"; however, theories of psychological essentialism would presumably not want to claim that we understand this property to be caused by an underlying essence. The theory of kind representation and formal explanation developed here provides a better characterization of the kind assumption required by psychological essentialism: The properties that are understood to be true of instances of a kind by virtue of their being the kinds of things they are (i.e., the properties for which we can provide formal explanations) are also candidates for the properties that are understood to be caused by an underlying essence in the relevant domain. The theory of kind representation and formal explanation developed here does not provide support for or against psychological essentialism; however, it can provide a crucial element of the representation of concepts within the framework of psychological essentialism.

The intrinsic connection between kind representation and formal explanation of the properties that characterize and individuate kinds is also likely to contribute to the bias toward inherent explanations that is described by research on the *inherence heuristic* (Cimpian & Salomon, 2014a, 2014b). This is the case because, in accessing a kind representation, not only will the properties that have a principled connection to the kind be retrieved and highly accessible, so will the fact that they are explained by the kind in question, and thus a potential inherent explanation of the properties will also likely be highly accessible (Prasada, 2014). As with psychological essentialism, the theory of kind representation and formal explanation developed here does not provide support for or against the inherence heuristic, but the representations it makes available are likely to engender a bias toward inherent explanations, as is proposed in the inherence heuristic (Cimpian & Salomon, 2014a, 2014b)

The nonmechanistic formal explanations licensed by kinds of different types may also contribute to the *illusion of explanatory depth*, which is the finding that we believe we have more explanatory knowledge of how things work than we actually do have (Keil, 2003). According to Keil (2003), one factor involved in creating this illusion is that participants often have high-level functional explanations for phenomena; however, participants often do not realize that the explanatory understanding they have does not translate into being able to formulate the mechanistic explanations that are required. The formal mode of explanation provides an additional mode that can contribute to our sense of understanding. The form of understanding it provides, however, is completely inadequate for causal mechanistic explanations. As such, if participants fail to recognize the irrelevance of this form of understanding, it could contribute to the illusion of explanatory depth. This may be especially likely to occur because the formal mode of explanation identifies certain things as causally relevant (e.g., as having a principled connection to the thing that is the efficient cause), but it does not provide any causal–mechanistic knowledge. One way to test whether formal explanations contribute to the illusion of explanatory depth may be to investigate whether producing formal explanations of the properties of objects prior to participating in an illusion-of-explanatory-depth task increases the magnitude of the illusion for those properties.

An important question for future investigation concerns the relation of the formal explanatory knowledge embodied in kind representations to core systems of knowledge and to implicit explanatory knowledge in infancy (Baillargeon, 2002; Spelke & Kinzler, 2007). It will also be important to investigate how the formal explanatory knowledge embodied in kind representations is related to explicit mechanistic understandings that are constructed during development and that are subject to conceptual change (Carey, 2009).

References

Baillargeon, R. (2002). The acquisition of physical knowledge in infancy: A summary in eight lessons The acquisition of physical knowledge in infancy: A summary in eight lessons. In U. Goswami (Ed.), *Blackwell handbook of childhood cognitive development Blackwell handbook of childhood cognitive development* (pp. pp. 46–pp. 83). Oxford, UK: Blackwell.

Carey, S. (1985). *Conceptual change in childhood Conceptual change in childhood*. Cambridge: MIT Press.

Carey, S. (2009). *The origin of concepts The origin of concepts*. New York: Oxford University Press.

Cimpian, A., & Salomon, E. (2014a). The inherence heuristic: An intuitive means of making sense of the world, and a potential precursor to psychological essentialism The inherence heuristic: An intuitive means of making sense of the world, and a potential precursor to psychological essentialism. *Behavioral and Brain Sciences, 37,* 461–480.

Cimpian, A., & Salomon, E. (2014b). Refining and expanding the proposal of an inherence heuristic in human understanding Refining and expanding the proposal of an inherence heuristic in human understanding. *Behavioral and Brain Sciences, 37,* 506–527.

Dewar, K., & Xu, F. (2010). Induction, overhypothesis, and the origin of abstract knowledge: Evidence from 9-month-old infants Induction, overhypothesis, and the origin of abstract knowledge: Evidence from 9-month-old infants. *Psychological Science, 21,* 1871–1877.

Diesendruck, G. (2005). The principles of conventionality and contrast in word learning: An empirical examination The principles of conventionality and contrast in word learning: An empirical examination. *Developmental Psychology, 41,* 451–463.

Gelman, R. (1990). First principles organize attention to and learning about relevant data: Number and the animate-inanimate distinction as examples. *Cognitive Science, 14,* 79–106.

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

Gelman, S. A. (2004). Psychological essentialism in children Psychological essentialism in children. *Trends in Cognitive Sciences, 8,* 404–409. doi:10.1016/j.tics.2004.07.001

Gelman, S. A., & Rhodes, M. (2012). "Two-thousand years of stasis": How psychological essentialism impedes evolutionary understanding "Two-thousand years of stasis": How psychological essentialism impedes evolutionary understanding. In K. S. Rosengren, S. K. Brem, E. M. Evans, & G. M. Sinatra (Eds.), *Evolution challenges: Integrating research and practice in teaching and learning about evolution Evolution challenges: Integrating research and practice in teaching and learning about evolution* (pp. pp. 3–pp. 21). New York: Oxford University Press.

Goodman, N. (1983). *Fact, fiction, and forecast Fact, fiction, and forecast*. Cambridge: Harvard University Press. Original work published 1955Original work published 1955.

Gopnik, A., & Meltzoff, A. N. (1997). *Words, thoughts, and theories Words, thoughts, and theories*. Cambridge: MIT Press.

Haward, P., Wagner, L., Carey, S., & Prasada, S. (2017, March). *Principled connections in kind concepts*. Poster presented at the 2015 Biennial Meeting of the Society for Research in Child Development, Philadelphia, PA.

Henry, J. (1997). *The scientific revolution and the origins of modern science The scientific revolution and the origins of modern science*. New York: St. Martin's Press.

Hull, D. L. (1965). The effect of essentialism on taxonomy: Two thousand years of stasis. Part 1The effect of essentialism on taxonomy: Two thousand years of stasis. Part 1. *British Journal for the Philosophy of Science, 15,* 314–326.

Keil, F. C. (1989). *Concepts, kinds, and conceptual development Concepts, kinds, and conceptual development*. Cambridge: MIT Press.

Keil, F. C. (1994). The birth and nurturance of concepts by domains: The origins of concepts of living things The birth and nurturance of concepts by domains: The origins of concepts of living things. In L. A. Hirschfeld, S. A. Gelman, & S. A. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture Mapping the mind: Domain specificity in cognition and culture* (pp. pp. 234–pp. 253). Cambridge: Cambridge University Press.

Keil, F. C. (2003). Folkscience: Coarse interpretations of a complex reality Folkscience: Coarse interpretations of a complex reality. *Trends in Cognitive Sciences, 7,* 368–373.

Keil, F. C., & Wilson, R. A. (2000). Explaining Explanation Explaining Explanation. In F. C. Keil & R. A. Wilson (Eds.), *Explanation and cognition Explanation and cognition* (pp. pp. 1–pp. 18). Cambridge: MIT Press.

Kelemen, D. (1999). The scope of teleological thinking in preschool children The scope of teleological thinking in preschool children. *Cognition, 70,* 241–272.

Kelemen, D., & Rosset, E. (2009). The human function compunction: Teleological explanation in adults The human function compunction: Teleological explanation in adults. *Cognition, 111,* 138–143. doi:10.1016/j.cognition.2009.01.001

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning over hypotheses with hierarchical Bayesian models Learning over hypotheses with hierarchical Bayesian models. *Developmental Science, 10,* 307–321.

Kuhn, T. S. (1977). Concepts of cause in the development of physics. In *The essential tension: Selected studies in scientific tradition and change* (pp. 21–30). Chicago, IL: University of Chicago

Lombrozo, T. (2009). Explanation and categorization: How "why?" informs "what?"Explanation and categorization: How "why?" informs "what? *Cognition, 110,* 248–253.

Lombrozo, T., & Gwynne, N. Z. (2014). Explanation and inference: Mechanistic and functional explanations guide property generalization Explanation and inference: Mechanistic and functional explanations guide property generalization. *Frontiers in Human Neuroscience, 8,* 700. doi:10.3389/fnhum.2014.00700

Macario, J. F., Shipley, E. F., & Billman, D. O. (1990). Induction from a single instance: Formation of a novel category Induction from a single instance: Formation of a novel category. *Journal of Experimental Child Psychology, 50,* 179–199.

Moravcsik, J. M. E. (1990). *Thought and language Thought and language*. London: Routledge.

Moravcsik, J. M. E. (1998). *Meaning, creativity, and the partial inscrutability of the human mind Meaning, creativity, and the partial inscrutability of the human mind*. Stanford: CSLI.

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

Opfer, J. E., & Gelman, S. A. (2001). Children's and adult's models for predicting teleological action: The development of a biology-based model Children's and adult's models for predicting teleological action: The development of a biology-based model. *Child Development, 72,* 1367–1381.

Prasada, S. (1999). Names for things and stuff Names for things and stuff. In R. Jackendoff, P. Bloom, & K. Wynn (Eds.), *Language, logic, and concepts Language, logic, and concepts* (pp. pp. 119–pp. 146). Cambridge: MIT Press.

Prasada, S. (2000). Acquiring generic knowledge Acquiring generic knowledge. *Trends in Cognitive Sciences, 4,* 66–72.

Prasada, S. (2014). The representation of inherent properties The representation of inherent properties. *Behavioral and Brain Sciences, 37,* 500. doi:10.1017/S0140525X13003853

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

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

Prasada, S., Ferenz, K., & Haskell, T. (2002). Conceiving of entities as objects and as stuff Conceiving of entities as objects and as stuff. *Cognition, 83,* 141–165.

Prasada, S., Khemlani, S., Leslie, S.-J., & Glucksberg, S. (2013). Conceptual distinctions amongst generics Conceptual distinctions amongst generics. *Cognition, 126,* 405–422.

Pustejovsky, J. (1995). *The generative lexicon The generative lexicon*. Cambridge: MIT Press.

Shipley, E. F. (1993). Categories, hierarchies, and induction Categories, hierarchies, and induction. In B. H. Ross (Ed.), *The psychology of learning and motivation The psychology of learning and motivation* (Vol. 30, pp. pp. 265–pp. 301). San Diego: Academic Press.

Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge Core knowledge. *Developmental Science, 10,* 89–96.

Vasilyeva, N., Wilkenfeld, D., & Lombrozo, T. (2015). Goals affect the perceived quality of explanations Goals affect the perceived quality of explanations. In D. C. Noelle, R. Dale, A. S. Warlaumont, J. Yoshimi, T. Matlock, C. D. Jennings, & P. P. Maglio (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. pp. 2469–pp. 2474). Austin: Cognitive Science Society.