Explanation and Confirmation in Inductive Inference

Philip M. Fernbach

Prelim Examination

Brown University

June 13, 2008

Explanation and Confirmation in Inductive Inference

Inductive inference, reasoning from premises or observations to a degree of belief in an uncertain conclusion, is a basic capability of human cognition. Philosophers address two types of inductions, which Hempel (1965) referred to as *explanation-seeking 'why' questions* and *confirmation-seeking 'why' questions*. Explanation-seeking why questions ask when a hypothesis or set of propositions *explains* a set of observations, or in other words, why something is the case. Confirmation-seeking why questions ask when a set of observations *confirms* a hypothesis or set of propositions, or in other words, why one should believe that something is the case. Darwin's theory of evolution by natural selection *explains* why finches in the Galapagos have small beaks after years of heavy rain. Observation of small-beaked finches after heavy rain *confirms* Darwin's theory.

Explanation and confirmation are clearly related, but there is no consensus on the nature of the relation. Attempts to describe both in a unified framework like deductive logic have failed, and work on the two has been separate. Studies of confirmation focus on the syntactic or statistical properties of arguments and specify *a priori*, general rules for determining the degree of support that evidence lends to a hypothesis. Explanation-based approaches focus on the content of arguments, exploring questions like the features of good explanations. This divergence is exemplified by the debate between proponents of the two most influential approaches in contemporary philosophy of science, Bayesian Confirmation Theory (BCT) and Inference to the Best Explanation (IBE). In BCT, degree of belief in a hypothesis is represented as a conditional probability of the hypothesis given the evidence. The key tenet of IBE is that explanation guides inference. Hypotheses are accepted as true on the basis of how much understanding they provide. Lipton (1991)

describes this difference as a distinction between *likeliness* and *loveliness*. Likeliness, the target of confirmation theories, refers to how likely a hypothesis is to be true. Loveliness refers to its explanatory virtue, how much understanding it provides. Proponents of BCT claim that likeliness is judged and hypotheses that are likely are accepted as good explanations. Proponents of IBE claim that the loveliness of a hypothesis determines whether it is accepted as likely to be true. Thus both approaches are touted as complete accounts of inductive inference but with converse assumptions about the relation between explanation and confirmation.

Psychologists have explored inductive inference using property induction tasks. For example a person might be asked to the rate the strength of an argument like:

Bears have sesamoid bones
Therefore dogs have sesamoid bones

where the statement above the line is an observation and the statement below it is a supposition. To assess the strength of the argument the reasoner must decide the degree of confirmation that the observation lends to the supposition. Theories of induction mirror the dichotomy in the philosophy of science. On one hand are Bayesian theories that treat confirmation as fundamental and specify probabilistic computations that apply to all arguments irrespective of content. On the other hand are approaches that focus on the role of specific knowledge and explanation in mediating inductive strength. A review of these theories will show that explanation is central to inductive inference. When deciding whether to project a predicate from one category to another people try to explain the predicate, and inductive strength is mediated by the extent to which the explanation applies to the conclusion. Theories of confirmation do not capture this aspect of induction and are therefore inadequate.

This could be taken to imply that loveliness guides likeliness and not vice-versa, but such a conclusion is facile. These results cannot rule out Bayesian theories because the analysis begs the question of what an explanation is. On one interpretation BCT can be construed as a kind of inference to the best explanation because it chooses among hypotheses that can be used to explain. Some have argued on these grounds that BCT is superior to IBE because it is a specification of what it means to infer the best explanation. In fact, this is a comparison of apples and oranges. BCT specifies a computation based on the rules of probability but it does not stipulate the primitives for the calculations, the hypothesis space, prior probability distribution and likelihood function, which depend on the context of the inference. Conversely, IBE is concerned with describing the specific knowledge that makes a hypothesis explanatory, but is not specified. I conclude that people do engage in IBE when making inductive inferences. A determination of whether BCT is descriptively correct awaits a theory of explanation.

1.1 Problem of Induction

Hume (1748) taxonomized inference into two kinds, *relations of ideas* and *matters of fact*. The former are deductions, *a priori* inferences from logical implication. Sound reasoning from the premises of an argument is guaranteed to yield a true conclusion. Hume referred to these as relations of ideas because they pertain only to the syntax of arguments and are not contingent on whether the premises refer to real objects or events. Matters of fact are inductions, inferences from experience to truths about the world, and are not about syntactic consistency. Hume argued that matters of fact cannot be demonstrably certain because they must be justified by other matters of fact, which leads to a pernicious regress. Inferring that the sun will rise tomorrow on the basis that

has risen in the past necessitates assuming that the future will be like the past, but on what basis is that assumption justified? Hume's 'problem of induction' has two aspects, normative and descriptive. On the normative side philosophers want to know how to characterize good inferential practice. On the descriptive side, philosophers and psychologists have tried to describe how people make inductive inferences.

Subsequent work on the problem of induction has centered on two fundamental questions, confirmation and explanation. Confirmation asks when observations lend inferential support to a hypothesis. Explanation asks when a hypothesis provides understanding of observations. According to the logical empiricist approach to the philosophy of science, confirmation and explanation are flip sides of the same coin. Hempel's (1965) deductive-nomological model of explanation exemplifies the logical empiricist approach. According to the model, explanation is a relation of logical implication. An observation is explained by a set of rules from which it can be deduced. The logical structure of explanation implies a theory of confirmation, hypothetico-deductivism (Popper, 1959). A hypothesis is confirmed by instances that follow deductively from the rules that constitute the hypothesis and falsified by observations that are contrary to its implications.

The simplest example of confirmation by logical implication, referred to as enumerative induction, is that a positive instance of a law should confirm that law. The deductive approach to confirmation espoused by the logical empiricists has often been attacked by showing that even in this simple case, relying purely on syntactic relations for confirmation can lead to absurd conclusions. For example, Hempel (1945) introduced the 'raven paradox:' Observing a black raven confirms the generalization that *all ravens*

are black. This generalization is logically equivalent to its contraposition, *all non-black things are non-ravens.* Upon observing a non-raven that is non-black, say a red shirt, a proponent of logical confirmation would have to concede that the observation confirms the generalization *all non-black things are non-ravens,* and by implication, it's contraposition, *all ravens are black.* Thus observation of a red shirt confirms *all ravens are black,* which is problematic because intuitively, observing shirts should not give any information about ravens.

Another example is Goodman's (1955) 'new riddle of induction,' which showed that syntax fails to distinguish between plausible and implausible predicates. It is natural to contend that observation of a green emerald supports the generalization that all emeralds are green. However, consider a different predicate, 'grue', meaning 'observed before the time *t,* and green or observed after *t* and blue.' The green emerald is also grue, assuming it is observed before *t*. Is it reasonable to accept the generalization that all emeralds are grue? Goodman contends that it is not. For one, the grue generalization implies that an emerald observed after *t* will be blue while the green generalization makes the opposite prediction. According to logical implication, the same observation equally confirms generalizations that make opposite predictions. Goodman (1955) showed that syntactically, the two-predicates are indistinguishable.

The failure of the logical approach to integrate confirmation and explanation has led to a bifurcation in the philosophy of science. On one hand, Studies in confirmation can be seen as heir to the logical empiricist approach because they have attempted to specify general rules or computations that avoid the pitfalls of hypothetico-deductivism. BCT (Fitelson, 2006), the most popular approach of this type is based on the rules of

probability rather than the rules of deductive logic. On the other hand, IBE (Lipton, 1991) has its roots in the philosophy of Peirce (1931) and Harman (1965) and tries to describe how specific knowledge supports the generation and evaluation of explanatory hypotheses.

1.2 Bayesian Confirmation Theory

BCT is based on a subjectivist interpretation of probability in which probabilities represent *epistemic credences,* or degrees of belief (De Finetti, 1980). The subjectivist interpretation is motivated by equating degree of belief to betting behavior and showing that a person who bets according to probabilities derived from the probability axioms cannot be liable to Dutch book, a conjunction of bets that is guaranteed to lose. The laws of probability can thus be seen as specifying a rational calculus for deriving probabilistic beliefs from other probabilistic beliefs in an analogous way that following the rules of logic guarantees coherence. Applying this idea to confirmation is natural because confirmation asks when an observation increases belief in a hypothesis. In the language of probability this is interpreted as asking the probability of the hypothesis conditioned on the observation. The formula for conditionalization can be derived from the probability axioms. According to Bayes' rule, the posterior probability, the probability of a hypothesis h_i conditioned on some observation, o, is

$$P(h_i \mid o) = \frac{P(o \mid h_i)P(h_i)}{\sum_{h \in H} P(o \mid h)P(h)} \quad (1)$$

where $P(h_i)$ is the *prior probability* of the hypothesis, the degree of belief that it is true before seeing any data. $P(o|h_i)$ is the conditional probability of the observation given the hypothesis, referred as the *likelihood* of the observation. The likelihood represents how probable it would be to observe o if h_i were true. The denominator normalizes across all

hypotheses, *H*, to generate a probability distribution. This enforces the constraint that the posterior probability of all hypotheses must sum to unity. Proponents of BCT claim that Bayes' rule dictates the normative standard for inference, and that it describes scientific practice. Some claim that is also describes ordinary inductive reasoning (Salmon, 2001).

BCT has been applied to the paradoxes of confirmation such as the raven paradox and the new riddle of induction discussed above. The Bayesian response to the raven paradox is to claim that observation of a non-black non-raven actually does confirm that all ravens are black, but that the degree of confirmation is very small. This can be justified on the basis of the likelihood of the observation given the hypothesis. Because the set of non-black things is so large observing a non-black thing that is not a raven is only very slightly more likely in a world where all ravens are black than in a world where some ravens are black and some are not (for a proof see Good, 1960). The Bayesian response to Goodman's new riddle of induction is to claim that plausible predicates have higher prior probabilities. This does not resolve the paradox because probability theory, like logic, does not furnish a way to distinguish green from grue (Strevens, 2004).

While BCT provides a formula for confirmation critics argue that it does not have inductive content. The constraints on inductions come from the primitives of the Bayesian calculation, but Bayesianism does not provide a general way to choose those primitives. All three inputs to the Bayesian calculation have been criticized in this way. Howson (2001) argues that the prior distribution can be fixed to yield any posterior distribution. Absent a general way to choose priors, Bayes' rule does not help to choose between hypotheses *pre hoc*. This is evidenced by the Bayesian resolution to the new riddle of induction. One could fix the prior probability to accord with one's belief about

which generalizations are more plausible but then the constraint on induction comes from explanatory considerations, not the *a priori* rules of probability. Miller argues that the likelihood function is also problematic. Calculating the likelihood distribution is easy when the observations are generated by a well-defined statistical model and the hypotheses are simple. Consider learning the distribution of ball colors in an urn by selecting balls at random one at a time. Ordinary inductions often involve complex hypotheses and data generation processes. Under those conditions, the likelihood distribution may be opaque. The third input to the Bayesian calculation, the hypothesis space, is also a target of criticism. Bayesianism provides a formula for choosing between hypotheses, but it is silent on where those hypotheses come from (Okasha, 2000).

1.3 Inference to the Best Explanation

IBE has it roots in the philosophy of Peirce (1931) who used the term 'abduction' to refer to a distinctive form of inference, central to everyday and scientific reasoning, by which hypotheses are generated and evaluated to explain observations. Josephson and Josephson (1994) describe abduction as following this pattern:

> *D* is a collection of data (facts, observations, givens)
> *H* explains *D* (would, if true, explain *D*)
> No other hypothesis can explain *D* as well as *H* does
> Therefore, *H* is probably true

Proponents of IBE argue that this pattern describes many important inferences, as when a jury reasons to a guilt or innocence on the basis of testimony, a detective reasons to a series of events from clues, a doctor reasons to the presence of a disease on the basis of symptoms, or a scientist reasons to the correctness of a theory on the basis of empirical observations. In all of these cases, the inference is diagnostic in the sense that it works backward from an observation to a hypothesis that *explains* that observation.

Pierce noted that the aspect of abduction that makes it a unique form of inference is that *H*, the explanatory hypothesis, is a novel entity that has to be added to the argument by the reasoner. In Peirce's words abduction is the only type of inference that "introduces something new (p. 171)." Thus unlike attempts to ground inferences in general syntactic or statistical rules, inference to explanation is inherently contextual, necessitating that the reasoner go beyond the syntactic information in the argument and draw on knowledge specific to the content of the argument. Harman (1965) popularized this idea as 'inference to the best explanation.' He argued against inductive logic on the grounds that good inductive practice prescribes taking account of the 'total evidence,' the observations or premises in the argument and any relevant background knowledge.

1.3.1 Explanation and Causality

The primary critique of IBE is that it is closer to a 'slogan' (Lipton, 2001) than a theory because it does not specify what constitutes an explanation. The most successful approaches to explanation in the philosophy of science implicate causal knowledge as the core of explanation (Salmon, 1987; Miller, 1987; Glymour, 2001). According to this view, an event is explained by delineating the causes of the event. The most influential contemporary articulation of IBE (Lipton, 2001) is also based on a causal approach to explanation.

The utility of a causal version of IBE can be seen by applying it to the paradoxes of confirmation discussed above as Miller (1987) does. Recall that according to the Raven Paradox, observing a red shirt appears to confirm that all ravens are black. According to IBE, any plausible causal explanation for the observation of the red shirt is completely irrelevant to the blackness of ravens. Likewise, IBE fails to confirm

Goodman's grue generalization. There are two possible explanations under which the grue generalization holds. Either, the emeralds that are not inspected before time *t* must change color, or some unknown factor must lead to observing only green emeralds before *t* and blue ones after. Neither explanation is plausible. Unlike the Bayesian resolution of the paradoxes, the IBE resolution does not rest on general statistical properties of the argument. Rather, it assumes that causal knowledge allows people to generate explanatory hypotheses that do or do not support the generalization.

2.1 Property Induction Paradigm

Having reviewed the major approaches to inductive inference in philosophy I now turn to psychology where there is also a split between theories of confirmation based on general rules for computing argument strength, and explanation-based approaches that focus on how specific knowledge mediates inference. Psychologists have investigated inductive reasoning within a paradigm introduced by Rips (1975) in which participants make judgments about arguments consisting of a premise or premises and a conclusion. A premise states that a particular category possesses a property and the conclusion is a supposition that based on the information in the premise or premises that some other category possesses the property. Here's an example from Sloman (1993):

Foxes secrete uric acid crystals
Beavers secrete uric acid crystals
Therefore Chihuahuas secrete uric acid crystals

'Foxes' and 'beavers' are the premise categories, 'Chihuahuas' is the conclusion category and 'secretes uric acid crystals' is the predicate, the property being projected. In Rips' (1975) original formulation premise and conclusion categories were basic level animal categories and all arguments involved projection of the property from one

category to another at the same level of abstraction. Subsequent studies have generalized the paradigm to arguments in which the conclusion is a member of the premise's super-ordinate category, and to other types of categories. Rips (1975) collected judgments of the percentage of the conclusion category's members that possessed the property given that all of the members of the premise category possessed it. Subsequent studies have used a variety of dependent measures including the strength of the argument, the probability of the conclusion or a forced choice between two arguments for strength. In many ways this type of task is similar to the

Another innovation introduced by Rips (1975) was the use of *blank predicates*, like 'secretes uric acid crystals.' A blank predicate is one that an average participant does not know to which categories is applies. The motivation for using blank properties was so responses would reflect the processes and representations underlying the inference as opposed to knowledge about a specific instance. Asking people to reason about properties that they know little about is a way to eliminate the possibility that people make their judgment based on background knowledge.

2.2 Similarity and Induction

From the introduction of the paradigm in the 1970's through the 1990's the major approaches to inductive reasoning were based on similarity measures between the categories of the arguments. The similarity-based approaches are not perfectly analogous to either confirmation or explanation-based theories in the philosophy of science, but they have more in common with confirmation theories. Like confirmation theories, similarity-based approaches attempt to account for inductive strength as a general computation that applies to all arguments irrespective of the property being projected. Further, like

confirmation theories, they specify mathematical models that yield quantitative calculations of inductive strength. Because of these features, the similarity models bear many similarities to the more directly analogous Bayesian confirmation models that followed them, with respect to their assumptions and predictions. Understanding these later approaches requires a cursory review of the similarity-based models (for more thorough reviews see Heit, 2000; Rips, 2001).

2.2.1 Similarity and Typicality

Rips' (1975) first study of argument strength was designed to test the importance of various similarity-based measures in judgments of argument strength, with similarity represented as distance in a low-dimensional space derived from multi-dimensional scaling (Shepard, 1980). In the studies, participants were asked to imagine an island with a small number of animal species. They were told that all of the members of one of the species had a contagious disease and were asked to estimate the percentage of members of the other species that had the disease. Rips found that judgments could be fit well with a model based on two distance measures, the distance from the premise to category to the conclusion category, representing the similarity of the premise and the conclusion, and the distance from the premise category to its immediately super-ordinate category, representing the premise typicality.

2.2.2 Similarity-Coverage Model

The similarity-coverage model proposed by Osherson et al. (1990) has been highly influential, and extended Rips' (1975) approach to arguments with more than one premise and to general arguments. Like Rips' (1975) model, the similarity-coverage model is based on two similarity-based judgments. As suggested by the name the model

is an average of 'similarity' and 'coverage,' weighted by a free parameter. 'Similarity' refers to the similarity between the premise and conclusion. In the case of more than one premise category, the maximal similarity is used. 'Coverage' is a generalization of Rips' (1975) notion of typicality and is a measure of the average similarity of the premise categories to all of the members of the immediately super-ordinate category that contains both the premise and conclusion categories. With a single premise this measure is high for typical categories. With multiple premises, it is high when the premise categories 'cover' the super-ordinate category. Similarity/coverage can account for the similarity and typicality effects reported by Rips (1975), but it also predicts many other well-established empirical phenomena, especially for arguments with multiple premises. Several important phenomena arise from the observation that inductive strength increases by adding premises that increase similarity or coverage. For instance, more diverse premises often increase inductive strength:

Hippos have sesamoid bones
Hamsters have sesamoid bones
Therefore all Mammals have sesamoid bones

Is judged stronger than:

Hippos have sesamoid bones
Rhinos have sesamoid bones
Therefore all Mammals have sesamoid bones

Likewise, adding premises usually increases inductive strength because additional premises increase similarity or coverage, as when:

Foxes have sesamoid bones
Pigs have sesamoid bones
Wolves have sesamoid bones
Therefore Gorillas have sesamoid bones

Is judged stronger than:

Pigs have sesamoid bones
Wolves have sesamoid bones
Therefore Gorillas have sesamoid bones

However, additional premises and diversity do not always increase strength. Non-monotonic effects of premise additions sometimes occur, as when:

Flies use serotonin
Therefore bees use serotonin

Is judged stronger than:

Flies use serotonin
Orangutans use serotonin
Therefore bees use serotonin

The similarity-coverage model predicts this phenomenon on the grounds that adding a premise that changes the super-ordinate category can decrease argument strength. In this case, 'flies' covers the category of insects better than 'flies' and 'orangutans' cover the category of animals.

The ability of similarity-coverage to account for so many phenomena is impressive, though the model does have drawbacks. For one, it assumes that people make use of pair-wise similarity judgments for the premise categories and all the members of the super-ordinate category when judging an argument. This implies, for instance, that making a judgment in which the super-ordinate category is 'animals' requires access to a large number of pair-wise similarity-judgments. The model also assumes that people represent a taxonomic hierarchy of categories and that the prescriptions of this hierarchy guide judgments. This is problematic in two ways. First, as discussed below, Sloman (1993, 1998) gives many examples where people's judgments do not respect class

inclusion relations. Second, the taxonomic hierarchy, while appropriate to inductions about animals and biological properties, may not be useful for describing all kinds of inductions. Thus the support for the model may be a function of the types of categories and predicates used in the materials. Blank predicates are meant to impart little knowledge, but the predicates used by Osherson et al. (1990), like 'has sesamoid bones' and 'uses serotonin' do bring to mind biological properties.

2.2.3 Feature-Based Induction

The feature-based model, proposed by Sloman (1993) resolves some of the weakness of the similarity-coverage model. The feature-based model represents the inductive process as a feature comparison between the premise and conclusion categories. Similarity is not used explicitly but to the extent that similarity is a function of feature overlap (cf. Tversky, 1978), it can account for many of the same phenomena as the similarity-coverage model such as premise diversity and premise monotonicity. The model is expressed as a neural network where input units represent values over a set of features and the output unit represents the presence or absence of the predicate. Two things determine the activation of the output unit, the feature overlap of the premise and conclusion and the number (in the binary case) or the 'magnitude' (more generally) of salient features possessed by the conclusion. Thus an argument is strong if there is a great deal of overlap of the features of the premise and conclusion and the conclusion has few additional features.

A virtue of the feature-based model is that it makes very few representational assumptions. The only input to the model is a vector of feature values. Unlike similarity-coverage, the model does not assume any category hierarchies, nor does it assume that

the inductive judgment has access to pair-wise similarity judgments for all of the categories in the domain. The inductive judgment is local in the sense that the only inputs to the model are the categories of the particular argument. This locality is beneficial for two reasons: First, fewer representational assumptions implies less computational demands and lends the model a degree of plausibility. Second, inductive judgments do not respect categorical hierarchies. People often commit inclusion fallacies (Osherson et al, 1990) as when they judge:

Robins have sesamoid bones
Therefore all Birds have sesamoid bones

To be stronger than:

Robins have sesamoid bones
Therefore Penguins have sesamoid bones

In the first argument the conclusion is a super-ordinate category that includes the conclusion of the second argument. As such, the second argument should never be judged stronger.

The similarity-coverage model can predict the inclusion fallacy on mathematical grounds by assuming that typical categories are judged more similar to their super-ordinate categories than they are to other atypical categories at the same level of abstraction, but committing the fallacy also necessitates that people ignore the logical implications of their representation, namely that a categorical hierarchy implies inclusion relations. The feature-based model avoids this tension by dint of its locality. Moreover the feature-based model predicts several novel inclusion fallacy phenomena which are not predicted by similarity-coverage such as the finding that

Animals have sesamoid bones

Mammals have sesamoid bones

Is rated stronger than:

Animals have sesamoid bones
Reptiles have sesamoid bones

and that neither is rated as perfectly strong. Sloman (1998) reports a host of phenomena like this. One weakness of the feature-based model is that it fails to predict non-monotonicity, because adding premises cannot decrease feature overlap or magnitude.

2.2.4 Relation between Similarity-Coverage and Feature-Based Model

One way to conceive of the difference between the two approaches to similarity-based induction is as a distinction between theories that ground induction in the structural relations between categories, like the similarity-coverage model, and those that ground it in the features of categories, like the feature-based model. I'll refer to these as the 'inside' and 'outside' focus respectively, because the distinction is broadly analogous to Lagnado and Sloman's (2004) distinction between 'inside' and 'outside' probability judgments. On an inside view the inductive process is local in the sense that it involves a comparison between the features of the categories in the arguments and does not include assumptions about the way those categories are related to each other or to other categories not mentioned in the argument. On an outside view, a key determinant of inductive strength is the way that the categories themselves are structured. For example, judgments about animals may be mediated by a taxonomic structure that embodies knowledge about the way different animals are related. The inside-outside distinction for induction is also roughly comparable to feature-based (Tversky, 1977) and spatial (Shepard, 1980) models of similarity. In the feature-based model, the similarity judgment is mediated by a local

feature comparison. On a spatial account, similarity is mediated by distance in a global, low-dimensional psychological space.

The similarity-coverage and feature-based models make many of the same predictions because they are both based on similarity judgments. For instance both approaches predict that inductive strength increases with increased typicality of the premises, increased diversity of the premises and increased similarity between the premises and conclusion. The differences between the models stem primarily from the different representational approaches, namely the inside versus outside focus. The feature-based model is better equipped to handle inclusion fallacies and other phenomena that violate the prescriptions of taxonomic hierarchies and it necessitates fewer computations. But it fails to predict non-monotonic effects, which are consistent with the taxonomy assumed by the similarity-coverage model.

2.3 Bayesian Models

Bayesian models, analogous to BCT in the philosophy of science, represent inductive reasoning as probabilistic inference, where the premises of the argument are treated as observations. Inductive strength is determined by the posterior probability of different 'hypotheses,' denoting the possible groupings of categories to which the predicate could apply. Shepard's (1987) 'universal law' for stimulus generalization was the first approach to represent an inductive problem in this way. In a stimulus generalization paradigm a stimulus is paired with a response and then the stimulus is varied systematically to see whether the response is elicited (Gutman & Kalish, 1956). For example a pigeon might be conditioned to peck at key based on light of a certain wavelength, and a generalization gradient is collected in a test phase by varying the

wavelength and measuring the response. Shepard conceived of stimulus generalization as a prototypical instance of inductive reasoning. In the learning phase the participant learns that a stimulus with a particular value or set of values is rewarded. In the test phase the participant has to decide based on the observations in learning whether a novel stimulus will generate a reward. Shepard's claim was that generalizations could be explained as the output of a Bayesian inference on a psychological similarity space. Generalization is mediated by an attempt to determine the 'consequential region' of that space, the stimulus values that will generate a reward. The major difference between Shepard's (1987) approach and Bayesian property induction models is in the nature of the hypothesis space. Shepard's model was applied to inductions about stimuli that varied along a small number of continuous dimensions and the participant generalized from some stimulus values to others. In the property induction paradigm, participants generalize from discrete categories to other categories. Thus the inductive problem is to determine the *consequential subset,* the set of categories to which the predicate applies, rather than a region of low-dimensional similarity space.

2.3.1 Qualitative Bayesian Model

Heit (1998) was the first to provide a Bayesian analysis to property induction. He gives the example of an argument in which the reasoner is asked to assess an argument in which the premise category is 'horses' and the conclusion category is 'cows.' For example consider the following argument:

Horses have sesamoid bones
Cows have sesamoid bones

The reasoner begins with four hypotheses as to the consequential subset of categories: i. that the predicate applies to neither cows nor horses; ii. That it applies to

both; iii. That it applies to horses but not cows and iv. That it applies to cows but not horses. A prior degree of belief is assigned to each hypothesis, on the basis of similarity. For instance, hypothesis ii is assigned a high prior because cows and horses are similar. Assigning the prior in this way is the equivalent of assuming that similar categories are more likely to share properties. The likelihood of each hypothesis is calculated simply by determining if it is consistent with the premise of argument. In this case, hypotheses i. and iv. are ruled out on the basis that horses have been observed to possess the property. Combining the likelihood and prior distributions according to Bayes' rule yields the posterior distribution over hypotheses.

Intuitively, degree of belief in the conclusion that 'cows have sesamoid bones' should be a sum of the hypotheses in which the conclusion is true, namely hypotheses ii and iv. Thus Argument strength is calculated by summing the posterior probabilities of those hypotheses. In the example, hypothesis ii is the sole remaining one that includes cows and it has a relatively high posterior because of its high prior. Thus the argument is rated as strong.

As should be obvious, most of the explanatory and predictive force of the model rests in the prior distribution, and the determination of this distribution is based on intuition and not strongly motivated. Thus the ability of the model to predict empirical phenomena again comes down to similarity, but in an even less-constrained way then the models discussed above. In that sense, the model is severely deficient from a descriptive standpoint. However, from a computational perspective, the idea of representing the problem in this way is a good one because it can provide a deeper explanation for the empirical phenomena. For example rather than saying that similarity is important to

inductive strength, Bayesian models begin to address the question of why it is important, namely because similar categories tend to belong to many of the same consequential subsets, or perhaps, that they are judged similar because they belong to many of the same consequential subsets.

2.3.2 Quantitative Bayesian Models

A number of the concerns associated with Heit's (1998) model have been addressed by a family of Bayesian models developed by Tenenbaum and colleagues (Tenenbaum & Griffiths, 2001; Sanjana & Tenenbaum, 2003; Kemp and Tenenbaum, 2003; Shafto et al., 2005) These models have made significant improvements to all three aspects of the Bayesian inference, the hypothesis space, the prior distribution, and the likelihood function. In Heit's (1998) model the hypothesis space is just all possible combinations of the categories of the argument and a prior distribution over the hypotheses is determined generically, based on similarity. Sanjana and Tenenbaum (2003) propose a motivated hypothesis space and prior distribution by creating a structured representation of the categories. They apply their model to a limited domain consisting of ten mammal species, and use a hierarchical clustering algorithm based on pair-wise similarity judgments to create a taxonomy of the animals. The hypothesis space is then supplied by clusters and unions of clusters in the taxonomy. This enforces a preference for hypotheses that are close in the taxonomy and hence for similarity without resorting to a generic distribution. The prior over the taxonomy is still a tricky issue, though reasonable priors have been created by relying on rational principles such as a preference for simpler hypotheses (Sanjana and Tenenbaum, 2003) or by defining a generative mutation process over the taxonomy (Kemp and Tenenbaum, 2003). With

regard to the likelihood calculation, Tenenbaum and Griffiths (2001) point out that the likelihood of a hypothesis given observations of positive instances is not always a simple indicator function as in Heit (1998), but is instead a function of how they data are assumed to be sampled. Given certain sampling assumptions, observation of a positive instance provides more evidence for a consistent specific hypothesis than for a more general hypothesis, a phenomenon that is referred to as the 'size principle.' Based on the size principle, the Bayesian model predicts novel empirical phenomena. For instance, the size principle predicts that given strong sampling, addition of premises within a range should weaken induction to a conclusion outside of that range.

Lions have sesamoid bones
Raccoons have sesamoid bones

Should be rated stronger than:

Housecats have sesamoid bones
Tigers have sesamoid bones
Lions have sesamoid bones
Raccoons have sesamoid bones

Evidence for such a phenomenon is mixed. Sanjana and Tenenbaum (2003) and Medin et al. (2003) found small non-monotonic effects of this type but Fernbach (2006) found that people usually considered arguments with more premise to be stronger even when the premises were similar to one another and task instructions implied strong sampling.

2.4 Evidence for Bayesian Models

There has been little work explicitly testing the Bayesian model. Heit's (1998) model is not testable because it does not make quantitative predictions. Sanjana and Tenenbaum (2003) compared fits of their Bayesian model to the predictions of the

similarity-coverage model using Osherson et al.'s data. Because the Bayesian model, like the similarity-coverage model, is based on a taxonomic hierarchy derived from similarity judgments it makes almost identical predictions to the similarity-coverage model so it is hard to distinguish the two models, though the Bayesian model does achieve slightly better fits. As discussed above, the evidence for novel empirical phenomena predicted by the Bayesian approach is mixed.

Work by McDonald, Samuels and Rispoli (1996) lends some indirect support to the Bayesian approach by showing that people are sensitive to the factors in the Bayesian analysis. They propose a hypothesis-assessment model, according to which property induction is a kind of hypothesis testing where the premise is treated as an observation and the conclusion is treated as one of several hypotheses advanced for the consequential subset. Drawing on the hypothesis assessment literature they identify three factors that should affect argument strength for general arguments. First, adding premises that instantiate the conclusion should strengthen an argument. Second, an argument with a smaller conclusion category should be more readily accepted. Third, an argument should be weakened to the extent that other conclusion categories come to mind as possible hypotheses. To test this idea they presented one group of subjects with an argument strength task with general arguments about animal categories. The other group saw just the premises of these arguments and was asked to generate hypotheses about possible conclusion categories. The analysis showed that the three predictors accounted for most of the variance in argument strength. All three of these predictors have analogues in Sanjana and Tenenbaum's (2003) Bayesian formulation. First, as the number of premises increases, inconsistent hypotheses are eliminated, leading to an increase in strength for

remaining hypotheses. Second, under the size principle, the likelihood function favors smaller conclusions over larger ones given evidence consistent with both. Lastly, more alternative hypotheses imply a larger number of hypotheses with posterior probability lowering the relative strength of each. The similarity-coverage and feature-based models can also account for the first two predictors on the basis of similarity, coverage and feature overlap, but since they calculate strength directly, they do not have anything to say about alternative hypotheses. This implies that some aspect of hypothesis assessment, as embodied in a qualitative framework or in a Bayesian one, captures a unique and important part of the induction process.

2.5 Summary of Confirmation theories

As in the philosophy of science, one of the major approaches to inductive reasoning in psychology is to specify general computations that specify how much confirmation a set of observation lends to a conclusion irrespective of the content of the argument. Similarity-based models specify a direct computation of inductive strength based on pair-wise similarity judgments or feature overlap. The similarity-coverage model and the feature-based model are two proposals for similarity models that differ with respect to their representational assumptions. The similarity-coverage model takes an outside view of induction and represents all of the categories in the domain in a taxonomic hierarchy. The feature-based model takes an inside view, only representing the categories mentioned in the argument. Bayesian models, analogous to BCT in the philosophy of science represent induction as a probabilistic inference intended to identify consequential subset of categories to which the predicate applies. A third confirmation-

based approach, hypothesis assessment, is couched in different language, but is consistent with the Bayesian formulation.

The confirmation-based theories make many of the same predictions. Argument strength increases with premise typicality and diversity, and with increased similarity between the premise and conclusion categories. These predictions are supported by argument strength experiments in which arguments consist of animal categories and blank biological properties. Empirical phenomena that are specific to particular theories also have some support.

2.6 Effects of Explanation

Confirmation theories can account for many phenomena associated with inductive reasoning about blanks predicates and animal categories. Do these theories generalize to other kinds of inductions? According to theories of IBE in the philosophy of science, confirmation theories fail because they do not take explanation into account. IBE posits that inductive inferences are mediated by knowledge beyond what is given in the syntax of the argument, and people accept hypotheses that explain their observations. Likewise, in the property induction literature there is a great deal of evidence that inductive reasoning often depends on specific knowledge about the property being projected and about the relation between the property and the categories in the argument.

In the following sections I will review three kinds of evidence for this claim. First, people reason differently about arguments depending on the property being projected. Second, experts and novices reason differently about the same arguments. Third, manipulating explanation directly influences argument strength in ways that are not accounted for by general theories of confirmation.

2.6.1 Property effects

In a study similar to Rips' (1975), Nisbett et al. (1983) explored inductive judgments about different types of properties. Participants were instructed to imagine that they were explorers on an island and to estimate the proportion of a particular category's members that possessed a property given one or more positive instances. For example they might be told that they had observed a single tribe member with dark skin and asked to estimate the proportion of tribe members with dark skin. The key manipulation was whether the property was something the experimenters assumed was homogeneous, like skin color, or something variable, like weight. They found a very large difference in judgments. Given a single instance of an obese tribe member participants gave inductive judgments below 40%, but given an instance of a dark-skinned tribe member, responses were about 95%. Since the inductive judgments both concerned projecting the property from one tribe member to others, confirmation theories predict that strength should be the same. Nisbett et al. (1983) explain this finding in terms of statistical knowledge. They assume that people know something about way that different properties are distributed in populations and use this knowledge to mediate the induction.

In a set of studies closer to the standard paradigm, Heit and Rubinstein (1994) also found property differences by varying whether the property was an anatomical or behavioral property and asking for inductive judgments about single-premise, specific arguments in which the premise category was anatomically similar or behaviorally similar to the conclusion. In all cases, the predicate was blank in that the participant did not know whether the conclusion possessed it, but it was sufficiently interpretable to be clearly related to anatomy or behavior, for example, 'its liver has two chambers that act

as one' as an anatomical property, and 'it usually travels in a back and forth or zig-zag trajectory' as a behavioral one. The key finding was that people were more willing to project anatomical properties between categories that are similar anatomically, and behavioral properties between categories that are similar behavioralllly. For instance, people projected 'feeds at night' more strongly from wolf to shark, two predators. Conversely, they we more willing to project 'its blood contains between 2% and 3% potassium' from trout to shark, two fish. Ross and Murphy (1999) report similar results in the domain of food.

Heit and Rubinstein (1994) interpret their results as indicating that people have two measures of similarity that are specially recruited depending on the domain suggested by the predicate. It is interesting to note however that participants appear to be sensitive to even more specific aspects of the arguments than whether they reflect anatomical or behavioral properties. Though the general trend was in the direction suggested by different measures of similarity, there were significant item-by-item differences. Heit and Rubinstein (1994) were aware of these issues and suggest, after Murphy and Medin (1985), that similarity itself may be a function of a background theory that identifies properties that are causally relevant to the predicate. This idea is not developed in any detail but Heit and Rubinstein (1994) appear to have a three-stage process in mind. First, relevant properties are identified through a causal analysis of the premise category, then similarity is calculated with respect to those properties deemed relevant, and finally, inductive strength is calculated as a function of similarity. The notion that induction is mediate by a process that identifies the causes of the predicate is

consistent with IBE. Referring to this process as a kind of similarity computation does not add anything to the analysis.

Several studies lend further support to such a process, but do so in a way that highlights the importance of specific properties or causal relations rather than some notion of similarity. Smith, Shafir and Osherson (1993) report a number of examples where induction of nonblank properties violates similarity. For instance,

Poodles can bite through wire
Therefore German Shepherds can bite through wire

Was judged stronger than:

Dobermans can bite through wire
Therefore German Shepherds can bite through wire

Dobermans and German Shepherds were judged more similar than Poodles and German Shepherds but Poodles supported a stronger induction, presumably because participants explained the predicate in terms of a property possessed by the premise, like strength. If Poodles are strong enough to bite through wire, then surely so are German Shepherds because they are much stronger than Poodles.

2.6.2 Expertise effects

A second set of studies exploring the role of knowledge in induction concerns the differences between experts and novices (Reviewed in Coley et al., 2005). Experts are knowledgeable about their domains of expertise and draw on this expertise to make inductions. As in the property effects described above, this implies that inductive strategies are a function of how much specific knowledge the reasoner has about the predicate and its relation to the categories of the argument.

Lopez et al. (1997) compared inferences about local mammal species by the Itzaj-Maya of Guatemala to inferences about the same species by U.S. undergraduates and found significant differences in ratings of argument strength. For instance, when judging whether all local mammals had a disease given that two local species had it, U.S. undergraduates consistently judged more diverse premises to support stronger inductions, as predicted by confirmation theories. Itzaj-Maya showed no such pattern, relying instead on local ecological knowledge as the basis for inductions. Justifications were not analyzed in detail, but some trends were reported. For instance, Itzaj-Maya participants sometimes rejected arguments for which they could not find a good causal or ecological explanation why the two premise categories should both have the disease. Profitt, Coley & Medin (2000) obtained similar results in a study of tree-experts, implying that expertise and not cross-cultural differences was responsible for differences in diversity-based reasoning.

Stepanova and Coley (2003) showed the effect of expertise within subjects by varying the domain of inductions. College students were asked to judge inductive arguments about alcoholic beverages, a domain in which they were assumed to be experts, and about animals, a domain in which they were novices. Knowledge about the predicate was also varied by using a blank predicate, possession of a chemical, and a non-blank one, causing sickness. The key finding was that property interacted with domain. Evidently, when reasoning about alcohol and sickness, participants used expert knowledge but not when reasoning about chemicals. In the domain of animals, participants' judgments were the same for both types of predicates.

2.6.3 Direct Manipulation of Explanation

Sloman (1994) manipulated explanations more directly. He compared arguments with non-blank predicates in which the explanation for the predicate was the same for the premise and conclusion categories and in which it was different, and found that like explanations increased inductive strength. For instance:

Many ex-cons are hired as body guards
Therefore many war veterans are hired as bodyguards

Was judged stronger than:

Many ex-cons are unemployed
Therefore many war veterans are unemployed

It is reasonable to assume that ex-cons are hired as bodyguards because they are tough and accustomed to dangerous situations, an explanation that also applies to war veterans. In the second argument the explanation for ex-cons' unemployment does not apply as naturally to war veterans. Again, the similarity of the premises and conclusions for the two arguments are the same and explaining the effect requires appealing to a specific property of the premise. Sloman (1994) also obtained judgments of prior probability of the conclusion and found that premises with like explanations increased strength over base-line levels implying that participants treated the premise as evidence to increase their belief in the conclusion. More surprisingly, he also found that premises with incompatible explanations decreased belief in the conclusion below baseline levels. In a follow-up paper Sloman (1997) showed that discounting occurred even for premises that were irrelevant to the conclusion such as:

Most cars have hoods
Therefore most jackets have hoods

Participants actually judged the conditional property of the conclusion given the premise to be lower than the prior probability of the conclusion. Though these effects are not well understood they point to an important influence of explanations on judgments.

2.6.4 Artificial Categories

The importance of explanations has also been shown by manipulating the causal structure of artificial categories. Lassaline (1996) showed that adding knowledge about a causal relation connecting a property possessed by the premise category to the predicate increased induction to a conclusion that also possessed that property. For example subjects rated:

> *Animal A has a weak immune system, skin that has no pigment, and dry flaky skin*
> *Animal B has a weak immune system and an acute sense of smell*
> *For animal B, a weak immune system causes an acute sense of smell*
> *Therefore, Animal A also has an acute sense of smell*
>
> *To be stronger than*
>
> *Animal A has a weak immune system, skin that has no pigment, and dry flaky skin*
> *Animal B has a weak immune system and an acute sense of smell*
> *Therefore, Animal A also has an acute sense of smell*

In a similar setup, Rehder (2006) compared inductions based on similarity or causal explanation more directly. He taught participants about novel categories, such as 'Kehoe Ants,' that possessed four made-up properties such as 'stinging bite,' and had them judge the strength of projection of a novel target property from one member of the category to another. He manipulated similarity by varying the number of features in common between the premise and conclusion individuals, and he varied the presence of an explanation by telling participants that one of the features caused the target property. In the explanation-present condition, the premise and conclusion shared the feature

responsible for the target property, while in the explanation-absent condition the conclusion individual did not have this feature. He found that similarity had no effect on inductions when an explanation was given. When the feature responsible for the target property was present in the conclusion, inductive strength was very high. In contrast, when the feature was absent, strength was very low. When no explanation was given, judgments were indeed a function of feature overlap. This implies that participants preferred to make their inductive judgment on the basis of causal knowledge as opposed to similarity (also see Rehder & Hastie, 2001).

2.7 Explanation-Based Theories

A variety of studies have shown effects of varying knowledge about the property and categories of arguments that cannot be accounted for by the confirmation-based theories. This work shows that in order to determine the confirmation that the premises lend to the conclusion, people attempt to explain the premises. Explanation involves going beyond the syntax of the argument and searching for the causes of the predicate in one's pre-existing knowledge. Inductive strength is mediated by the extent to which good explanations also apply to the conclusion. In other words, people engage in IBE when assessing argument strength. The reliance on IBE leads to many empirical phenomena. Varying the predicate can have drastic impacts on argument strength even when the structure of the arguments remains constant. Experts reason differently about their domain of expertise than do novices. Manipulating explanations directly using real-world or artificial materials strongly influences induction. In the following sections I will review three theories have been proposed to explain these effects.

2.7.1 Relevance Theory

Medin et al. (2003) propose a relevance theory of induction, inspired by Sperber and Wilson's (1986) pragmatic theory. The key idea is that when assessing an inductive argument, participants assume that the premises are chosen so as to be relevant to assessing the strength of the argument. The principle of relevance has two aspects. First, salient properties of the premises are deemed most important, and salient properties that are shared by multiple premise categories are reinforced. In other words, the reasoner assumes that experimenter is selecting the premise categories so as to give hints about what properties might be relevant to the induction. Second, any causal relations between the premise and conclusion categories are deemed relevant to the induction. For instance, in an induction from bananas to monkey, though monkeys and bananas are quite different, the fact that monkeys eat bananas is relevant to inductive strength. These assumptions lead to several predictions of novel empirical phenomena. Here I focus on two of these predictions that are of particular interest.

If premises are selected so as to highlight certain shared features than an argument will be weakened when this feature is not possessed by the conclusion category this leads to non-monotoncity via property reinforcement. Medin et al. (2003) predict that

Brown Bears have sesamoid bones
Therefore Buffalo have sesamoid bones

Will be rated stronger than:

Polar Bears have sesamoid bones
Grizzly Bears have sesamoid bones
Brown Bears have sesamoid bones
Therefore Buffalo have sesamoid bones

According to the principle of relevance, the reasoner assumes that the selection of three bear categories in the second argument is informative that a salient property of bears

explains the predicate. Thus an induction to a category like ‘buffalo’ which does not share that property is weak. As discussed above, a Bayesian model with a strong sampling assumption, also predicts this kind of effect. This highlights an interesting correspondence. Assumptions about the intentions of the experimenter in selection of the premises is analogous to sampling assumptions in the Bayesian framework because both ideas concern how evidence is selected. Moreover, the commitment to an assumption of informativeness in the relevance theory is analogous to an assumption of strong sampling advanced in several Bayesian models (e.g. Tenenbaum & Griffiths, 2001), which leads both theories to predict a non-monotonic effect of adding similar premises, when the conclusion category is dissimilar.

People do not always demonstrate this effect. Medin et al. tested for non-monotonicity via property reinforcement by varying the number of similar premises in the argument (2 vs. 4 in one case and 3 vs. 5 in another), and results were generally in line with predictions. On average, participants preferred the arguments with fewer premises. However, the effect did not hold up across all test items and in fact significant effects were found in the opposite direction for some items. As discussed above, Fernbach (2006) also found that many participants did not show such a non-monotonic effect. He varied sampling assumptions by changing the instructions across subjects. Though some participants did show the non-monotonic effect, most preferred arguments with more premises. These finding cast doubt on the idea that participants always bring the same pragmatic assumptions to bear on induction problems, which calls into question the project of identifying general principles of relevance in the way intended by Medin et

al. (2003). Clearly though, pragmatics do play a role, as evidenced by the non-monotonic effect shown by some participants and for some test items.

A second prediction of the relevance theory is that of causal asymmetry. For example:

Carrots have property 'X'
Therefore Rabbits have property 'X'

Will be rated stronger than:

Rabbits have property 'X'
Therefore Carrots have property 'X'

In the first argument there is a causal path that flows from the premise category to the conclusion category that provides a good explanation for the transmission of a property. In the second argument the relation is reverse the direction of projection and as such the argument should be judged weaker. Medin et al. (2003) tested arguments of this type and found that judgments were stronger for inductions in the direction of transmission. One question that arises from this finding is about the normative status of this asymmetry. Because causal relations support diagnostic inference in addition to predictive inference (Pearl, 2000), the second argument might be judged quite strong on the grounds that if rabbits have the property, and eating carrots is a plausible cause, then it is likely that carrots have the property. Indeed the normative strength of the diagnostic inference depends on assumptions about other possible causes. Consider a case where there are no alternative causes for the property. In that case, the diagnostic inference should be perfectly strong. Fernbach, Darlow and Sloman (in preparation) tested this idea by asking people to judge inductions about property transmissions. They varied the predicate across conditions to create situations where there were or were not alternatives explanations and

asked for judgments of the strength of the causal and diagnostic inferences. The design of the experiment with an example item is shown below:

	Diagnostic Inference	Causal Inference
No Alternatives	*Baby is drug addicted* *Mother is drug addicted*	*Mother is drug addicted'* *Baby is drug addicted*
Yes Alternatives	*Baby has dark skin* *Mother has dark skin*	*Mother has dark skin* *Baby has dark skin*

They found that when there were no alternatives, like in the case where a mother's drug addiction is the only plausible cause of her baby's addiction, the diagnostic inferences were rated as stronger than the causal inferences. When there were alternatives the opposite was true. This is counter to the prediction of relevance theory, that inferences in the causal direction are always stronger. Evidently, a simple rule like causal asymmetry is insufficient to capture inductive reasoning about property transmissions.

2.7.2 Structural Alignment

A model based on structural alignment was proposed by Lassaline (1996) and further developed by Wu and Gentner (1998). Inspired by models of relational mapping (e.g. Gentner, 1983) this approach extends structural alignment theories of similarity (e.g. Markman & Gentner, 1993) to inductive argument strength. The key idea is that both similarity and inductive strength are a function of a structural comparison between the premise and conclusion categories but that the two judgments rely on different comparisons. Specifically, the categories are conceived as structured representations of attributes and relations. This is roughly comparable to features and causal relations in a causal model representation, though relations are construed more broadly to encompass both causal and other kinds of relations. In assessing an inductive argument there are two

types of critical relation, 'shared relations' and 'binding relations.' Shared relations are those that are shared by the premise and conclusion categories. Binding relations are relations of the premise that are assumed to be causally responsible for the predicate. The model proposes that similarity is a function of shared relations, while inductive strength is a function of binding relations. Thus categories are similar to the extent that they have shared relations, but induction from one to the other is supported only by binding relations. Lassaline (1996) gives the example of a cat with poor vision because it is old and scratches incessantly because of fleas. In assessing the projection of the predicate 'scratches incessantly,' the relation between fleas and scratching is clearly the binding relation. An animal that shares many features with this cat might be similar but it is not a strong candidate for projection of the predicate unless it too has fleas. Lassaline (1996) and Wu and Gentner (1998) report several studies that confirm this prediction.

Structural alignment is consistent with IBE because a binding relation in a relational structure is roughly analogous to an explanation in a causal structure. But the theory assumes conditions where the binding relation is known *a priori. M*any inductive arguments present a predicate that represents new information about the premise category. The process by which explanations for the predicate are generated and evaluated is not addressed.

Another weakness of the approach is that it explains induction in terms of structural alignment however alignment is not a part of the inductive judgment in any substantive sense. Inductive judgments are a function of a single shared binding relation not a structural comparison. The reason this is a problem is because the explanatory power of conceiving of similarity as a relational mapping is not inherited by the model

and the model is dangerously close to merely stating a well-known empirical fact, namely that causal relations influence inductive strength, without providing anything in the way of a process or computational model.

2.7.3 Causal Reasoning

Rehder (2007) has applied a similar idea to a causal model representation of categories, which he refers to as 'generalization as causal reasoning (GCR).' Like relevance theory and structural alignment his approach is to identify qualitative predictions that result from conceiving of the problem in this way and then test for them empirically. For instance, people should rate inductive arguments strong on the basis of diagnostic evidence for the predicate. For example if a premise category has some disease that causes symptoms, then a reasoner should judge that a conclusion category is likely to have that disease if it also has the symptoms. Empirical support for this idea comes from Hadjichristides et al. (2004), who showed that central features were more projectable than non-central features between similar animals. In other words, inductions were strong when the premise and conclusion category shared many surface properties and the predicate being projected was assumed to be a cause of those properties in the premise category. Another prediction of GCR is that induction should be sensitive to prospective reasoning in which one reasons from the presence of the causes of the predicate to the presence of the predicate itself. This is comparable to the effects reported by Lassaline (1996) in which a shared binding relation increases inductive strength.

GCR is the approach that is most directly analogous to IBE because it proposes that induction is the outcome of a causal reasoning process that evaluates the predicate in light of the causal structure of the premise categories. Like the structural alignment, GCR

suggests some qualitative predictions that are well supported, but it doesn't propose a computational model.

2.8 Summary of Explanation-Based Theories

I have reviewed three theories that attempt to account for how people use knowledge about the categories and predicates of arguments to determine inductive strength. Relevance theory draws on an analogy to pragmatics and assumes that people treat the premises of the argument as informative about relevant properties of the categories. Theories based on structural alignment and causal reasoning, more directly comparable to theories of IBE described earlier, assume that inductive strength is mediated by a process that explains the predicate in terms of the causal or relational structure of the categories in the argument.

The general weakness of these theories is the same as the weakness of theories of IBE in philosophy. All of them make qualitative predictions but fail to provide a well-specified model that could yield quantitative predictions for a variety of arguments. Beyond that, Relevance theory's predictions have mixed empirical support. The theory's commitment to general pragmatic principles misses some of the intricacies of human reasoning like sensitivity to assumptions about how data are sampled, and the effect of causal structure on judgments of causal and diagnostic inferences. The predictions of structural alignment and GCR are well supported but they are unsurprising, given the well-known explanation effects discussed above. Both theories deal with cases where the explanation is already known and show that the presence of an explanation mediates inductive strength, a fact that is well-established in the literature. Neither approach addresses how explanations are generated and evaluated in a computational framework.

Conclusions

Given the review of explanation effects in induction we are now in position to assess whether IBE or BCT is a better account of human inference. It would be tempting to conclude that IBE is superior on the grounds that confirmation theories cannot account for people's tendency to go beyond general properties of arguments and make inductions on the basis of knowledge specific to the categories and predicates in the argument. IBE is intended to do just that because it grounds inference in causal explanations.

I would like to argue that this conclusion is not warranted. The effects of explanation clearly show that Bayesian models of the type introduced above are inadequate to account for the range of inductions that people make. However, BCT can be construed as a kind of inference to the best explanation because it specifies a computation for choosing between hypotheses on the basis of new evidence. In principle, Bayes' rule could describe how people choose between explanatory hypotheses. On this interpretation, BCT would not be an incorrect account of inference. It would just be incomplete. A Bayesian model would have to be supplemented with a method for deriving the primitives for the calculation from some representation of knowledge. At the same time, IBE does not fare much better at accounting for the effects of explanation. The idea that explanation is central to inference is descriptively correct, but without a fully-specified model, IBE is limited. The explanation-based theories in psychology share this weakness. This suggests that deciding between IBE and BCT requires a closer examination of the relation between the two approaches.

In the philosophy of science there is a vigorous debate on this issue. Lipton (2001) describes the difference by distinguishing between the 'loveliness' and the

'likeliness' of an explanation. The loveliness of an explanation is the extent to which it provides understanding. The likeliness of an explanation is its probability of being true irrespective of whether it provides understanding. For example, the absence of food is a likely explanation of the famine in India, while the drought is a lovely explanation. According to Lipton, inference to the best explanation means inference to the loveliest explanation. Loveliness and likeliness are often correlated, but it is loveliness and not likeliness that determines which hypotheses are accepted. Proponents of BCT claim that confirmation is calculated based on epistemic credences, degrees of belief. Explanatory considerations may play a role in fixing these credences, but inference is guided by posterior degree of belief in the truth of the hypotheses. Thus confirmation is based on likeliness and not on loveliness. Hypotheses that are confirmed can subsequently provide explanations for the observations that confirmed them (Salmon, 2001).

For the purposes of this paper a key question is whether this difference can be cast as a testable psychological question. One way of doing this is suggested by Lipton (2001). He proposes that it may be easier for people to access judgments of loveliness than it is to access the numerical probabilities necessary for a Bayesian calculation, but that since loveliness and likeliness are often correlated, judgments of loveliness can guide an updating process that approximates Bayesian inference. This resolution is analogous to saying that BCT and IBE stand in relation to one another as a computational model stands to an algorithmic model (Marr, 1980). The flip-side of this interpretation is that if inferences are guided by loveliness then there should be conditions under which people's judgments systematically diverge from the prescriptions of BCT. People should be willing to endorse hypotheses that are explanatory beyond what is prescribed by Bayes'

rule. If, as suggested by Salmon (2001), the causal relation is in the other direction and a Bayesian calculation of posterior probability determines judgments of loveliness, then there should be greater correspondence to normative standards.

Given the centrality of explanation to human inference, the key tenet of IBE is certainly correct. Fleshing out IBE by identifying the processes and representations that underlie abductive inferences will surely be a worthwhile endeavor for psychologists. A theory of inference will require specifying precisely how explanations are used to determine confirmation. It may turn out that people follow Bayes' rule, or it may be that people use some other confimation function, like a heuristic based on loveliness. In either case, confirmation theory will have to be informed by a theory of explanation. The utility of BCT as a descriptive account of human cognition is less clear, because it is contingent on showing that people really do respect normative standards when choosing explanatory hypotheses.

References

Coley, J. D., Shafto, P., Stepanova, O. & Barraff, E. (2005). Knowledge and category-based induction. In W. Ahn, R. L. Goldstone, B. C. Love, A. B. Markman & P. Wolff (Eds.), *Categorization inside and outside the laboratory: Essays in honor of Douglas L. Medin.* Washington, DC: American Psychological Association.

De Finetti, B., (1980). Foresight. Its logical laws, its subjective sources. in H. E. Kyburg, Jr. and H. E. Smokler (eds.) *Studies in Subjective Probability*, Robert E. Krieger Publishing Company.

Fernbach, P. M. (2006). Sampling assumptions and the size principle in property induction. *Proceedings of the 28th annual conference of the cognitive science society.*

Fitelson, B. (2006) The Paradox of Confirmation. *Philosophy Compass*, 1 (1), 95–113

Gentner, D. (1983). Structure mapping: A theoretical framework for analogy. *Cognitive Science*, 7, 155-170.

Glymour, C. (2001). *The mind's arrows: Bayes nets and graphical causal models in psychology*. Cambridge, MA: Bradford Books.

Good, I. J. (1960) The Paradox of Confirmation, *The British Journal for the Philosophy of Science*, 11, 42, 145-149

Goodman, N. (1955). Fact, fiction, and forecast. Cambridge, MA: Harvard University Press.

Gutman, N. & Kalish, H. I. (1956). Discriminability and stimulus generalization. *Journal of Experimental Psychology, 51 (1),* 79-88.

Hadjichristidis, C., Sloman, S. A., Stevenson, R., & Over, D. (2004). Feature centrality and property induction. *Cognitive Science, 28(1)*, 45-74.

Harman, G. H. (1965). The inference to the best explanation. *The Philosophical Review 74(1)*, 88-95.

Hempel, C. G. (1945). Studies in Logic and Confirmation. *Mind*, 54, 1-26.

Hempel, C. G. (1965). Aspects of Scientific Explanation. New York: Free Press.

Heit, E., & Rubinstein, J. (1994). Similarity and property effects in inductive reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20 (2), 411-422.

Heit, E. (1998). A Bayesian analysis of some forms of inductive reasoning. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition*: Oxford University Press.

Heit, E. (2000). Properties of inductive reasoning. *Psychonomic Bulletin & Review*, 7 (4), 569-592.

Howson, C. (2001), *Hume's Problem: Induction and the Justification of Belief*, Oxford: Oxford University Press.

Hume, D. (1748). *An enquiry concerning human understanding*. Oxford: Clarendon.

Jospehson J. R. & Jospephson, S. G. (1994). *Abductive Inference: computation philosophy, technology*. Cambridge University Press: New York, NY.

Kemp, C., & Tenenbaum, J. (2003). Theory-based induction. *Proceedings of the Twenty-Fifth Annual Conference of the Cognitive Science Society.*

Lagnado, D., & Sloman, S. A., (2004). Inside and outside probability judgment. In D. J. Koehler & N. Harvey (Eds.), *Blackwell hand-book of judgment and decision making*. Oxford, UK: Blackwell Publishing.

Lassaline, M. E. (1996). Structural alignment in induction and similarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(3), 754-770.

Lipton, P. (1991) *Inference to the best explanation*. London: Routledge.

Lipton, P. (2001) Is explanation a guide to inference? A reply to Wesley C. Salmon. In G. Hon & S. S. Rakover (Eds.), *Explanation: Theoretical approaches and applications.* Dordrecht: Kluwer.

Lopez, A., Atran, S., Coley, J. D., Medin, D. L., & Smith, E. E. (1997). The tree of life: Universal and cultural features of folkbiological taxonomies and inductions. *Cognitive Psychology*, 32(3), 251-295.

Markman, A. & Gentner, D. (1993). Splitting the differences: A structural alignment view of similarity. *Journal of Memory and Language, 32,* 517-535.

Marr, D. (1982). *Vision*. San Francisco: W.H. Freeman.

McDonald, J., Samuels, M. & Rispoli, J. (1996). A hypothesis-assesment model of categorical argument strength. *Cognition, 59,* 199-217.

Medin, D. L., Coley, J. D., Storms, G., & Hayes, B. K. (2003). A relevance theory of induction. *Psychonomic Bulletin & Review*, 10(3), 517-532.

Miller, R. W. (1987). *Fact and method*. Princeton, NJ: Princeton University Press.

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

Nisbett, R. E., Krantz, D. H., Jepson, C. & Kunda, Z. (1983). The use of statistical heuristics in everyday inductive reasoning. *Psychological Review* 90 (4), 339-363.

Okasha, S. (2000) Van Fraassen's critique of inference to the best explanation. *Studies in History and Philosophy of Science, 31 (4),* 691-710.

Osherson, D., Smith, E. E., Myers, T. S., & Stob, M. (1994). Extrapolating human probability judgment. *Theory and Decision*, 36, 103-126.

Osherson, D. M., Smith, E. E., Wilkie, O., Lopez, A., & Shafir, E. (1990). Category-based induction. *Psychological Review*, 97, 185-200.

Peirce, C. S. (1931). *Collected papers*. C. Hartshorn and P. Weiss (Eds.), Cambridge: Harvard University Press.

Pearl, J. (2000) *Causality*. Cambridge: Cambridge University Press.

Popper, K. (1959). *The Logic of Scientific Discovery*. London: Hutchinson.

Proffitt, J. B., Coley, J. D., & Medin, D. L. (2000). Expertise and category-based induction. *Journal of Experimental Psychology*: *Learning, Memory, and Cognition,* 26(4), 811-828.

Rehder, B. & Hastie, R. (2001). Causal knowledge and categories: The effects of causal beliefs on categorization, induction and similarity. *Journal of Experimental Psychology: General, 130 (3)*, 323-360.

Rehder, B. (2003). A causal-model theory of conceptual representation and categorization. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 29(6), 1141-1159.

Rehder, B. (2006). When similarity and causality compete in category-based property induction. *Memory & Cognition 34*, 3-16.

Rehder, B. (2007). Property generalization as causal reasoning. In A. Feeney & E. Heit (Eds.), *Inductive reasoning: Cognitive, mathematical and neuroscientific approaches.* Cambridge: Cambridge University Press.

Rips, L. J. (1975). Inductive judgments about natural categories. *Journal of Verbal Learning and Verbal Behavior*, 14 (6), 665-681.

Rips, L. J. (2001). Necessity and natural categories. *Psychological Bulletin*, 127 (6), 827-852.

Ross B. H., & Murphy, G. L. (1999). Food for thought: Cross-classification and category organization in a complex real-world domain. *Cognitive Psychology,* 38, 495-553.

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

Salmon, W. C. (2001a) Explanation and confirmation: A Bayesian critique of inference to the best explanation. In G. Hon & S. S. Rakover (Eds.), *Explanation: Theoretical approaches and applications.* Dordrecht: Kluwer.

Sanjana, N. E., & Tenenbaum, J. B. (2003). Bayesian models of inductive generalization. *Advances in Neural Information Processes 15.*

Shafto, P., Kemp, C., Baraff, E., Coley, J. D., & Tenenbaum, J. (2005). Context sensitive induction. *Proceedings of the Twenty-Seventh Annual Conference of the Cognitive Science Society.*

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting and clustering. *Science*, 210 (4468), 390-398.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, 237(4820), 1317-1323.

Sloman, S. A. (1993). Feature-based induction. *Cognitive Psychology*, 25, 231-280.

Sloman, S. A. (1994). When explanations compete: The role of explanatory coherence on judgments of likelihood. *Cognition*, 52, 1-21.

Sloman, S. A. (1997). Explanatory coherence and the induction of properties. *Thinking and Reasoning*, 3(2), 81-110.

Sloman, S. A.. (1998). Categorical inference is not a tree: The myth of inheritance hierarchies. *Cognitive Psychology*, 35, 1-33.

Sloman, S. A., & Lagnado, D. A. (2004). The problem of induction. In R. Morrison & K. Holyoak (Eds.), *Cambridge handbook of thinking and reasoning*. New York: Cambridge University Press.

Sperber D. & Wilson, D. (1986). *Relevance: Communication and cognition*. Oxford: Blackwell.

Stepanova O. & Coley, J. D. (2003). Animals and alcohol: The role of experience in inductive reasoning among college students. *Paper presented at annual meeting of the Psychonomic Society.*

Strevens, M. (2004), Bayesian Confirmation Theory: Inductive Logic, or Mere Inductive Framework? *Synthese*, 141, 3, 365-379.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity and Bayesian inference. *Behavioral and Brain Sciences*, 24, 629-640.

Tversky, A. (1977). Features of similarity. *Psychological Review*, 84 (4), 327-352.

Wu, M. L., & Gentner, D. (1998). Structure in category-based induction. *Proceedings presented at the 20th Annual Conference of the Cognitive Science Society.*