A. Gopnik (2000). Explanation as orgasm and the drive for causal understanding: The evolution, function and phenomenology of the theory-formation system. In F. Keil & R. Wilson (Eds.) *Cognition and explanation.* Cambridge, Mass: MIT Press.299-323.

Explanation as orgasm and the drive for causal knowledge: The function, evolution and phenomenology of the theory-formation system.

Alison Gopnik

Dept. of Psychology

University of California at Berkeley

Berkeley, CA.

This research was supported by NSF grant DBS9213959. I am very grateful for helpful discussions with Henry Wellman, John Flavell, Clark Glymour, John Campbell, Dacher Keltner, Daniel Povinelli, Lucia Jacobs and Eric Schwitzgebel, as well as the editors and reviewers. Address for correspondence: Alison Gopnik, Dept. of Psychology, University of California at Berkeley, Berkeley, CA, 94720.

There is a lust of the mind, that, by a perserverance of delight in the continual and indefatigable generation of knowledge, exceedeth the short vehemence of carnal pleasure.
Thomas Hobbes

What is explanation? Both our ordinary everyday concept of explanation and refinements of the concept in the philosophy of science seem to have a curiously circular quality. Theories are good because they explain things, but explaining things turns out to be an awful lot like having theories of them. We explain something, we are told, when we characterize it in terms of some set of abstract underlying laws and entities. But this characterization seems to reduce to having a theory itself. Alternatively, we may say that we explain something when it satisfies our explanation-seeking curiosity. But then explanation-seeking curiosity seems only to be definable as that curiosity which explanation satisfies. The ordinary concept of explanation seems to involve both a kind of knowledge, hence the link to theories, and also a distinctive phenomenology, hence the link to the feeling of curiosity. This sort of amalgam of cognition and phenomenology is quite characteristic of our ordinary folk psychological concepts, of course.

What I want to propose in this paper is not really a further explication of either our ordinary notion of explanation or the notion as it is used in science. In particular, I do not think there is some set of necessary and sufficient features that will account for all the things we call "explanation", anymore than there is a set of necessary and sufficient features that will account for all the things we call "bird". Instead, I want to suggest that the everyday notion of explanation might be a jumping off point for a non-circular and interesting cognitive concept of explanation. Moreover, thinking about explanation suggests some interesting, if still untested, hypotheses about how our cognitive system might work. The concept of explanation I will propose, like the everyday concept, includes both cognitive and phenomenological elements but I will propose a particular, and more precise, relation between them. I will suggest that explanation may be understood as the distinctive phenomenological mark of the operation of a special representational system. I will call this system the theory-formation system. This system was designed by evolution to construct what I will call "causal maps". Causal maps are abstract, coherent, defeasible representations of the causal structure of the world around us. Moreover, the distinctive phenomenology of the theory-formation system impels us to action as well as to knowledge, it reflects a sort of theory-formation drive. Hence the title of this paper. My hypothesis will be that explanation is to theory-formation as orgasm is to reproduction. It is the phenomenological mark of the fulfillment of an evolutionarily determined drive. From our phenomenological point of view, it may seem to us that we construct and use theories in order to achieve explanation or have sex in order to achieve orgasm. From an evolutionary point of view, however, the relation is reversed, we experience orgasms and explanations to ensure that we make babies and theories. Moreover, I will suggest that the distinctive phenomenology of explanation may be important methodologically. By using phenomenological evidence we may be able to identify when and how the theory-formation system is operating.

The theory-formation system

Within the past ten years developmental psychologists have increasingly used the model of scientific theory change to characterize cognitive development ((Carey, 1985);(Keil, 1987);(Gopnik, 1988);(Gelman & Wellman, 1991);(Wellman, 1990)). I have called this idea the theory theory. (For recent expositions see(Bartsch & Wellman, 1995);(Gopnik & Wellman, 1994);(Gopnik & Meltzoff, 1997)). This approach has been consistently productive in explaining the child's developing understanding of the mind and the world. One way of interpreting the theory theory is as a claim about a special and distinctive set of human representational capacities. These capacities are most visible in scientists and in children, but they are part of every human being's basic cognitive equipment. On this view science is successful not because it invents special new cognitive devices (though, of course, this is part of what science does) but because it capitalizes on a more basic human cognitive capacity. The analogy to science has two aspects. First, children's knowledge is structured in a theory-like way, and second, that knowledge changes in a way that is analogous to theory change in science.

We have suggested that this theory formation system may have evolved specifically to allow human children to learn. Human beings' evolutionary advantage stems from our ability to adapt our behavior to a very wide variety of environments. In turn, this depends on our ability to learn swiftly and efficiently about the particular physical and social environment we grow up in. The long, protected immaturity of human children gives them an opportunity to infer the particular structure of the world around them. We suggest that the powerful and flexible theory-formation abilities we see in childhood evolved in order to make this learning possible. On this view, science takes advantage of these basic abilities in a more socially organized way and applies them to new types of problems and domains. Science is thus a kind of epiphenomenon of cognitive development. Our motto is that it is not that children are little scientists but that scientists are big children.

The theory theory is still relatively new and controversial and there will not be space here for a detailed explication or defense of the idea (for such an explication and defense see Gopnik & Meltzoff, 1997). Instead, I will assume that the general idea is correct, that is, that there are general mechanisms of cognitive development that are very much like the mechanisms of theory change in science and develop some hypotheses about explanation within that general framework. Moreover, thinking about theory-formation in the context of explanation may also point to some interesting hypotheses about the details of the theory-formation system itself. Much of the work that has been done in the field so far, including my own work, has primarily been devoted to elaborating on the similarities between scientific theory change and cognitive development. We have constructed lists of features of science, its abstract, coherent structure, its interwoven web of laws and entities, its capacity for dynamic change, and pointed to similar features of children' s developing understanding of the world around them. If we assume that the general parallel to science is correct, however, we can go on to the question of developing a more detailed understanding of the theory-formation system as a system in its own right. In what follows I will point to some possible features of the system that are especially relevant to explanation and have not been sufficiently emphasized before.

Theories as causal maps. One way of thinking about the theory formation system is that it is a way of uncovering the underlying causal structure of the world from perceptual input. In this respect, the theory formation system is analogous to our systems for representing objects in space. The visual system takes retinal input and transforms it into representations of objects moving in space. While these representations of objects are not perfectly veridical, they at least approach a greater level of veridicality than the retinal input itself, and presumably this fact explains the evolution of perception and spatial cognition. Similarly, we can think of the theory formation system as a system that takes the input of the perceptual system and transforms it into representations of the underlying causal structure of the world. The theory formation system is designed to go beyond perceptual representations just as perceptual representations go beyond sensory input. The representations that result, representations formulated in terms of abstract theoretical entities and laws, more closely approach the actual causal structure of the world than the representations of perception. Again, this is true both at the level of the exercise and use of theories, and at the level of theory change. To apply a theory to some pattern of evidence is to assign the evidence a particular causal representation. Theories change in the face of evidence in order to give better causal representations.

In our earlier formulations of the theory theory, the causal character of theory formation was simply seen as one feature among many. I think this is probably partly because scientific theories have been our model. Science may include cases of theories, and explanations, that we need not think of as causal, though clearly causal claims play an important, indeed a central role in most scientific theories and explanations (see e.g.(Cartwright, 1989)(Salmon, 1984). It increasingly seems to me, however, that uncovering causes is the central feature of the theory-formation system from an evolutionary point of view, and that other features of theories derive from it, just as uncovering the spatial character of moving objects seems to be central to the visual system.

Of course, theory formation is far from being the only way that the cognitive system can uncover causal structure. Just as there are many ways in which perceptual systems detect the spatial character of objects, and no creature could survive without some abilities to detect the external world, so we might think of even the most primitive conditioning capacities as a kind of causal detection device. However, just as our spatial cognition is different from the simple detection systems of other creatures, so it seems plausible that our systems for detecting causal structure are also different.

Work on evolutionary cognition suggests an interesting analogy. It is well known that different species of animals, even closely related species, may use quite different strategies and systems to uncover the spatial character of the world around them. Some of an animal's understanding of space may be hard-wired in the perceptual system, for example, we may be hard-wired to translate two-dimensional retinal information into three-dimensional representations.

But animals may also use information to learn about the specific character of their spatial environment. In particular,(O'Keefe & Nadel, 1978)) pointed out that some species do this by keeping track of the effects of their movements in the world, and using that information to guide their future movements. If turning first to the left and then to the right leads to a food reward, animals will repeat that sequence of movements. They use a kind of egocentric spatial navigation system. However, other species use what O'Keefe and Nadel (1978) called "spatial maps". These species also learn from their movement through space. However, they do so by constructing a coherent, non-egocentric and often quite complex picture of the spatial relations among objects, a kind of map. As they move through the world they update the information in this spatial map and they may revise and change it as they learn more about their environment, spatial maps aren't hard-wired. Animals with this kind of representation system have some real advantages. Animals with spatial maps, for example, can find food in a maze they have previously explored even if they are placed in a different starting point, and so can't simply reproduce their earlier movements .

We can make a parallel distinction between a kind of egocentric causal navigation system and a kind of causal map (see(Campbell, 1994), 1994). One interesting possibility, in particular, is that other animals primarily understand causality in terms of the effects of their own actions on the world. In a process like operant conditioning they record the causal effects of their actions on events in the world and modify their actions accordingly. In contrast, human beings seem to equate the causal powers of their own actions and those of objects independent of them. They construct and update causal maps of their environments.

Causal maps have the same advantages as spatial maps. Once we represent the causal relations among ourselves, our conspecifics and objects, we can intervene in a much wider variety of ways to get a particular result. For example, we may imitate what we see another animal do, since we assume the causal effects of our actions will be like theirs. We may be able to do this even if we have never performed the action ourselves before.

Causal maps also let us use tools in an insightful rather than just a trial and error way. If we understand the causal relations among objects, independent of us, we can immediately use one object to cause another to do something, even if we have never used the tool that way before.

In fact, a number of recent empirical studies of primate cognition suggest that human beings may be specialized in just this way. While chimpanzees, for example, are extremely intelligent and very good at detecting contingencies between their actions and the effects of those actions, they seem unable to either imitate the solutions of other animals or to use tools insightfully ((Povinelli , in press)(Tomasello & Call, 1997)). They don't seem to construct causal maps.

We can think of a theory as precisely this sort of causal map. A theory postulates a complex but coherent set of causal entities, the theoretical entities, and specifies causal relations among them, the laws. Just as a spatial map allows for new predictions and interventions in the world, so a causal map allows for a wide range of causal predictions and interventions, including experiments. And just as theories are revisable in the light of new experience, rather than hard-wired, so causal maps, like spatial ones, could be updated and revised.

Recent work in developmental psychology in the context of the theory theory suggests that children are sensitive to the underlying causal structure of the world, and seek to form new causal representations at a much earlier age than we had previously supposed. There is evidence for causal understanding even in infancy ((Leslie & Keeble, 1987; Oakes & Cohen, 1994) Gopnik & Meltzoff, 1997). By 2 1/2 or three children show extensive causal reasoning both about living things and about psychological processes ((Gelman & Wellman, 1991)). Moreover, nine-month-old infants, apparently unlike chimpanzees, can learn a new action on an object by imitation ((Meltzoff, 1988)). 18-month-old infants, again apparently unlike chimpanzees, can use tools insightfully ((Piaget, 1952);(Gopnik & Meltzoff, 1984))

In recent empirical work, we have shown that, by three, children will override perceptual information in favor of causal information when they classify objects ((Gopnik & Sobel,)1997). In a series of experiments we showed children a "blicket detector". The blicket detector is a machine that lit up and played music when objects were placed upon it. Perceptually identical objects were placed on the machine with differential causal effects, some made the machine light up and some did not. Similarly, perceptually different objects might have the same effect on the machine. Children were shown an object that had set off the machine and were told it was a "blicket". Then we asked children to show us the other blicket. Even three-year-olds were willing to override perceptual features in favor of causal powers, they said that the other object that had set off the machine was a blicket, even when it was not perceptually similar to the original object. Moreover, a control condition demonstrated that this was genuinely causal reasoning. The children did not classify objects together when the experimenter simply held them over the machine and pressed a button that made the machine light up. A mere association between the object and the event was not enough to influence the children's categorization. Like scientists these young children seemed to organize the world in terms of the underlying causal powers of objects and to seek explanations of new causal relations.

Causal maps and computation: Could theories be Bayes nets?

One question about the theory theory has always been whether there is any perspicuous way that it could be implemented computationally. In earlier work, we have simply seen that as a question for the future, no computational representation that we knew of seemed particularly likely as a candidate. It is still true that this is very much a question for the future but there is a recent candidate that may be interesting. The computational formalism called Bayes nets have precisely been used to represent networks of causal relations among events. Moreover, there are formal results which suggest how such structures may be derived from empirical evidence of the sort that would be available to children (such as conditional dependencies and independencies among events). Similarly, formal results show how these representations may be used to predict future events, to interpret current evidence and to design appropriate experimental interventions (Glymour, this volume). Finally, at least some adult judgments of causal powers seem to be well represented by this formalism (Cheng, this volume). It is conceivable that the causal maps we have described are represented as Bayes nets, and that children use similar algorithms to learn the structure of these nets and to use them for prediction and control.

Domain-specificity and generality. Another important feature of the theory-formation system is that it combines domain-specific and domain-general mechanisms. We have proposed that infants have innate and quite specific theories of a number of particular domains. On our view, however, these initial theories are subject to revision and change, and the inductive mechanisms that lead to that change may be quite generally applicable. As a result of differences in their initial theories, and in patterns of evidence, the specific content of children's later theories may be also be quite different in different domains. Similarly, in science, the basic entities and laws of physics may be quite different from those of evolutionary biology. On the other hand, just as in science, the processes of hypothesis-testing, prediction, falsification and evidence-gathering will be quite similar across domains. Moreover, as in science, the assumption that there is some underlying causal structure to be discovered remains constant across domains. Conceptual change that is the result of these mechanisms may and often will take place within a particular domain. However, there may be radical restructurings of the domains with development, again, just as in science. One well-known if controversial example is the emergence of a distinctive folk-biology in the school-age years(Carey, 1985). In our own work we have suggested that children initially have a theory of action that includes aspects of both folk physics and folk psychology (Gopnik & Meltzoff, 1997).

Change. The dynamic properties of theories are also distinctive. The representations of perceptual systems seem to be relatively fixed, at least in adulthood, and perhaps in childhood as well. For example, many fundamental features of the visual system such as basic object segregation and distance perception appear to be in place at birth. When the systems do change as a result of experience they seem to do so in a fairly restricted set of ways. Typically, this process is described in terms of processes like triggering or parameter-setting, rather than in terms of the inductive inferences that are the result of new experiences in science.

In contrast, it is part of the very nature of theory-formation systems that they are perpetually in flux. In general, the perceptual system seems to work by taking in sensory input and trying to assign some coherent representation to that input. However, if the system can't find such a representation, it simply stops. The theory formation system also seeks to find a coherent causal representation of perceptual input. However, when the system fails to find such a representation, in enough cases and over a long enough period, it restructures both the very procedures it uses to assign causal representations, and the kinds of causal representations it assigns. In other words, the theory changes. The system takes all forms of evidence into consideration, and seeks a consistent causal account of objects on that basis. The theory-formation system is perpetually faced with counter-evidence and perpetually revises theories on that basis.

The representations that the perceptual system will come up with are highly constrained. The theory-formation system is much less constrained, in the most extreme cases, for example, it may come up with the representations of relativistic physics or quantum mechanics. It is a representational system that both computes representations from inputs and also, and in consequence, alters the way it computes new representations from new inputs.

In recent empirical work, we have explored the dynamic character of children's changing conceptions of the world. In particular, in a series of training experiments, Slaughter & I showed that we could induce general conceptual changes in three-year-olds' understanding of the mind by presenting them with relevant counter-evidence to their earlier theories(Slaughter & Gopnik, 1996). Children who received evidence that was conceptually relevant showed a new understanding of the mind, and extended that understanding to contexts that were very different from the contexts in which they had been trained.

Exploration, experimentation, and the theory drive. If we think for a minute about the dynamic features of theory change, moreover, we can see that using theories, and, to an even greater extent, changing theories, generally involves active intervention in the world. Schwitzgebel has suggested that theories are closely connected to what he calls a "drive to explain" ((Schwitzgebel, 1997)).

Let's return to our earlier analogy with spatial maps. Creatures who use such maps also display distinctive patterns of spatial exploration. A rat who constructs spatial maps, for example, will systematically explore a new spatial environment, even if that exploration has no immediate pay-off. Presumably the expenditure of energy involved in free-ranging exploration has its pay-off in the long term predictive advantages of constructing a spatial map.

Similarly, we might expect that creatures who depend on constructing causal maps would intervene in the environment in a way that lets them determine its causal structure. The most obvious example of such intervention is the process of experimentation. In experimentation we systematically act on the world in a way that is designed to obtain evidence relevant to the theoretical problems we are trying to solve. Sometimes we may experiment to see how a particular piece of evidence should be interpreted in terms of an established theory. Sometimes, we may do so in search of a new, more adequate theory. Sometimes, particularly in organized science, this process of experimentation is designed to carefully and systematically elicit particular pieces of evidence. Often, however, the process is more akin to exploration, to what scientists disparagingly call a fishing expedition. This sort of experimental intervention in the world is notoriously one of the most powerful ways of determining its causal structure.

We see both extensive causal exploration and even some more systematic experimentation in children's spontaneous play. Piaget, for example, charted how object manipulation and play were related to cognitive change in infancy(Piaget, 1962). Indeed, Piaget defined "play" as the process of assimilation, that is, what we would now call the process by which evidence is interpreted in terms of existing theories. We have suggested that infants fondness for "drop the spoon" games at 15 months is related to their changing conception of space, that the earlier hide and seek games are connected to object-concept understanding and that the later "terrible twos" behavior is related to an understanding of differences in desires (Gopnik & Meltzoff, 1997). In each of these cases infants actively try to produce new phenomena, phenomena that are at the leading edge of their theory formation, in an apparently experimental way.

In fact, the degree to which infants and children actively and spontaneously explore the world is almost a cliché of parenting; we talk about how toddlers "get into everything" or how preschoolers are always asking "why?" We "childproof" our houses to try to keep this exploratory behavior under control. While we take this for granted, it is a striking fact about childhood ecology. These exploratory and experimental behaviors require enormous expenditures of energy and they have little obvious function, in fact, they may be, superficially at least, quite dysfunctional. Not only is the baby expending enormous energy on getting to the light-bulb or the lipstick, we adults expend enormous energy trying to keep him away from it. Interestingly, work from a very different tradition of developmental psychology, namely attachment theory, supports this picture of a fundamental exploratory drive. In fact, in its original formulation, attachment between infants and mothers was supposed to function precisely to mediate between infants need for protection and security and their equally strong need to explore and manipulate objects(Bowlby, 1969).

If active intervention in the world is necessary to infer its causal structure, then there needs to be some sort of motivational system to bring about such intervention. In formal organized science, of course, we have an elaborate set of incentives and rewards to ensure that such intervention takes place. Children, however, and to a lesser extent ordinary adults, seem to continue to explore and experiment with the world quite independently of such external incentives. Children, in particular, spontaneously attempt to interpret patterns of evidence in terms of the underlying causal representations of their theories, they spontaneously reorganize their theories when this sort of interpretation fails, and they spontaneously expend energy on the sort of active exploration and experimentation that this entails.

There is a payoff for this activity in the long run, of course. Getting a veridical causal map of the world allows for a wide range of accurate and nonobvious predictions, and these accurate predictions, in turn, allow one to accomplish other types of goals that are more directly related to survival. The relation between assigning the causal interpretations and making the useful predictions may be quite long-term and indirect, however (as scientists are always assuring congressmen). Again, the analogy to sexual drives should be obvious. Nature ensures that we do something that will be good for us (or at least our genes) in the long run, by making it fun (or at least compelling) in the short-run.

Theory-formation and the experience of explanation.

The phenomenology of explanation. What I have proposed above and elsewhere, then, is that there is a special representational system that has a number of distinctive qualities. How is explanation related to the operation of this theory-formation system? We could simply identify explanation with the successful operation of the theory-formation system. In particular, we could define explanation as a relation between theories and evidence. In the terms I've been using, we might say, very roughly, that a theory explains evidence when it assigns the evidence a particular causal representation. There is a long tradition in the philosophy of science, dating back to Hempel (Hempel, 1965) that follows this line.

Intuitively, however, this way of treating explanation seems to leave something out. Explanation is a goal-directed human activity. It depends on what is relevant or important to the explainer, it satisfies a special kind of explanation-seeking curiosity, it answers "why?" questions. Again there is a tradition of pragmatic accounts of explanation in the philosophical literature that emphasize this aspect of explanation (e.g.(Bromberger, 1965);(Van Frassen, 1980)). As is often true in the philosophy of science there seems to be little purchase between the two traditions, logical and pragmatic.

While pragmatic accounts are rarely phrased in this way, I think what they point to has as much or more to do with phenomenology than it does with pragmatics per se. What the purely logical view leaves out is that there is something that it is like to have or seek an explanation. I want to suggest that there is a distinctive phenomenology of explanation.

Such phenomenological claims are, of course, difficult to justify initially except by appeal to intuitions. Moreover, in the case of sophisticated adults almost any particular experience will reflect a complex mixture of different types of phenomenology. Visual experience may reflect extensive implicit inferences as well as reflecting the operation of the visual system itself. The experience of an emotion like anger may run the gamut from cold withdrawal to helpless depression to irresistible rage. Nevertheless, it seems right to say that there is a phenomenology of anger and that that phenomenology is consistently related to a particular set of psychological functions, and that there is a phenomenology of vision that is related to the operation of the visual system. In fact in the case of anger there are even suggestions that there is a "basic emotion", an evolutionarily determined complex of facial expression, psychophysiology, and phenomenology(Ekman, 1992).

In the same way, I want to suggest that there is a distinctive phenomenology associated with explanation. The phenomenology involves both the search for explanation and a recognition of the fact that an explanation has been reached. We might call them the "hmm" and the "aha". In English they seem to be expressed by "why?" and "because". These experiences are obviously close to what we more broadly call curiosity or interest but they are not identical with them. We may engage in purely exploratory behaviors (the desire to open the locked cupboard, say, or climb the mountain, or see around the bend) that have no "aha" at the end of them. Often they are connected to goal-directed or problem-solving behavior but they do not simply reflect the desire satisfaction that comes from achieving a goal. We may blunder our way to the exit, or use trial and error to find the right key to delete the file, and be happy we have done so, but without any "aha".

Conversely, we may experience a splendid moment of illumination as we realize just exactly why it is thoroughly impossible to get what we want.

This explanatory phenomenology also appears to be strikingly domain general. We seek and are satisfied by explanations of physical objects, animate beings, psychological agents, and even social groups. We seek and are satisfied by explanations in terms of physical laws, biological processes, reasons, or rules. At least first-person, the aha of understanding why the wiring doesn't work seems quite similar to the aha of understanding why the electrician won't tell you why the wiring doesn't work. Even in children "Why?" and "Because" seem to cut across domains in this way ((Schult & Wellman, in press);(Wellman, Hickling, and Schult, In Press)).

Moreover, explanation, unlike many other experiences, seems to combine some of the properties of both cognitive and motivational phenomenology. Like vision, but unlike, say, anxiety or depression, or even hunger or lust, explanation seems intrinsically referential, an explanation must be of or about something in particular (we can't usually experience free-floating explanation, or even free-floating explanatory curiosity, anymore than we can experience free-floating vision). In fact, explanation, even more than vision, seems to require some sort of propositional representational capacity.

But explanation also, and unlike vision, has some of the phenomenological character of a motivational or drive system. We not only know an explanation when we have one, we want explanations, and we are satisfied when we get them. Even in adults, the "hmm" is, to varying degrees, an unsettling, disturbing, and arousing experience, one that seems to compel us to some sort of resolution and action (the two great resources by which popular fiction holds our attention are titillation and mystery, nothing like unsatisfied fundamental drives to keep the pages turning). Conversely, finding an explanation for something is accompanied by a satisfaction that goes beyond the merely cognitive.

In children, the drive for explanation may even override other more obvious and straightforward motivations. We have suggested that in "the terrible twos" children are conducting experiments to understand the nature of differences in desires, even though the immediate consequence of those tests is maternal rage. This may even be true in adults, as when, in Proust's Swann's Way, Swann compulsively tests Odette in search of her secret life, in spite of the emotional and practical pain this will cause him, a rather advanced case of the terrible twos.

It even seems possible that some aspects of explanatory phenomenology might qualify as a kind of "basic emotion". Surprise and interest, phenomena very closely related to the "hmm" are, in fact, often taken to be basic emotions. There is some evidence for a distinctive and universal facial expressions associated with these states that are distinct from the mere reflex of a startle response, or from other emotions like anger or fear. The "aha" in contrast is often accompanied by a positive expression of joy. This expression appears to be less clearly distinct from the expression of other positive emotions, but this is characteristic of positive expressions in general (see Ekman, 1992). In our own work with children, even with infants, we see a distinctive set of affective responses and facial expressions that accompany exploration and problem-solving. In our experiments, children who are in the intermediate stages of theory formation often exhibit a great deal of puzzlement and even distress, furrowed brows, pursed mouths. This contrasts with the behavior of these same children on easier tasks, and with the behavior of children who are firmly in the grip of an earlier or later theory. Children who are presented with problems that are relevant to a newly formed theory, in contrast, often display intense satisfaction and joy.

This sort of "cognitive emotion" has been surprisingly neglected in the psychological literature, perhaps because of the old oppositions between emotion and cognition, or perhaps because it is more common and dramatic in children than in adults. Nevertheless, evidence of this sort of phenomenon appears in variety of quite disparate contexts, even in the psychology of adults. A "theory drive", for example, seems to be involved in the Zeigarnik effect, or in the social psychologist's notion of a "need for closure" (see e.g.(Kruglanski, 1989)). I suggest that this sort of experience is at least an important part of what we talk about when we talk about explanation.

The contingency of explanation. So far I have claimed that there is a special representational system, the theory-formation system, and that it is accompanied by a kind of theory drive. I have also suggested that there is a distinctive set of experiences that are at least part of what we mean by explanation. Now I want to talk about the relation between the cognitive system and the phenomenology. It should be clear by now that I think that the explanatory phenomenology, the "hmm" and the "aha", is closely related to the operation of the theory formation system. The "hmm" is the way we feel when we are presented with evidence to which the theory has not yet assigned a causal representation. The "aha" is the way we feel when such a representation is assigned, either by applying the theory or by revising it.

But isn't this just returning to the circularity I began with? I want to suggest that is not. In our folk psychology, and indeed in conceptual analysis in philosophy, the connection between phenomenology, psychological structure and function often appears to be transparent and necessary. Before we knew in detail about the visual system, it might have seemed obvious that visual experience and object perception were one and the same thing. Similarly, accounts of explanation often seem to move back and forth between the description of the phenomenology of explanation and its cognitive function, thus, for example, the tension between logical and pragmatic accounts in philosophy. I want to argue that, in fact, the relation between the phenomenology of explanation and its cognitive function is quite complex and contingent.

To see this, let us return to the example of vision. 100 years of perceptual research have shown us that while visual cognition and visual experience are closely related they are also conceptually and empirically separable. It is possible to have a system that functions cognitively like the visual system, but that lacks visual phenomenology, as in the case of computer vision or blindsight. Alternately, visual phenomenology may be produced by systems outside the visual system itself, as in the case of images, hallucinations and certain kinds of illusions, such as the "top-down" illusion of experts that they "see" things that they actually infer. The same point may be made even more obviously about the case of sexual phenomenology. Even though the canonical example of sexual phenomenology is connected with sexual activities that lead to reproduction, it is notoriously true that sexual phenomenology is in another sense only contingently related to those activities. Sex need not be attended by desire, and desire may be associated with everything from shoes to conversation to deep-sea diving equipment.

In the examples I was just discussing, the phenomenology of vision or sexual desire might occur in the absence of any perceptual or sexual activity. In other rather different cases, both the phenomenology and the appropriate activity might occur, but the system might not successfully fulfill its evolutionary function. While the function of the visual system is to obtain a veridical picture of the outside world, in practice, the system will often end up with non-veridical representations. The fact that there are visual illusions is not an indicator that perception, in general, is unveridical, or an argument against the idea that the perceptual system evolved because it is veridical in general and over the long run. Again the case is even clearer for sexual phenomenology. While the very existence of sexual phenomenology and activity depends on its reproductive function over the long evolutionary run, there may be no connection at all for individual humans in the grip of desire, and most experiences of desire will not result in reproduction. There is an interesting additional aspect of this point that is not as frequently considered and may be particularly relevant here. A system may evolve, in particular, to serve a function at one developmental phase of an organism's life and yet continue to operate at some other phase. We continue to have orgasms after menopause, and have breasts and wombs even when we are not pregnant.

Thinking about explanation in this way may help to resolve what seem like puzzles and paradoxes. What I am suggesting is that the phenomenology of explanation is, in the canonical case, connected with the operation of a distinctive cognitive system, the theory-formation system. Moreover, that theory-formation system evolved because in general and over the long run, and especially in childhood, it gives us a more veridical picture of the causal structure of the world around us. In a particular case, however, explanatory phenomenology may be divorced from the operation of this cognitive system. Perhaps the most dramatic cases of this, the equivalent of visual hallucinations or sexual fetishes, are certain types of mystical experiences. Some cases of mystical experience seem to simply involve being washed over by a sort of generalized positive affect. But, in at least some such cases, the experience is more pointedly cognitive, the mystic suddenly experiences all the positive affective phenomenology of explanation with no apparent cognitive content. Suddenly it all becomes clear, all at once everything makes sense. Something like this also seems to happen in certain kinds of paranoia. Less dramatic but still striking instances of this are the "middle of the night" solutions which dissolve as we decipher our scribbled bedside notes.

Conversely, it may be possible to engage in something like theory use and theory formation without explanatory phenomenology. It might be argued that automated theorem-provers , of the sort whose results are published in physics journals, do just that. Sadly, the same may be true for the scientist who has been thoroughly corrupted by the system of social incentives, so that the greed for the Nobel utterly outweighs the joy of discovery. Indeed, given the complex social organization of science it may be that a whole group of scientists scattered over many places comes upon the correct theoretical answer to a question without any single scientist experiencing any phenomenology at all. (It is striking, and comforting, however, to see the phenomenology of explanation persist even in relatively sophisticated and socially organized scientists. The scientists in the recent Mars probes almost without exception described their joy and excitement by saying it felt like being a little kid again. None of them said it felt like getting a raise).

It is also possible that the theory formation system may genuinely operate, and the related explanatory phenomenology may occur, without achieving a more veridical causal map. The function of theory-formation may be to obtain veridical causal maps, in general and over the long run, and particularly in childhood, but this is perfectly compatible with the idea that the products of theory formation are often not veridical. The function of sex is still to reproduce even if reproduction doesn't occur in the vast majority of cases of sex. These cases would be more like visual illusions than hallucinations, more like having sex on the pill than like fetishism.

Some of the notorious cognitive illusions offered by Kahneman and Tversky and others may be instances of this sort of case. Magical, mythical and religious explanation, and certain types of social explanation, may also be examples. This may help resolve the otherwise puzzling question of whether having a bad explanation or a pseudo-explanation is the same as having no explanation at all. From the sort of normative cognitive view of philosophy of science, this may indeed be true. From the psychological point of view I am developing here, however, genuine explanation, and indeed genuine theory formation might take place, and yet the outcome might be normatively deficient, even very normatively deficient, and even very normatively deficient much of the time. This is perfectly consistent with the view that the system evolved because, in general, and over the long run, and especially in childhood, it gives us veridical information about the causal structure of the world.

It appears that one of the differences, perhaps the most important cognitive difference, between organized science and spontaneous theory-formation is precisely that science contains additional normative devices that are designed to supplement the basic cognitive devices of the theory-formation system, and to protect them from error. We might think of science as a kind of cognitive optometry, a system that takes the devices we usually use to obtain a veridical picture of the world and corrects the flaws and distortions of those devices. The fact that most people over forty wear glasses, is not, however, usually taken as an indictment of the visual system. In fact, the analogy might even be taken further, perhaps science compensates for our deteriorating adult theory-formation abilities the way optometry compensates for our deteriorating adult vision. By twenty, most of us have done all the theory-formation evolution requires, by forty, most of us have done just about everything that evolution requires.

On this view, then, the relation between explanation and theory-formation is close and principled, but is not circular. We can have theory-formation without explanation and vice-versa. Nevertheless, most of the time and overall, explanatory phenomenology will be attended by theory-like cognition, and there is a principled psychological reason for this.

Methodological issues.

This view of explanation as the phenomenological mark of a cognitive process also has methodological implications. Like other cognitive scientists of my generation, I grew up a functionalist. The basic tenet of functionalism was that cognitive science would proceed by characterizing the input and output to the human mind and provide a computational account of the relations between them. The difficulties, both practical and philosophical, of implementing that project have, however, become increasingly clear. Mere functional information seems to deeply underdetermine the possible computational accounts of the mind. Many have turned to neuroscience for salvation in this dilemna. But while neurological accounts may indeed help us to untangle competing accounts of mental structure in low-level cognition like vision and motor control, they appear to be much less applicable to higher-order thinking, reasoning and problem-solving.

I want to argue that there is another source of evidence in cognitive science. The evidence comes from phenomenology; from the internal structure of our conscious experience. Recently there has been a great deal of speculation in cognitive science about Capital-C Consciousness, the Big Problem of how phenomenology is possible at all, and how it relates in general, to the functional structure of the mind. We do not seem close to a solution. For the entire history of cognitive science, however, specific relations between conscious experience and function have been the source of some of the most productive work in the field, even though the cognitive scientists who use this evidence have kept pretty quiet about it.

The most striking example is vision science, arguably the most productive branch of psychology in this century. Psychophysicists, of course, never adhered to a strictly functionalist program, though often they pretended to, and continue to pretend to. (A famous and spectacularly hard-nosed psychophysicist I know regularly shows slides of various psychophysical functions he is investigating. The slides have initials in the corners, identifying the subjects, and establishing that this is an objective scientific enterprise. But somehow, the initials always correspond to those of the psychophysicist and his coinvestigators). Psychologists in psychophysics and perception always began from the structure of conscious percepts and then produced accounts of the relation between those percepts and underlying functional, computational and, most recently, neurological structure. In fact, the phenomenology in some sense even defined the field of inquiry. We don't need an elaborate set of inferences to work out whether a particular phenomenon is due to the visual system, as opposed, say, to the auditory or kinesthetic system. The evidence of phenomenology itself gives us at least a very good first pass at an answer. More generally, without this phenomenological evidence our understanding of vision would be severely impoverished.

Importantly, however, while the psychophysicists crucially used phenomenological evidence, they never assumed that that evidence constituted mental structure itself. Rather they outlined quite complex and contingent relations between the phenomenology of visual experience and the functional relations between the mind and the world. Indeed it is arguable that the first great cognitive revolution in psychology, sixty years before Chomsky and Bruner, and evident in psychologists as diverse as Freud, Piaget and the Gestaltists, came when psychologists began to treat phenomenology as evidence for psychological structure, with complex relations to underlying theoretical structures that were themselves unconscious. This contrasts with the practice of earlier philosophers of mind, from Descartes to Hume to Brentano, who still assumed that the theoretical entities of psychology would themselves be conscious experiences.

I would argue in a similar way that, in spite of the contingent relation between explanatory phenomenology and theory-like representations, we can use the phenomenology as a guide to the underlying psychological structures. In particular, a purely functionalist account may make it difficult to discriminate the theory-formation system from other types of representational systems. For example, as Irvin Rock elegantly demonstrated, the formal structure of perceptual representations may involve “inferences” about underlying theoretical entities from the "data" of sensation. When the moon looks larger at the horizon than the zenith, it is because the perceptual system draws a set of "inferences” about the relation between size and distance. Similarly, and more generally "modular" representational systems may mimic the functional architecture of theories. In fact, modularity theorists will sometimes talk about "theories" in just this way, a "theory of mind module"(Leslie, 1987), in this sense, is not an oxymoron. Chomsky famously characterized the highly automatic, unconscious and indefeasible processes of syntactic processing as a kind of knowledge, if not quite as a theory.

Elsewhere I have suggested that developmental evidence may be crucial in discriminating between modules and theories. Another important type of evidence may come from examinations of the phenomenology of theories, and from explanation, in particular. While modules may mimic the architecture of theories they are strikingly lacking in explanatory phenomenology. There is no internal "aha" when we look at the large moon on the horizon, just a big moon. Similarly, we do not seem to be driven to parse in the way that we are driven to explain, we just hear (or don't hear) a meaningful sentence. Again, I do not want to suggest that explanatory phenomenology is a necessary indicator of theory-like processes, just that it is a reliable indicator. In particular, it may be that very well-practiced and well-established theories often lose their explanatory character. Nevertheless, in general it seems to me that theory-like knowledge in adults will at least potentially support explanatory phenomenology. Even in the case of a well-practiced theory like folk psychology it should be possible to formulate and answer "why" questions and “because” answers, and to experience the “hmm” and “aha”.

These ideas may have particularly interesting implications for developmental psychology. In the past, developmental psychologists have been rather shy about attributing phenomenology to infants and children, perhaps because we have enough trouble getting our colleagues to agree that children have minds at all. The conventional phrase, for example, is to say that children have "implicit" rather than "explicit" theories. This shyness, however, seems more political than rational. There is every reason to believe that infants and children have rich phenomenal experiences even if those experiences are in many ways different from our own. Do they experience explanatory phenomenology?

Recent work by Wellman and his colleagues suggests that at least one index of explanatory phenomenology, explicit linguistic explanations and requests for explanation, is in place much earlier than we previously supposed. In adults, we think of explicit “why” questions and “because” answers as the quintessential index of explanation, just as we take color reports to be an index of visual phenomenology. In an analysis of children's spontaneous speech recorded from the CHILDES data base(Wellman et al., In Press), they found explanations and requests for explanation as early as two years old, almost as soon, in fact, as the children could speak at all. These explanations also changed, and changed in interesting ways, as the children grew older. The changes seemed to reflect changes that had independently been attributed to theory-formation. For example, two-year-olds who were more likely to explain behavior in terms of desires and perceptions, while three and four-year olds began to explain behavior in terms of beliefs. A similar shift occurs in children's predictions about behavior at about the same time (Bartsch & Wellman, 1995).

Are there other indicators of explanatory phenomenology in young children? The fact that explanatory phenomenology in adults is accompanied by distinctive affective and motivational states may provide a clue. We suggested above that we see very similar patterns of affect, action and expression in young children and even in infants. Even infants show characteristic patterns of surprise, interest and joy, and characteristic attempts at exploration and experimentation in some circumstances.

These patterns may be an interesting tool in sorting out whether infants form theories and which of their other behaviors are indicative of theory-formation. In particular, there are currently interesting discrepancies in the literature between the inferences we can draw from infants' active behavior and the inferences we can draw from their visual attention. The most well-known case is the discrepancy between the object-knowledge that children demonstrate in their search behavior and the knowledge they demonstrate in paradigms where they are faced with an "impossible event" and their looking times are recorded (see(Baillargeon, 1993),(Spelke, Breinlinger, Macomber, and Jacobson, 1992)). The children's behavior in the looking time paradigms is often described in terms of explanatory phenomenology, the children are said to be surprised or puzzled by the unexpected event, or to have predicted that the event would not take place and to be registering the violation of their prediction. Similarly, the children's later search behavior has also been described in these terms; children predict that the object is in a location, search for it there and are surprised and puzzled when it is not there. However, neither the child's actions by themselves nor their visual attention by itself necessarily supports this interpretation. That fact has been at the center of alternative attempts to resolve these discrepancies. The child's looking time may reflect some automatic modular perceptual preferences that are detached from theory-like knowledge. Alternately, the child's actions may be the result of some automatic habit rather than the result of an inference about the object’s location.

Searching for signs of explanatory phenomenology in either case might be helpful. Are children genuinely surprised by impossible events? Do they furrow their brows or show distress? Do they show signs of exploratory behavior that might be relevant to the conceptual problems such events pose? Do they smile if a more coherent explanation of the event is made available to them? If these behaviors do accompany search or visual attention, we are more licensed in concluding that they reflect something like the operation of a theory-formation system.

More generally, being able to identify the operation of the theory-formation system in this way, with reasonable reliability if not perfect accuracy, could be helpful as a first step in working out the functional, and eventually, computational structure of the system. Imagine if we tried to do vision science by calculating every possible relation between visually accessible information and behavior. By using phenomenology as a mediating form of evidence we can narrow the psychological problems to something more tractable, though still very hard. The same might be true for the psychological problem of characterizing theory-formation. For example, the drive-like phenomenology of explanation may lead us to think of theory-formation as a more active exploratory process than we might have done otherwise, exploring the temporal unfolding of the transition from “hmm” to “aha” may give us clues about how the theory-formation system works on-line, and so on But obviously these are still just suggestions for a future research program.

Finally, understanding the nature of the psychological process of theory-formation and explanation may contribute to the more traditional normative questions of philosophy of science. We may both learn how evolution constructed the best causal inference device we know about, and also see how the limitations of that device can be corrected and supplemented. In this way, explanation might actually explain things.

References

Baillargeon, R. (1993). The object concept revisited: New directions in the investigation of infants' physical knowledge. In C. Granrud (Ed.), Visual perception and cognition in infancyCarnegie Mellon symposia on cognition. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Bartsch, K., & Wellman, H. M. (1995). Children talk about the mind. New York, NY: Oxford University Press.

Bowlby, J. (1969). Attachment and loss. New York: Basic Books.

Bromberger, S. (1965). An approach to explanation. In R. J. Butler (Ed.), Analytic philosophy (pp. 72-105). Oxford: Blackwell.

Campbell, J. (1994). Past, space and self. Cambridge, Mass.: MIT Press.

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

Cartwright, N. (1989). Nature's capacities and their measurement. Oxford. New York: Clarendon Press. Oxford University Press.

Cheng, P.

Ekman, P. (1992). An argument for basic emotions. Cognition and Emotion, 6(3/4), 169-200.

Gelman, S. A., & Wellman, H. M. (1991). Insides and essence: Early understandings of the non-obvious. Cognition, 38(3), 213-244.

Glymour, C.

Gopnik, A. (1988). Conceptual and semantic development as theory change. Mind and Language, 3, 163-179.

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

Gopnik, A., & Meltzoff, A. N. (1984). Semantic and cognitive development in 15- to 21-month-old children. Journal of Child Language, 11(3), 495-513.

Gopnik, A., & Sobel, D. Reexamining the role of causality in children's early categorization of objects. Society for Research in Child Development .

Gopnik, A., & Wellman, H. M. (1994). The theory theory. In L. Hirschfield, & S. Gelman

(Eds), Mapping the mind: Domain specificity in cognition and culture (pp. 257-293. xiv, 516). New York: Cambridge University Press.

Hempel, C. G. (1965). Aspects of scientific explanation, and other essays in the philosophy of science. New York: Free Press.

Keil, F. C. (1987). Conceptual development and category structure. In U. Neisser (Ed), Concepts and conceptual development: Ecological and intellectual factors in categorization. Emory symposia in cognition, 1 (pp. 175-200. x, 317). New York: Cambridge University Press.

Kruglanski, A. (1989). Lay epistemics and human knowwledge: Cognitive and motivational biases. New York: Plenum Press.

Leslie, A. M. (1987). Pretense and representation: The origins of "theory of mind.". Psychological Review, 94(4), 412-426.

Leslie, A. M., & Keeble, S. (1987). Do six-month-old infants perceive causality? Cognition, 25(3), 265-288.

Meltzoff, A. N. (1988). Infant imitation and memory: Nine-month-olds in immediate and deferred tests. Child Development, 59(1), 217-225.

O'Keefe, J., & Nadel, L. (1978). The hippocampus as a cognitive map. New York: Oxford University Press.

Oakes, L. M., & Cohen, L. B. (1994). Infant causal perception. In C. Rovee-Collier, & L. P. Lipsitt (Eds), Advances in infancy research, Vol. 9 . Norwood, NJ: Ablex.

Piaget, J. (1952). The origins of intelligence in children. New York: International Universities Press.

Piaget, J. (1962). Play, dreams, and imitation in childhood (Norton library . New York: Norton.

Povinelli, D. (in press). Folk physics for apes? New York: Oxford University Press

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

Schult, C., & Wellman, H. (in press). Explaining human movements and actions: Children's understanding of the limits of psychological explanation. Cognition.

Schwitzgebel, E. (1997). Children's theories and the drive to explain. Science and Education.

Slaughter, V., & Gopnik, A. (1996). Conceptual coherence in the child's theory of mind.

Child Development, 67(6), 2967-2989.

Spelke, E. S., Breinlinger, K., Macomber, J., & Jacobson, K. (1992). Origins of knowledge. Psychological Review, 99(4), 605-632.

Tomasello, M., & Call, J. (1997). Primate cognition. New York: Oxford University Press.

Van Frassen, B. (1980). The scientific image. Oxford: Oxford University Press.

Wellman, H. (1990). The child's theory of mind . Cambridge, Mass.: MIT Press.

Wellman, H., Hickling, A., & Schult, C. (In Press). Young children's explanations: Psychological, physical and biological reasoning . In H. Wellman, & K. Inagaki (Eds.), Children's theories . San Francisco: Joosey-Bass.