Reconstructing the Evolution of Mind

Daniel J. Povinelli

Since Darwin, the idea of intellectual continuity has gripped comparative psychology. Psychological evolution has been viewed as the accumulation of gradual changes over time, resulting in an unbroken chain of mental capacities throughout the diversity of life. Some researchers have even maintained that no fundamental psychological differences exist among species. An alternative model argues that a rather profound new psychology related to mental state attribution may have evolved recently in the primate order. The author explores recent experimental research from chimpanzees, rhesus monkeys, and children that is consistent with this second model of psychological evolution. Drawing on the fields of developmental, comparative, and social psychology, as well as evolutionary and developmental biology, the author outlines a research agenda aimed at reconstructing the evolution of metacognition.

Evolution has no compunctions about embarrassing those bent on trusting their intuitions. Intuition relies on subjective judgment, biases in perception, and unequal weighting given to equivalent occurrences. Evolution, in contrast, makes no subjective judgments and perceives only differential survival and fertility. Equivalent occurrences are mechanistically given equivalent weightings. Thus, intuitions about what ought to be cannot protect one against that which evolution has produced. As a case in point, consider how the deep-seated intuition that there is continuity among species affects casual observations of the natural world. For instance, a superficial examination of the feathers of birds might lead one to assume that feathers are derived from mammalian hair, or perhaps vice versa. In fact, neither is true. Avian feathers are derived from the scales of the reptilian ancestors of birds; mammalian hair, in contrast, is a novel evolutionary structure bearing little developmental homology to reptilian scales (Elias & Bortner, 1957; Lillie, 1942). Yet intuition, which understandably seeks to find continuity, would lead us to believe otherwise.

Such instances of apparent morphological discontinuity are really no mystery; evolution is historical, replete with the errors that history implies. Extinction, coupled with the imperfections and inherent limitations of the fossil record, virtually assures that there will be many instances in which contemporary diversity will appear discontinuous. A cursory examination of the major transformations that have occurred in body plans (reflected in the classification of distinct phyla) is enough to demonstrate that discontinuous patterns do not necessarily imply discontinuous processes.

Perhaps nowhere in biology is the temptation to trust intuitions about continuity more seductive than in the field of animal behavior and, in particular, with respect to the issue of intellectual continuity. Unlike many other areas of biology, the history of comparative psychology has been marked by an inability to meet a central goal of the evolutionary sciences: reconstructing the timing and order in which particular characteristics appeared during the history of life. One explanation for this inability is the charge that psychologists have adhered to the "phylogenetic scale" instead of treating psychology as part of the radiation of animals outward in all directions (Hodos & Campbell, 1969). Others have defended the opposite position, arguing that this is not a failure of comparative psychology at all but rather that our inability to identify the emergence of interspecific psychological differences reflects a fundamental reality of nature—namely, that such differences do not exist (see McPhail, 1987).

In this article I sketch a preliminary synthesis among the fields of evolutionary biology, developmental biology, developmental psychology and epistemology, and comparative psychology in an effort to further understand the issue of continuity and discontinuity in psychological evolution. This attempt can be viewed as complementary to Parker and Gibson's (1979) previous effort at reconstructing psychological ontology. However, in contrast to their efforts at using comparative ontology of Piagetian stages of development as a means of providing an adaptive scenario for human sociocultural evolution, I examine

Lewis P. Lipsitt served as action editor for this article.

This article has benefited from many conversations with Gordon G. Gallup, Jr., Sandra deBles, Lucian Hughes, Laurie Godfrey, and John Cant. I also thank Jon Marks and an anonymous reviewer for providing a number of helpful comments. The preparation of this article was supported by National Institutes of Health Grant No. RR-03583-03 to the University of Southwestern Louisiana New Iberia Research Center.

Correspondence concerning this article should be addressed to Daniel J. Povinelli, Laboratory of Comparative Behavioral Biology, University of Southwestern Louisiana, New Iberia Research Center, 100 Avenue D, New Iberia, LA 70560.

1 Evolutionary biologists distinguish between homology—structures or capacities that are similar between species because of common descent from an ancestor that possessed that character (e.g., the finger bones of reptiles and mammals)—and convergence—similarity that results not from common descent but from evolutionary "re-inventions" of the same feature as the result of the same functional demands (e.g., the wings of birds and bats). Finer distinctions are possible and important (e.g., the distinction between convergence and parallelism), but I do not address them here.
the evolution of the capacity for self-conception and mental state attribution in order to understand the issue of psychological discontinuity in both ontogenetic and evolutionary history. In addition, I hope to demonstrate how this approach can break the description-to-argument cycle that has historically paralyzed comparative psychology's effort to assess whether nonhumans have an awareness of the mental world by replacing it with a more meaningful cycle of prediction-data-collection. And although the ultimate resolution of the issue is far less important than achieving the needed conceptual and methodological improvements, it may turn out that a profoundly new psychology may have evolved quite recently in the history of the primate order. In its wake it may have left two fundamentally distinct groups of life: those who know that the mental world exists and those who do not.

Darwin's 'Mistake'

A unifying goal that links many diverse fields of biology is what evolutionary biologists call phylogenetic reconstruction. The challenge is simple, although implementing it is often difficult. In looking at the past and present diversity of life on earth, evolutionary reconstructions ask, can we inferentially deduce the sequence in which species, and their specific traits, evolved? Issues of continuity plague researchers who attempt to reconstruct phylogenies of both living and extinct organisms. When a novel character is identified, is it a primitive trait that has been lost in other lineages? Or instead, is it a recently derived trait that is present in only a handful of species? Much progress has been made in recent decades in systematizing approaches to phylogenetic reconstruction, but intuition still plays a major role in making judgments about whether particular morphological characters are primitive or derived.

Psychological evolution is a special case in point, but it suffers from the slightly different problem that intuition has led many researchers into seeing continuity connecting most vertebrate forms of life (Darwin, 1871/1982a; Griffin, 1976; Romanes, 1883). Part of the reason for this was recognized by Romanes (1882), who realized that unlike morphological structures, psychological structures are visible only through their "behavioral ambassadors." Thus, identifying the sequence of the evolution of novel innovations in psychological traits has been all but impossible because many investigators, looking at the behavioral evidence to date, do not see any instances of stepwise evolutionary innovations (cf. Bitterman, 1975). This state of affairs can be traced at least to Charles Darwin himself, who believed that anything less than complete intellectual continuity could be used as evidence against his theory of evolution (Darwin, 1871/1982a).

Darwin's legacy is a direct descendant of his publication of The Origin of Species (1859/1982b). In it, naturalists found a logically compelling argument for the "how" of evolution. Darwin proposed that the fodder for the astounding goodness of fit between organisms and the specific habitat in which they lived could be found in the variation that existed between individuals. The bottom line, as he saw it, was differential survival and fertility. Individuals that possessed slight variations that helped them survive and reproduce in a particular environment would outreproduce those that did not. Over time, these new traits, if significant enough, would become fixed as a new part of the species. Thus, through a process of natural selection, the bodily structures of animals were fine-tuned to fit their environments. Life evolved not upward along the rungs of a ladder but outward in all directions like the branches of a tree.

Darwin (1859/1982b) was convinced that his ideas about natural selection of morphology could eventually throw important light on the gradual accumulation of mental power as well and thus inform us about "the origin of man and his history" (p. 373). For years he had collected notes and refined his ideas on the evolution of humans, but he had intentionally refrained from publishing them because, as he put it: "I was afraid that I should thus only add to the prejudices against my views" (Darwin, 1871/1982a, p. 389). However, by the late 1860s Darwin sensed that many of the younger, aspiring biologists were placing increasing faith in his ideas, and it thus seemed like an opportune time to publish a treatise that applied his theory of natural selection to a specific species, Homo sapiens, in greater detail. But by beginning with humans, Darwin came face to face with a set of difficult issues that were at the heart of the Victorian rejection of the very idea of evolution.

The remarkable physical similarities between humans and animals convinced Darwin that they shared a common ancestry in the distant past. He realized that others would not be persuaded by similarities in gross morphology alone. After all, according to traditional wisdom, what truly separated humans from animals was not their bodies but their minds. On the one hand, Darwin knew that he must deal with the apparent discontinuity implied by human mental prowess, but, on the other hand, he felt that it was a requirement of his theory that he be able to show varying degrees of those same mental abilities in animals. Darwin (1871/1982a) outlined the problem in the opening paragraph of the third chapter of The Descent of Man:

We have seen in the last two chapters than man bears in his bodily structures clear traces of his descent from some lower form; but it may be urged that, as man differs so greatly in his mental power from all other animals, there must be some error in this conclusion. . . . If no organic being excepting man had possessed any mental power, or if his powers had been of a wholly different nature from those of the lower animals, then we should never have been able to convince ourselves that our high faculties had been gradually developed. (p. 445)

The paradox was real for Darwin. If natural selection resulted in the origin of new species through the gradual accumulation of small adaptations, then the absence of those gradual differences between the most closely related species was evidence against his theory of natural selection. He knew that if he suddenly postulated that mental evolution was governed by a different set of laws from
that of physical evolution, this would raise a cloud of suspicion around the entire theory. Indeed, Darwin already stood aghast at Alfred Russell Wallace (co-discoverer of the principle of natural selection), who balked at the human threshold and invoked divine intervention as the origin of human intellect (Gould, 1980). But fortunately, even in the largely uncharted realm of mental experience and intelligence, Darwin could already find evidence for an evolutionary continuity:

It can be shown [sic] that there is no fundamental difference of this kind. We must also admit that there is a much wider interval in mental power between one of the lowest fishes, as a lamprey or lancelet, and one of the higher apes, than between an ape and man; yet this interval is filled up by numberless gradations. (Darwin, 1871/1982a, p. 445)

Darwin pressed for psychological continuity between humans and other organisms for a number of different reasons but none more so than a strictly pragmatic one: Contemporary intuition (his own included) dictated that where there had been evolution, there must also be continuity. Indeed, the style of presentation of his argument for natural selection in the Origin makes this clear. And at one important level, Darwin was correct. Natural selection acted on variants that, although different from each other along some characteristic, were nonetheless members of the same species. From a distance, then, continuity in evolution was obvious. But at another level, Darwin was mistaken. In any particular case, the failure to find continuity among living species was not evidence against his theory at all. The process of extinction alone was enough to selectively filter out both morphological and psychological continuity so that living representatives of ancient diversity might look as if some of their particular traits had arisen through spontaneous generation. Darwin, of course, recognized this and appealed to the fossil record as a means of filling up the "numberless gradations." But when it came to mental powers, Darwin had no such recourse. Psychology does not fossilize. The unfortunate truth is that when continuity in mental functioning is no longer apparent in living species, it cannot be resurrected by paleontologists.

When it came to mental powers, then, Darwin's mistake was fundamental. The mere fact that it was he, not a member of some other species, who sat to write the Origin might have been enough to signal that dramatic psychological changes may have occurred quite recently in the evolution of the primates. That key traces of this diversity may well have been lost to extinction, both before and after this process began, should not have alarmed him at all. That underlying continuity might be obscured during periods of rapid evolutionary change is an idea that might have otherwise occurred to a man who was not so besieged by the prevailing intuition of his day (see Preuss, in press, for a discussion of this issue from the perspective of the neurosciences).

**Discontinuity in Psychological Evolution and Ontogeny**

Many comparative psychologists view the issue of continuity in psychological evolution as an outdated term from an archaic, sterile academic debate. Like the "nature–nurture" debate before it, they see the "continuity–discontinuity" debate as simplistic and serving little useful function. There is much to recommend this wariness. For example, the ape-language experiments seem to have nearly drowned in the continuity–discontinuity issue, long before their promise was fully realized. In this sense, casting aside troubling terminology can often be extraordinarily helpful. The problem, of course, is that at other times, long after the terminology is abandoned, the substantive issues remain as stark and daunting as ever. Indeed, the abandonment of troublesome terminology is often merely a sign that an academic truce has been reached to ignore fundamental conceptual disagreements. In the context of the ape-language experiments, for instance, it was not really the continuity–discontinuity issue that was the root of the field's turmoil. It was the general failure to define language (and experimental methodology) precisely enough to determine where similarities and differences (continuities and discontinuities) existed between ape and human (Premack, 1976, 1986).

It is important to address the criticism that the phrase psychological continuity may mean different things to different investigators. Some might claim that continuity implies nothing more than the ability to discover the historical antecedents of a particular capacity. For them, continuity implies that there are no qualitative differences in psychology between nearest living relatives and that apparent differences start to emerge only as larger and larger monophyletic groups are examined. I believe that these issues are critically important, as a later portion of this article bears out quite clearly. But in the context of the continuity debate, this argument may merely turn out to be a convenient excuse for our failure to confront a basic conceptual issue: Do species differ with respect

---

2 Darwin was engaged in a debate with Wallace that pivoted on the issue of humanity's apparent intellectual isolation from the rest of the natural world. Wallace (1870), unlike Darwin, held out for the uniqueness of human intellect. Wallace believed that natural selection could not have produced such a marvelously complex device as the human brain because it contained a seemingly infinite store of latent potentials for advancement. This, Wallace contended, violated his assumed axiom of natural selection that all structures were produced to meet immediate needs. Thus, the only alternative mechanism he could envision as having produced humanity's intellect was the divine hand of some "Supreme Intelligence."

3 Eldredge and Gould (1972) proposed a specific theory that explains this kind of discontinuity both in the fossil record and in cases of apparent discontinuity among living organisms. Their theory of "punctuated equilibrium" argues that major transitions (speciation events) are confined to relatively short durations during the history of a given lineage. The vast majority of a lineage's history is marked by periods of stasis in which little change occurs. They contrasted this model with what they saw as the traditional view of "phyletic gradualism," inherited from Darwin, in which speciation events occur gradually throughout the lifetime of a given lineage.

4 Studies of fossil endocasts from brain cases represent an obvious objection to this claim. However, although much information can be gleaned from them, their resolution at present is relatively poor, and it is possible that an understanding of the evolution of specific psychological traits may never be rescued from them.
to the presence or absence of basic psychological traits? Discovering ancestral antecedents of particular structures or capacities does not eliminate the crucial issue of final function. The homologies between the skeletal structures of a bat's wings and the primitive Bauplan of the first four-legged amphibian that crawled out of the Devonian sea do not in any way reduce the marvelous difference that separates the two: Bats can fly; most land-dwelling quadrupeds cannot. On this issue I side with Goldman-Rakic and Preuss (1987), who pointed out that when it comes to similarity and difference, "Evolution does not accord priority to one over the other, and comparative psychology should not either" (p. 667). (For a similar conclusion, see also discussion of psychological discontinuity in Premack, 1986.)

The same interplay between change and function is true in the context of psychological ontogeny as well. To many, the field of developmental psychology must appear to be a tired, century-long debate between those who see complete continuity in psychological development from birth to adulthood and those who advocate stage-like theories of development (see Brainerd, 1978). However, such a naive view misses both the merits and limitations of both positions. Just as immature organisms are developing the morphological structures necessary to perform later functions of which they are initially incapable, so too can psychological ontogeny be thought of as the construction of capacities (structures) that will later provide the organism with tools for behavioral performances impossible at earlier points in development. Thus, when development is viewed across stretches of time, it should not be surprising that qualitative shifts in performance capacities can be demonstrated. By the same token, it is also perhaps not surprising that more fragile performances can be elicited at earlier time points (for historical examples, see Gelman, 1979). The issue becomes one of local, domain-specific investigations that seek to determine at what point early demonstrations of competence are implicated in some tangible behavioral patterns, at what point they are tapping into unused (but developing) capacities, and at what point these early demonstrations are simply not really measuring the capacities they are believed to be measuring.3

**Gallup's Challenge**

Just over a decade ago, Gordon Gallup (1982) returned to the issue of psychological discontinuity following his previous discovery that chimpanzees are capable of recognizing themselves in mirrors (Gallup, 1970). Unlike a number of species of monkeys that he originally tested, chimpanzees spontaneously learned to use mirrors to explore previously unknown parts of their bodies, such as the head and anal-genital region. Orangutans soon made their debut into the circle of species capable of self-recognition (Lethmate & Ducker, 1973). But in the decade following this discovery, a number of researchers attempted, unsuccessfully, to demonstrate this capacity in other species of nonhuman primates. At present count, nearly 20 of these species, including gorillas, have been more or less systematically examined for signs of self-recognition and have shown little evidence for the capacity (recent reviews have been provided by Gallup, 1991; Povinelli, 1991).

Gallup (1970, 1975, 1977) repeatedly proposed that self-recognition can be interpreted as an index of self-awareness or the ability of an organism to become the object of its own attention. In other words, he proposed that the behavior of self-recognition in a mirror is possible only in organisms that possess some (unspecified) self-concept. Gallup also speculated that the repeated and robust failures to elicit self-recognition in other primates might best be interpreted as a cognitive deficit in these species—in other words, they are incapable of reflecting on their own existence. This position has drawn repeated criticisms, especially from those committed to seeing intellectual continuity among the primates or even among all mammals (see Desmond, 1979; Eglash & Snowden, 1983; Fox, 1982). Others, myself included, have objected to Gallup's interpretation of the negative evidence concerning the failures to find self-recognition in gorillas, lesser apes, and monkeys on strictly logical grounds, arguing that there does not appear to be a necessary reason that self-aware organisms must recognize themselves in mirrors (Povinelli, 1987). Others have offered explanations for self-recognition that require less of a self-concept than Gallup implied (Goustard, 1983; Jaynes, 1978; Mitchell, in press). But from the standpoint of intuition, perhaps the most troubling issue concerns the repeated failures to find clear and unambiguous evidence of self-recognition in most gorillas (Ledbetter & Basen, 1982; Suarez & Gallup, 1981). This is especially true in light of the rather compelling evidence for self-recognition in at least one home-reared gorilla that had been instructed in a version of American Sign Language (Patterson, 1984; for a discussion, see Povinelli, 1987). I shall return to this issue in greater detail later.

All along, Gallup acknowledged the potential deductive weakness of certain aspects of his arguments (see Gallup, 1977, as well as Gallup, 1979). Nonetheless, as the diversity of primate species unsuccessfully tested for self-recognition continued to widen and as the amount of time and techniques used on them increased, Gallup found it difficult to resist the conclusion that a fundamental cognitive difference related to self-conception separated chimpanzees, orangutans, and humans from...
most other forms of life. Reflecting on this state of affairs, Gallup (1982) proposed a model in which he attempted to provide a more meaningful and predictive framework for the self-recognition debate. In other words, he set about outlining a theory he believed could provide an independent test of his hypothesis that species differences existed with regard to self-awareness. In brief, Gallup argued that if self-recognition is an index of self-awareness, then chimpanzees and orangutans (and possibly some gorillas) ought to be capable of introspection. This introspective capacity enables organisms to reflect on some unspecified set of their own mental experiences. Gallup reasoned that such knowledge of self provides a means by which organisms can begin to inferentially reason about similar mental experiences in other organisms.6

Thus, Gallup has repeatedly predicted that chimpanzees and orangutans ought to regularly engage in what social psychologists refer to as self- and social attribution—the process of attributing mental states to oneself and to others. In essence, Gallup's argument boils down to proposing that self-recognition is an empirical marker of what Premack and Woodruff (1978) referred to as a "theory of mind," the ability of organisms to make inferences about the mental states of others. In an ingenious and landmark set of experiments, Premack and Woodruff offered evidence that an adult female chimpanzee was able to attribute intention to actors she observed struggling to solve staged problems. In contrast, Gallup believes that lesser apes and monkeys—indeed, perhaps the remaining diversity of life itself—ought to be incapable of reasoning about mental states such as intention, knowledge, and belief in either themselves or others. Gallup has predicted that unlike chimpanzees, which behave in ways that suggest they reason about unobservable mental states and invoke them as explanations of behavior (e.g., Premack & Woodruff, 1978), organisms incapable of self-recognition do not. The specific details of Gallup's model have been described and discussed elsewhere, and I refer the unfamiliar reader to these sources for an elaboration of the model (Gallup, 1982, 1983, 1985; Povinelli, 1991).

Gallup and Suarez (1986) later extended this basic framework in the context of the development of self-recognition in young children, arguing that children who are not yet old enough to display self-recognition ought to likewise fail to attribute mental states to others and be incapable of a variety of introspectively based social strategies. In contrast, children who are capable of self-recognition should begin to show the first signs of introspectively based social behavior. A number of years later, Lewis, Sullivan, Stanger, and Weiss (1989) adopted a remarkably similar model in the context of their ongoing research into the development of self-recognition in children. They reported, for instance, that the self-conscious emotions in young children emerge hand-in-hand with the onset of self-recognition.

Gallup's challenge, therefore, has been to propose that self-recognition is an index of a much broader and conceptually profound change in psychology that evolved relatively recently in a small branch of the primate order. His model is silent on the issue of what the cognitive precursors of self-conception were, and developmentally speaking, are.7 In the context of both the human and nonhuman self-recognition literature, there have been attempts to identify the precursors of the behavior of self-recognition from both a developmental and evolutionary perspective (Bertenthal & Fischer, 1978; Dixon, 1957; Eglash & Snowdon, 1983; Mitchell, in press). But regardless of whether the precursors of self-recognition can be specified in a manner that produces a set of coherent and testable propositions, Gallup's central prediction has remained unaddressed: Those organisms capable of self-recognition ought to engage in self- and social attribution; those incapable of self-recognition should not. Translated into the terms of evolutionary biology, Gallup has argued that the common ancestor of the gibbons and the great ape–human clade had not yet evolved the ontogenetic system that specifies the construction of the self- attributional system.

Until recently, the procedures for testing Gallup's model may have seemed obscure; but with the explosion of research in developmental child psychology following Premack and Woodruff's (1978) theory-of-mind article, the door has been opened for such investigations (for recent reviews of theory-of-mind research in children see Perner, 1991; Wellman, 1990). Thus, in an ironic way the field of comparative psychology, which spawned a renewed interest in the development of the child's awareness of the mental world, has come full circle. The task now facing comparative psychologists is to adapt verbal paradigms from theory-of-mind research in young children to nonverbal procedures applicable to nonhuman primates.8

---

6 Gallup's (1982, 1983, 1985) argument that knowledge of self provides a mechanism for knowledge of others followed a long-standing epistemological tradition (Adams, 1928; Schutz, 1962; Smith 1759/1961; Stewart, 1956). Although Gallup's argument that humans use the self as a source of inferential knowledge of others was not new, what was strikingly original was his application of this idea in the context of comparative psychology. The extent to which this argument has merit remains unsettled (see Gopnik, in press).

7 George Butterworth (in press) recently clearly articulated the distinction between self-perception and self-conception in the development of the child. His conclusion was that despite a growing understanding of each of these separate aspects of the development of the self, it is still impossible to specify the ontogenetic links between the two in any convincing, causal manner. Thus, it remains unclear how the child's capacity for self-perception (which is present in the neonate and presumably in the fetus as well) gives rise to the profound developmental change at 18 to 24 months of age that many psychologists see as the onset of self-conception or self-awareness (see, e.g., Kagan, 1981).

8 This should not be viewed as a one-way street. Nonverbal tests that are developed for use with chimpanzees, for instance, may be able to provide a vital source of information for developmental psychologists, who traditionally are concerned about relying exclusively on a child's verbal responses (see Flavell, 1977). Indeed, in our own laboratory we are currently investigating a potential nonverbal measure of self-knowledge assessment that can be used across human and nonhuman species (Povinelli, Perilloux, & Bierschswale, 1993). With regard to young children, it has the distinct advantage of not requiring the child to answer, or react to, verbal inquiries.
It is important to clarify a central misconception that has plagued Gallup’s efforts to introduce self-recognition as a phylogenetic (and developmental) marker of the onset of self-conception. On both intuitive and theoretical grounds, some comparative psychologists have objected to one or more of Gallup’s assumptions. However, in doing so they have, for the most part, failed to understand that the model is simply a way of generating strong (and hence quite falsifiable) predictions about the relationship between self-recognition and underlying psychological processes. Thus, it is no longer a matter of intuitive preference for one theoretical position over another but rather a matter of hypothesis testing. Once the formalism of a predictive framework has been introduced into a scientific debate, argument alone will no longer suffice. Other theoretical models that differ in their interpretation of the meaning of self-recognition are possible; and the ultimate fate of them, along with Gallup’s model, will be determined by the extent to which they produce novel predictions about heretofore unknown data sets (see Mitchell, in press, and reply by Gallup & Povinelli, in press).

In a very different context, evolutionary biologist Leo Buss (1987) remarked, “Evolution is historical, and our most difficult task is not to identify evolutionary innovations ... but to identify their sequence and consequences” (p. 183). For nearly a decade, Gallup’s framework has offered a set of rich and risky predictions concerning the consequences of the evolution of the psychology supporting self-recognition—a set of predictions that can now be addressed in a straightforward, empirical fashion.

Do Phylogenetic Differences in Social Psychology Exist?

Several years ago I embarked on a series of comparative experiments that were originally designed to test Gallup’s model of the evolution of self- and social attribution. As a preliminary attempt at falsifying the model, I settled on a test case comparison between chimpanzees and rhesus macaques, two species notorious for their ability and inability, respectively, to recognize themselves in mirrors (see Gallup, 1985). The idea behind these investigations was to develop a series of nonverbal experimental procedures that could tap into an organism’s ability to attribute mental states to others and then to conduct a closely parallel version of each task with the two species. Gallup’s model predicted that chimpanzees ought to do well on some (unspecified) subset of such tasks, whereas rhesus macaques ought to uniformly fail them. In a very real sense, the approach I adopted was to pit two explanatory frameworks against each other: learning theory (such as that espoused by Hull, Spence, and Skinner) versus attribution theory (Mead, Heider).

Role Reversal as a Measure of Cognitive Empathy

The first set of experiments was designed to tap into a form of social attribution known as role taking or cognitive empathy, which develops in children at approximately 18 to 24 months of age (Dunn & Dale, 1984; Flavell, Botkin, Fry, Wright, & Jarvis, 1975; Watson & Fischer, 1980). This ability presumably underpins young children’s ability to engage in a variety of forms of pretend play in which they switch from one social role to another (Mead, 1934), as well as speech acts in which the child essentially adopts “motherese” (Sachs & Devin, 1976). Such acts seem to involve the attribution of mental states to oneself (and others, at later stages) in that they allow individuals to put themselves into the psychological perspectives of another person. Watson and Fischer have interpreted this behavior as the child’s understanding (or attribution of) personal agency.

The experimental attempt to operationalize this form of role taking was derived from a study conducted in the early 1960s by Mason and Hollis (1962), who studied the development of social communication in young rhesus monkeys. They constructed an apparatus that allowed pairs of infant monkeys to cooperate to obtain food rewards. During testing, the pair of animals sat across from each other with an apparatus between them (see Figure 1A). For each pair of animals, one had access to a series of handles, which, if pulled, extended two identical food trays to within reach of each subject. During the course of their experiment, Mason and Hollis modified the apparatus so that each food tray contained a shield that prevented the monkey with access to the handles (the operator) from seeing into it. In contrast, the other monkey (the informant) could see into the food trays but had no way of manipulating the apparatus. Eventually, the subjects learned to “cooperate” with each other, with the informant giving some signal (either sitting in front of the baited pair of food trays or reaching for it) and the operator using this cue to pull the correct handle, thus rewarding both animals. The researchers conducted a series of transfer tests, only one of which is relevant here. In this test, after each pair of subjects had learned their respective roles as operator or informant, the monkeys were transferred into the opposite cages, and Mason and Hollis looked for evidence of transfer. That is, would the monkeys perform their new roles with the same accuracy as their old roles? When the subjects’ roles were reversed, their performances uniformly dropped to chance levels, and they only gradually relearned the new roles.

Mason and Hollis (1962) were not interested in issues of self- or social attribution, role taking, cognitive empathy, or theory of mind. However, it occurred to me that the task itself could be used as a means of inferring to

---

9 There are those researchers who do not believe that true mental state attribution underpins these behaviors. Perner (1991), for instance, believed that children do not conceive of the mind in a truly representational manner until they are approximately four years old. However, other frameworks exist that allow young children some understanding of the psychological states of themselves and others (see Miller & Aloise, 1989, for a review). Indeed, even researchers (e.g., Perner, 1991) who hold out for a strong conceptual change between three and four years of age believe that younger children do have some limited understanding of the mental lives of themselves and others.
what extent an organism is capable of engaging in social role taking. Organisms that see the task solely from a behaviorist standpoint ought to perform quite poorly on role reversal, despite having learned to carry out their own original role with near-perfect accuracy. On the other hand, organisms that not only utilize cues from their partner successfully but also comprehend the meaning of such cues within the context of the task, ought to understand what to do when the roles are reversed. In other words, they should immediately understand the need to perform the opposite role in order to successfully complete the reciprocal exchange. The monkeys that Mason and Hollis tested were socially isolated infants, and thus their results could not legitimately be reinterpreted in hindsight. But I interpreted Gallup’s model as clearly predicting that socially competent, appropriately aged chimpanzees should show immediate evidence of role reversal, whereas comparable rhesus macaques should not.

In order to test this prediction, one adult and three subadult chimpanzees and four adult rhesus macaques were used as subjects. All of the subjects were housed in social groups, and all had extensive experience with traditional learning studies. I constructed two modified versions of the apparatus used by Mason and Hollis, which differed from each other only in size and complexity (see Figure 1A for a schematic diagram). Using each apparatus, my colleagues and I set about replicating the portion of the Mason and Hollis (1962) experiment described above, with one major modification. Instead of pairing the subjects with each other, we paired them with human experimenters in order to standardize the experiences that each animal received. For each species, we divided the subjects in half. Half of them were trained as operators (with the human partner as informant) and tested for role reversal comprehension as informants; the other half were trained and tested in the opposite order. We used this design so that in the event that we obtained clear signs of role reversal, we could be somewhat assured that the results were not simply an artifact of one role being inherently easier to perform than the other. The procedures used with the chimpanzees and macaques were nearly identical. The only major difference was that the macaque subjects who were initially trained as operators showed a profound difficulty in learning to respond to the pointing gesture of their human partners. We were therefore forced to intervene and use elaborate shaping procedures to train them to attend to the cue of pointing until they performed at near-perfect levels.

The results from these experiments differed dramatically by species. Three of the four chimpanzees we tested showed clear and immediate evidence of comprehending their new role on reversal, despite the fact that they had encountered some difficulty in learning their original role. None of these three subjects showed a decrease in performance from their last session of 20 trials of training in their old role to the first session of 20 trials in their new role. Indeed, two of them displayed appropriate behaviors from Trial 1 forward, making only one and two mistakes about half way through the session. The fourth subject showed only mild evidence of transfer, and his performance decreased significantly on reversal. In striking contrast to the overall pattern of the chimpanzee experiment, none of the macaques showed evidence of

---

Figure 1
Apparatus Used in Role Reversal and Seeing-Knowing Experiments

![Diagram of apparatus](image)

---

10 This difference is intriguing in its own right because comprehension of pointing has often been interpreted in terms of an understanding of intention (for a discussion of this issue from a comparative perspective, see Povinelli, 1991). However, pointing emerges at such a young age in children, and its topography appears to be so strikingly universal (that is, extending the index finger), that it may well be fairly hard-wired, with little meta-awareness behind the act at early ages. Our research with three- and four-year-old children fits this model because even young three-year-olds appear not to be aware of certain aspects of the communicative function of pointing, despite the fact that they are capable of using it successfully (Povinelli & deBlois, 1992b).
transfer into their new roles. On confronting the role reversal procedure, they monitored us as we slowly rotated the apparatus but then simply sat watching, as if waiting for the old cues to appear. Some of the monkeys even went as far as to search for the handles, which in their case were no longer present. Only after a number of sessions (and only after elaborate shaping procedures in the case of those that had to respond to pointing), did the monkeys learn their new roles. A complete description of the procedures and results of these experiments are provided by Povinelli, Nelson, and Boysen (1992) and Povinelli, Parks, and Novak (1992). Taken collectively, these experiments point to potential species differences in the psychological capacity for role taking.

**The Perception–Knowledge Relationship**

The second set of experiments was devised after I read the now-classic report by Hienz Wimmer and Josef Perner (1983) on the attribution of false belief in young children. I had been struggling to operationalize the distinction between the mental states of guessing and knowing that Premack and Woodruff (1978) had discussed in their landmark theory-of-mind article. Gallup (personal communication, October 30, 1986) had also stressed the potential importance of determining whether chimpanzees could understand the implications of sensory obstruction; for example, could they model the visual experiences of someone who is blindfolded? After I read Wimmer and Perner’s report, it struck me that with only relatively minor modifications in our role reversal paradigm, we could pose the following question (behaviorally) to our subjects: What is the psychological difference between someone who has seen an event occur and someone else who has not? In other words, we could ask the subjects whether they understood how knowledge states arose and in particular whether they understood the psychological connection between seeing and knowing (Wimmer, Hogrefe, & Perner, 1988).1 We have referred to the general phenomenon of whether an individual understands the causal connection between sensory input and knowledge formation as the perception–knowledge relationship (Povinelli & deBlois, 1992b).

The final procedure we selected capitalized on the fact that during the role reversal experiment all of the subjects (either during training or after role reversal) had learned to use human pointing as an indicator of which among several food trays contained a slice of banana. For our test of whether chimpanzees understood the causal psychological connection between seeing and knowing, we slightly modified the apparatus by placing opaque upside-down cups over the food trays, hiding their contents (see Figure 1B). The subjects were then allowed to watch two experimenters act out different roles, which we conveniently referred to as the *guesser* and the *knower.*12 At the start of each trial, both experimenters stood across from the subjects on the informant side of the apparatus. Next, following a randomized schedule, the guesser signaled the subject that he was going to leave the room and then did so, shutting a door behind him.

After the door was closed, the knower showed the subject a piece of food and then proceeded to hide it under one of the cups. The subjects could see that the knower was hiding the food, but a cardboard screen prevented them from seeing which cup was baited. The knower also used a number of false baiting procedures on each trial to prevent the chimpanzees from using other cues to locate the reward. After the food was hidden, the knower again stood behind the apparatus and waited until the guesser returned from outside the room. Next, both the guesser and the knower simultaneously pointed to different cups. The knower, of course, pointed to the correct cup. The guesser, in contrast, pointed to a previously determined incorrect cup. The cardboard screen was removed by a neutral third experimenter, and the subject was allowed to pull one of the handles in order to look under one of the cups. Each subject was given 10 trials per day.

The results of these experiments again produced a clear contrast between the chimpanzees and the rhesus monkeys. To begin, two of the four chimpanzees showed a clear and stable (although by no means perfect) discrimination between the guesser and knower within the first five days of testing. A third chimpanzee showed the relevant discrimination later in testing. The final chimpanzee, although she showed an overall significant preference for the knower across the six initial weeks of testing, never displayed a consistent preference. Although intriguing, this result could potentially be explained in terms of associative learning. It was quite possible that the chimpanzees were not reasoning about mental states at all; rather they may have been using behavioral strategies such as “pick the person who stays in the room” or “pick the person who hides the food” in order to solve the task.

In order to pit this learning explanation against the attributional one, a simple transfer test was conducted in which the key environmental variables that the chimpanzees might have been using were altered. The guesser now never left the room but instead stood passively beside the knower behind the apparatus. The knower, mean-

---

1 Although we were not aware of it at the time, two related papers were in press that had direct bearing on our own research. Wimmer, Hogrefe, and Perner (1988) had conducted the analogous research with three- and four-year-old children, and Premack (1988) provided a preliminary report on a similar experiment conducted with four juvenile chimpanzees (see below). Reading the results of these experiments later, along with Flavell’s research on Level 1 and Level 2 perspective taking (Flavell, Everett, Croft, & Flavell, 1981) greatly assisted in clarifying the interpretation of our own research.

12 Although I have chosen to retain the labels *guesser* and *knower,* I fully recognize that this experiment taps less into an organism’s ability to discriminate between the mental states of guessing and knowing than it does its ability to understand the connection between seeing and the resulting mental state of knowing. As we pointed out in our original report: “A question left unresolved by our study is to what extent the chimpanzees were capable of understanding the concept of guessing. Although they clearly responded preferentially to the knower, their comprehension of what the guesser was doing remains uncertain” (Povinelli, Nelson, & Boysen, 1990, p. 209). In this sense, it is unfortunate that we chose to title the report “Inferences About Guessing and Knowing by Chimpanzees (Pan troglodytes).”

May 1993 • American Psychologist
while, never hid the food but simply watched as a third experimenter appeared and baited one of the cups. While this baiting procedure was occurring, the guesser placed a paper bag over his head to prevent him from seeing where the food was hidden. Finally, as before, both of the experimenters pointed, and the chimpanzees selected between them. Each chimpanzee was given 30 trials using this procedure. The results of this transfer test revealed that the three subjects that had shown a preference for the knower during the original procedure maintained that preference. However, even this result does not rule out extremely rapid learning on the part of the chimpanzees. In addition, Premack (1988) provided a preliminary description of an experiment similar to our own in which two of his four chimpanzees made the relevant discrimination within 24 trials. Although he used procedures somewhat different from ours, his findings essentially mirror those that we reported. On the basis of these results, my colleagues and I adopted the working hypothesis that chimpanzees do understand the perception–knowledge relationship, at least in the visual modality. A complete description of the experiment and its results has been provided by Povinelli, Nelson, and Boysen (1990).

The rhesus macaques, it will be recalled, were also trained during the course of the role reversal experiment to respond to human pointing with near-perfect accuracy. Thus, by all appearances, they began the seeing–knowing experiment with the same technical expertise as their chimpanzee counterparts. This similarity, however, proved to be only superficial. In sharp contrast to the chimpanzee results, we found no evidence that our four adult rhesus macaques understood the relationship in question. Despite over 12 weeks (and in one case, 16 weeks) of extended training, none of the macaques ever showed a preference for the individual who had hidden the food or who watched as a third experimenter hid it. Indeed, we used a number of procedures that were explicitly designed to assist the macaques in making the relevant discrimination, without success. We also tested a number of potential procedural explanations for the macaques’ poor performance. For example, it was possible that the macaques had an attentional problem in the context of the task or that they understood the relationship between seeing and knowing but simply could not properly associate the pointing hands with the appropriate experimenters. It was also possible that the monkeys had some specific learning difficulty within the context of the task. However, further experimentation, and a careful analysis of the data, cast doubt upon each of these potential explanations. A complete description of this experiment and its results has been reported by Povinelli, Parks, and Novak (1991).

Admittedly, the evidence was negative. However, our elaborately unsuccessful attempts to prod the macaques into performing like the chimpanzees led us to adopt the working hypothesis that rhesus monkeys, unlike chimpanzees, do not understand the seeing–knowing relationship. To those wary (for good reason) of attempts to prove the null hypothesis, it is important to point out that when adopted as a working hypothesis, claims that species differences exist in a specific psychological domain are as falsifiable (and hence as useful) as claims that they do not. Immediately after we had conducted the experiments with the macaques, we were encouraged as to the generality of our findings in the knowledge attribution experiment by the publication of a paper by Dorothy Cheney and Robert Seyfarth (1990a), who reported on a series of experiments suggesting that macaque mothers may not attend to knowledge states of their infants when warning them about the presence of food or danger.

Young Children: Calibrating Nonverbal Measures of Mind

Although the results of the experiments described above provided clear contrasts between chimpanzees and rhesus monkeys on tasks that presumably were measuring mental state attribution, an obvious alternative explanation for our results (especially those with the chimpanzees) had not yet been explored. Perhaps our intuitions about what these tasks were measuring were simply wrong, no matter how sound those intuitions seemed to us. In other words, a skeptic appealing to learning theory could reasonably contend that the task in question was not really measuring the presence of the psychological capacity that we believed it was. What was needed was an independent method for assessing the validity of our tasks. After pondering this concern, we adopted a version of a technique long advocated by David Premack: using with young children tasks originally designed for apes (Premack, 1975; Premack & Woodruff, 1978; Premack & Dasser, 1991; see also Köhler, 1927). In particular, however, we implemented a novel version of this technique. Using what was then known about young children’s development of a theory of mind, we made specific predictions about how children of different ages ought to perform on our tasks.

For example, by the time we had completed our seeing–knowing experiments with chimpanzees and macaques, there was a reasonable body of evidence suggesting that whereas many four-year-old children understand the perception–knowledge relationship, most young three-year-olds do not (see Perner, 1991 for a recent summary of the evidence on this point). Thus, we reasoned that if our nonverbal task was a valid measure of the seeing–knowing relationship, then many four-year-olds should perform quite well, whereas most young three-year-olds should not. In order to test this prediction, Sandra deBlois and I tested a number of three- and four-year-old children

---

13 We have undertaken several attempts to falsify our working hypotheses concerning the absence of role taking and knowledge attribution in macaques. For example, we used our role reversal procedures with an adult female rhesus monkey that had for many years shown evidence of spontaneous pointing, very similar to instances observed in captive chimpanzees. We thus hypothesized that given her “natural” pointing capacities she might, like chimpanzees, comprehend the role reversal task immediately, thus falsifying our species differences hypothesis. However, she did not (Hess, Novak, & Povinelli, in press).
using procedures closely modeled after the chimpanzee and macaque work described above (Povinelli & deBlois, 1992b). Indeed, we even went as far as to construct an exact replica of the monkey apparatus to use with the children.

Each child was tested for a single session consisting of 10 trials each.14 The child was not given any verbal instructions other than the minimal amount necessary to convey the idea that he or she was about to play a game with the two experimenters, involving the “hider” (the knower) and the “leaver” (the guesser), and that the object of the game was to find a surprise that would be hidden under one of the cups. After we had determined that the subjects would respond accurately to the pointing of each experimenter, the formal testing began. At the start of every trial, each experimenter announced the role that he or she would play. The leaver left the room and waited outside (just as the guesser had done with the chimpanzees and macaques), while the hider hid the surprise under one of the cups (just as the knower had done).

The results of this experiment provided clear support for our hypothesis. Using a number of independent measures, most of the four-year-olds understood the task without difficulty. These successful children easily met (and exceeded) our predetermined criterion of 7 out of 10 choices for the hider, averaging almost 9 correct choices. In addition, they also showed a significant run of correct choices for the hider. Finally, all of these successful children were correct on their first trial. In contrast, only one of the three-year-olds met our 7 out of 10 criterion; this child did not show a significant run of correct choices for the hider, nor was she correct on her first trial. The unsuccessful three- and four-year-olds averaged only 4 out of 10 correct choices, and none of them showed a significant run for the hider. In addition, half of them were correct on Trial 1 and half were incorrect, exactly what one would expect if they were responding at random. Finally, during the middle of one of the trials we asked the children, “Does the Leaver know where the surprise is?” Ten out of the 11 successful four-year-olds answered correctly by saying no or shaking their heads. In contrast, the unsuccessful children appeared to answer this yes or no question at random. These findings allowed us to make a stronger claim concerning the perception–knowledge experiments with macaques and chimpanzees.15 In particular, we argued that success on the task requires an organism to understand how visual perception leads to knowledge.16 A complete description of the methods and results of this experiment is reported in Povinelli and deBlois (1992b). We have not yet conducted an experiment with young children using the role reversal paradigm; thus it would be premature to conclude definitively that the test is measuring role-taking capacities known to be present in two-year-old children. However, recent research by Brownell and Carriger (1990) has confirmed our suspicions by demonstrating that children younger than 18–24 months of age are unable to solve an analogous cooperative task.

With proper prudence, some researchers will object to this method of proceeding, arguing that looking for qualitative developmental transitions in young children as benchmarks to calibrate nonverbal mental state attribution tasks assumes stagelike transitions in the child’s development of a theory of mind that, in fact, are not present. For example, they would point to the success of three-year-olds on certain tasks of understanding false belief but not on others as a clear indication that much of what passes for “transitions” are really artifacts of task demands (e.g., Chandler, Fritz, & Hala, 1989; Hala, Chandler & Fritz, 1991). However, several points need to be kept firmly in mind. To begin, as Premack and Dasser (1991) have pointed out, such interspecific comparisons will always involve measuring task performance. The extent to which the tasks themselves tap into changes in underlying psychological capacities is a matter for empirical prediction. In addition, demonstrating precocial

---

14 Because we were interested in the children’s understanding of the process of knowledge formation in themselves as well as others, we first tested the children for comprehension about the formation of their own knowledge states. The results from two separate experiments closely parallel those reported previously by Gopnik and her colleagues (Gopnik & Graf, 1988; O’Neill & Gopnik, 1991; see also Wimmer, Hogrebe, & Perner, 1988) and suggest that young three-year-olds do not understand how their own knowledge states arise (see Povinelli & deBlois, 1992b).

15 Those who are intimately familiar with theory-of-mind research in young children may object to this claim on the grounds that there currently exists a considerable degree of controversy surrounding whether three-year-olds are truly incapable of understanding the seeing-knowing relationship (i.e., Pillow, 1989; Pratt & Bryant, 1990). I believe that a large portion of the discrepancy results to date may be the result of different investigators looking at fairly different age ranges. This issue, along with potential methodological problems, is discussed elsewhere (Povinelli & deBlois, 1992b).

16 The pattern of the four-year-olds’ success presents some difficulties for our working hypothesis about the chimpanzees’ comprehension of the task. There were obvious differences between the chimpanzees and the four-year-olds in the robustness of their performances. Some have argued that this adds support to the learning theory interpretation of our results (Cheney & Seyfarth, 1992). In fact, this conclusion is erroneous. The chimpanzees were performing at ceiling levels at the conclusion of the experiment (and had been for most of the testing). Thus, their performance level (about 70% correct) was not uniquely explained by learning theory. Learning theory would have to resort to ad hoc explanations such as species differences in short-term memory recall. But the very same explanations could account for why an organism operating with a theory of mind would perform at those levels. Thus, this aspect of the chimpanzees’ performance is silent with respect to the learning versus attribution issue. Regardless of whether chimpanzees understand the seeing-knowing relationship, they clearly differ from normal children in attentional and emotional domains, and these differences may account for the problem under either framework. It is also important to point out that three-year-olds do not perform better when provided repeated trials such as the chimpanzees were given (see Povinelli & deBlois, 1992b). Whiten (1993) has suggested that the learning aspects could have been disentangled if we (in our transfer test) had allowed the knower to wear a paper bag in a way that did not obscure the eyes. In a similar vein, Heyes (in press) has suggested a control in which the knower would momentarily place the paper bag on his or her head after the food is hidden. However, both approaches still leave the subjects with a discriminable cue to learn the distinction instead of attributing it. The proper experimental design requires a large enough sample size of chimpanzees to pit a control (learning) group against an experimental (attribution) group (see Povinelli, 1991, for details of the experimental design).
instances of a given capacity does not necessarily do justice to the phenomenon under investigation. As task demands are simplified, researchers run the risk of no longer measuring what they set out to measure (see Goldman-Rakic & Preuss, 1987). Second, distinct models that stipulate the processes that build the psychology of an organism capable of attributing various mental states should compete with each other for their ability to explain the most empirical observations. In this fashion, conceptually at least, there is no real tension between those who choose to localize the end points of such transitional constructions and those who seek to understand the tools with which the structures are constructed. Thus, although there is a very real conceptual problem within the transitional zones of determining to what extent the organism has access to the capacity in question, researchers can attempt to pinpoint broad transitions that can be critical for interspecific comparisons.

Reconstructing Minds of the Past

The experiments reported above were conducted with the aim of subjecting Gallup's (1982) model of mind to a preliminary test. In this sense, the predictions of the model we tested successfully withstood a rigorous attempt at falsification. Chimpanzees performed as if they attributed mental states to others; rhesus macaques did not. Our findings point to the potential heuristic value of Gallup's framework. However, in addition to having these direct implications for Gallup's model, our experiments raise even broader questions concerning the prospects of studying mental state attribution from a comparative perspective. First, to what extent do the results of these experiments provide support for Gallup's claims that self-recognition is an empirical marker of mental state attribution? In what ways can the model be modified to improve its testability? Second, what methodological improvements are necessary to provide more compelling evidence for the presence of mental state attribution in nonhuman primates? Finally, what can investigations like the ones previously described contribute to reconstructing the evolution of the human mind? By considering these questions, I hope to outline a research agenda unifying developmental and comparative psychology with developmental and evolutionary biology.

Self-Recognition as an Index of Mental State Attribution

In order to fully appreciate the difficulties in interpreting the results of our test of Gallup's (1982) model, it is necessary to keep in mind the fact that the model can be viewed as both an evolutionary argument about the expected distribution of mental state attribution in primates and a developmental argument concerning the ontogeny of this capacity in young children (or young chimpanzees, for that matter). This raises some troubling issues for the model. For instance, the fact that an 18- or 24-month-old child can be fully capable of self-recognition and yet have relatively little understanding of the mental experiences of others suggests that the presence of self-recognition is not a sufficient indicator of complex social attribution. Many important attributional processes may develop only after the onset of self-recognition (e.g., the perception–knowledge relationship, the appearance–reality distinction, an understanding of false belief, Level 2 perspective-taking, second-order attribution; Wimmer & Perner, 1983; Flavell et al., 1981; Flavell, 1986; Perner & Wimmer, 1985; Wimmer et al., 1988). Thus, from a developmental standpoint, to be self-aware (in the sense of being able to recognize yourself in a mirror) is not necessarily the same as being able to attribute intentions and beliefs to others or yourself in the same way as adults or even older children do.

In this context, the failure of rhesus macaques to show evidence of understanding the perception–knowledge relationship becomes only circumstantial evidence that Gallup's (1982) model is valid. It is quite possible that although macaques perform poorly on tasks that require sophisticated attributional abilities, they may fare far better on tests that require attributional abilities more comparable with those possessed by two-year-old children. Thus, a logical next step in attempting to falsify Gallup's model would be to test chimpanzees and macaques for attributional capacities that emerge in young children coincident with the emergence of self-recognition. To some extent our role reversal experiments already speak to this issue, because the task presumably taps into an early developing attributional capacity.

However, simply because self-recognition emerges in children before the onset of many important aspects of self- and social attribution, this in no way invalidates Gallup's (1982) proposal. A revised version of the model, and one implied to some extent by Gallup and Suarez (1986), could predict that self-recognition is an empirical marker of the beginning of a complex developmental sequence involving multiple feedbacks between self- and social attribution. Thus, although the yes-or-no presence of self-recognition may not be isomorphic with all forms of mental state attribution (as suggested in Gallup's original formulations of his model), it could be seen as an easily detectable, correlated measure of a nearly invariant sequence of ontogenetic events. In other words, the emergence of the self-attributional capacities that underpin self-recognition may set in motion (or merely co-vary with) a series of cascading ontogenetic constructions related to an understanding of the mental world. Recent research with young children strongly supports this ver-

---

17 Although this may be a conceptually valid strategy, the difficulty arises in that relatively little is known about the attributional capacities of two-year-old children (see Miller & Aloise, 1989, for a review).

18 If this is correct, it is tempting to speculate that for each species capable of self-recognition, there may well be a mature and immature form of the capacity giving rise to the behavior. As the organism develops, its capacity for self-recognition may not change in any initially obvious ways, but its underlying appreciation of the image may well undergo dramatic shifts as the self-system becomes increasingly differentiated through time (e.g., Mounoud & Vinter, 1985). We are currently exploring this possibility in an age-graded colony of chimpanzees ranging from two-year-olds to adult animals.
sion of Gallup's framework by demonstrating that the emergence of mirror self-recognition seems tightly correlated with the development of behaviors that appear to depend on primitive forms of social attribution (Asendorpf & Baudonniere, 1993; Bischof-Kohler, 1988; Johnson, 1982).

My colleagues and I have recently launched a broad-based research program explicitly designed (in part) to test the ontogenetic predictions of the revision of Gallup's (1982) model that I am advocating here. If the model is of heuristic value, young chimpanzees ought to display little or no evidence of mental state attribution until after they show evidence of self-recognition. A demonstration to the contrary would neatly falsify Gallup's model. Such a research program is methodologically viable given that recent evidence strongly suggests that most chimpanzees do not display evidence of self-recognition until six–eight years of age (Povinelli, Rulf, Landau, & Bierschwale, in press). Thus, physically mature subjects can be tested on the same tasks both before and after the onset of self-recognition. For instance, we have recently tested six three- and four-year-old chimpanzees on our seeing-knowing task, and the results largely conform to the predictions of the model (Povinelli, Rulf, & Bierschwale, in press).

Autistic children provide an interesting test case for the idea that self-recognition can be decoupled from some but not all forms of mental state attribution. Autistic children older than approximately five years of age display clear signs of self-recognition (Neuman & Hill, 1978). A flurry of recent research on the autistic child's theory of mind, however, has revealed striking differences between children who are autistic and those who are not. Autistic children show some early forms of social attribution (Level 1 perspective-taking and possibly the perception-knowledge relationship), but they do not appear to reach a stage at which they understand false belief (Baron-Cohen, Leslie, & Firth, 1985, 1986; Hobson, 1984; Leslie & Firth, 1988; Perner, Firth, Leslie, & Leekman, 1989). Thus, self-recognition remains correlated with some but not necessarily all forms of social attribution. It is possible that autistic children may suffer from a breakdown in the development of certain aspects of the self-attribution developmental system, which shows up in parallel ways in their capacity for social attribution. The broader issue, however, is that Gallup's (1982) model could be falsified if self-recognition could be shown to exist in children (or other organisms) who possess no capacity for self- or social attribution, or vice versa.

Comparative Ontogeny of Mental State Attribution

This asynchrony between the development of self-recognition and many forms of social attribution in children raises at least two logical possibilities for the attributional capacities of apes that show clear evidence of self-recognition. On the one hand, it is possible that these apes develop only the attributional capacities that are tightly correlated with the onset of self-recognition in human children. However, it is also possible that self-recognition in these apes signals the beginning of the development of more complex attributional processes, as in human children. This in no way should be interpreted as meaning that the behavior of self-recognition causes mental state attribution. Rather, it may be one of the earliest signs of its onset. Thus, it is possible that the underlying self-attributional process that supports self-recognition induces the onset of a concurrently developing epigenetic program of self- and social attribution.

As Gopnik (in press) has pointed out, however, it remains a matter for empirical research to determine if analogous self- and social attributional processes develop jointly or emerge relatively sooner or later with respect to each other. In essence, I am proposing that research efforts seek to determine to what extent humans and great apes share a homologous developmental program implicated in the emergence of these cognitive capacities.

Some may object to this method of proceeding, arguing that I appear to believe that chimpanzees, for instance, are developmentally arrested young children, or that young children are relatively less hairy chimpanzees. However, a careful reading of our work will show that this claim is unfounded. As we have pointed out, for strictly procedural reasons if comparative psychologists are ever to make progress toward closure on whether any nonhuman primates attribute mental states, we must have a basis for comparison. Tasks of mental state attribution differ from traditional learning tasks precisely because at present there is only one species that we know for certain possess attributions capacities: humans. Thus, in order to know the mind of the chimpanzee or macaque, we must first know our own (Povinelli & deBlois, 1992a). Conversely, I would add that if we cannot understand the development of our own minds, then we have no hope of understanding the minds of apes. This highlights the need for methodological improvements in nonhuman theory of mind research. Elsewhere, I have made concrete proposals for the design of such experiments (Povinelli, 1991).

There is also a second, theoretical reason for advocating such comparisons. If our goal as evolutionary biologists is to reconstruct the sequence in which particular mental capacities evolved and to understand the consequences of this evolution, then we must compare the ontogenies of such capacities in closely related organisms (see Parker & Gibson, 1979). Just as evolutionary biologists compare morphological ontogeny as a means of inferring homology, and hence as a tool in reconstructing descent (e.g., Luckett, 1975), so too can this method be used as a means of distinguishing between homology and...

\[\text{It should be noted that if this account proves correct, the results of our (and Premack's, 1988) seeing-knowing experiment might have to be accounted for in terms of traditional learning theory. Another possibility exists, however. Chimpanzees (perhaps like three-year-old children) may attribute knowledge and act on the basis of these attributions without doing so correctly. Thus, they might solve the problem through trial and error but appeal to mental states as opposed to environmental variables.}\]
convergence in psychological structures. Such an approach fully recognizes that each species is also likely to possess its own unique specializations (autapomorphic traits).20

The Loss of Self-Recognition in Gorillas

Comparative ontogeny may also be applied fruitfully in the context of the fact that whereas most gorillas tested to date have not shown compelling indications of self-recognition (Ledbetter & Basen, 1982; Suarez & Gallup, 1981), at least one home-reared gorilla clearly does (Patterson, 1984).21 Chimpanzees, gorillas, humans, and orangutans all share a common ancestor, but orangutans appear to have diverged from this common stock much earlier than did the African apes and humans. Nonetheless, orangutans, but not gorillas, show full signs of self-recognition (Lethmate & Dicker, 1973). From a cladistic perspective, this is strong evidence that the common ancestor of the great ape–human clade was capable of self-recognition (see Figure 2). To date, the evolutionary treatment of this intuitively puzzling situation, including my own, has been rather simplistic (e.g., Gallup, 1985; Povinelli, 1987). We have simply noted that it appears that the most parsimonious explanation is that gorillas have undergone a secondary loss of the capacity. Although this may be correct, I believe that a richer understanding of this issue probably lies in a careful consideration of the effect of environmental perturbations on development.

As the result of the pioneering work of Hampé (1960), experimental embryologists have long known that simple environmental perturbations during the course of early ontogeny can result in the appearance of ancestral morphological characteristics (see Frazzetta, 1975). In discussing Hampé’s experiments, which demonstrated that the legs of chicken embryos can be made to revert to the ancestral condition of Archaeopteryx (the first bird) by simple experimental interventions, Alberch, Gould, Oster, and Wake (1979) noted

It is probable that the genetic capacity for producing these ancestral traits was never lost during more than 200 million years of avian evolution but merely “turned off” by failure of the inducing ribula to establish contact with the metatarsals. (p. 311)

Other examples make the same point (see Frazzetta, 1975; Kollar, 1972; Raikow, 1975; see especially Kollar & Fisher,

20 This leaves open the question of which of the living species of the great ape–human clade most closely resembles the common ancestor of the group with respect to its epigenetic program related to mental state attribution. As a working hypothesis, we have argued that given the recent and rapid change in the size and organization of the human brain beginning approximately 2 million years ago with the appearance of Homo habilis (Falk, 1987; Holloway, 1975; Tobias, 1971), the cognitive processes of the living great apes may be more similar to the common ancestor than are those of modern humans (Povinelli & Godfrey, in press).

21 Some may object to this characterization of the state of affairs of the self-recognition capacity in gorillas, citing instances of facial movements in gorillas or isolated anecdotes of what appear to be self-directed behaviors on the part of a particular gorilla (e.g., Parker, 1990). In our laboratory, we have to date investigated nearly 50 young chimpanzees and their reactions to mirrors. All show at least some instances of what might otherwise be considered weak signs of self-recognition. Yet these same subjects display these same behaviors while not in front of the mirror as well. Thus, these weak instances cannot be used as uniquely positive evidence of self-recognition. In clear contrast, subjects mature enough for self-recognition display clear, sustained, and unambiguous signs of self-exploration while carefully monitoring their reflections (see Povinelli, Ruff, Landau, & Bierschwalke, in press).
Indeed, Alberch et al. concluded that “latent capacities for the generation of ancestral structures probably exist in the genomes of all specialized animals” (p. 311).

Recent research on the cognitive development of gorillas suggests that, relative to chimpanzees, at least, they have undergone heterochronic changes in the onset of quadrupedal locomotion relative to the acquisition of particular cognitive structures (Antinucci, 1989, chap. 6). In particular, their locomotion appears to mature faster, thus potentially interfering with further growth in cognition. Other aspects of their maturation appear to have speeded up as well, relative to chimpanzees and orangutans (see Watts & Pusey, in press). Thus, specific changes in the onset and timing of particular sensorimotor cognitive pathways may have led to the secondary loss of self-recognition in gorillas. Selective pressures favoring earlier physical maturation in gorillas may have resulted in the secondary loss of a suite of cognitive structures implicated in the capacity for self-conception. In this context, Patterson’s (1984) evidence for self-recognition by a home-reared gorilla that has been instructed in a version of American Sign Language can be viewed as being analogous to an investigation in the field of experimental embryology. In this case, the experimental intervention has been a potentially profound perturbation involving intensive tutoring on numerous cognitive tasks during early ontogeny (see Patterson & Linden, 1981).

The general hypothesis that I am advancing here is that strong environmental inputs may affect psychological ontogeny in the same manner as they affect morphological ontogeny and thus may be capable of resulting in the expression of apparently lost ancestral psychological traits (in this case, the self-attributorial capacities that underpin self-recognition; Povinelli, in press). To accept this hypothesis as having potential merit, one need only accept two simple assumptions: first, that changes in neural structures underlie the evolution and ontogeny of cognitive capacities and, second, that normal epigenetic interactions determine the course of brain development. If true, then gorillas may well still possess the basic genetic instructions for the psychological capacities underlying self-recognition, which have been “turned off” because of heterochronic changes in other aspects of their development. Experimental interventions such as those received by Patterson’s (1984) gorilla, may be capable of reinstating their psychological capacity for self-recognition. From a heuristic standpoint it is important to note that this model of the loss of self-recognition in gorillas explicitly predicts that the same experimental perturbations that Patterson’s gorilla received would not have a similar effect on macaques, for instance, because they descended from a lineage in which this capacity never existed.22

Summary: The Promise of Ontogeny

Ultimately, comparing the ontogeny of psychological capacities should allow evolutionary psychologists to reconstruct the order in which particular features of mental state attribution evolved (Povinelli & Godfrey, in press).

By determining what (if any) characteristics of a theory of mind each living species possesses, it will be possible to specify what the likely features of any given common ancestor were. In doing so, a major goal of evolutionary psychology will have been achieved—understanding the sequence in which features of mental state attribution evolved. From the perspective of modern evolutionary biology, the question at least now seems clear: What portions of domain-specific epigenetic programs do organisms share as a result of common descent, and where in various lineages have innovations been introduced at various points into these programs?

Psychological Continuity Revisited

In the course of preparing this article, I was reminded by a colleague (Gordon Gallup) that the surest proof that psychological discontinuities can exist between species can be found in the context of within-species comparisons of ontogeny (see Goldman-Rakic & Preuss, 1987). Reconsider the case of three- and four-year-old children: If very young three-year-olds, for instance, truly do not understand how knowledge states arise, then consider how different their perception of the social universe must be from our own. They must, by definition, be left in a world in which knowledge appears miraculously (Leslie, 1987) or perhaps even according to some ad hoc rules that we as developmental psychologists have not yet begun to specify. Even more striking, consider how different that same three-year-old must be from a 15-month-old infant who may not even have psychological access to the concept of knowledge. Indeed, that same 15-month-old may not even have access to his or her own mind.

Species discontinuities in psychology can be usefully thought of in this same light. Within closely related taxa, each species may well represent its own unique constellation of ontogenetic end points. For the vast majority of species, these end points may not seem so profoundly different to us, although they certainly must be to the organisms themselves, because some subset of the resulting morphological and psychological differences are what define species boundaries (Paterson, 1985). But when it comes to groups of species in which major innovations in psychology have occurred, the ontogenetic end points are potentially of profound interest. If psychological evolution, like morphological evolution, can be meaningfully thought of as innovations in ontogeny, then comparative psychologists should have no trouble accepting as a possibility that a major evolutionary change occurred quite recently in primate history. This does not require adhering to any form of recapitulation theory per se. It requires only accepting an evolutionary framework that attempts to reconstruct likely ancestral ontogenies

22 Of course, the opposite effect is also possible. Strong, novel environmental inputs into the cognitive epigenetic system may lead to the expression of traits of descendant species as well. In a related vein, Morin and deBlos (1989) argued that Gallup underestimated the effect of giving chimpanzees exposure to mirrors. They speculated that chimpanzees who gain the ability to recognize themselves may undergo significant cognitive change vis-à-vis those who do not.
by means of comparing the similarities and differences of extant ontogenies.

The past few hundred million years of higher metazoan evolution may have been characterized by organisms that had "discovered" how certain categories of reality, such as time, space, and causality, govern the physical world. With the advent of even a rudimentary brain, many of these organisms can, in some sense, be said to have been operating on the basis of "intentions" and "knowledge" stored in their neural circuitry. Such operations require (in an implicit sense) that an organism take into account the relationships between time and space and hence that they react as if they had a concept of causality. But what remained to be discovered by such organisms was the existence of these mental concepts and the understanding that they can be causative agents of behavior. In other words, what remained to evolve was an awareness of the reality of the mental world—the evolution of metacognition. Such capacities can easily be seen to be advantageous because they can construct more accurate descriptions (or models) of reality. After all, organisms do operate on the basis of simple-to-complex representations of reality stored in their brains, even if they are incapable of realizing it. To become aware of such representations does not require a detailed understanding of the biological hardware of the system in which these mental states are encoded. An abstract understanding of the mere existence of such states would be of sufficient advantage. The task awaiting empirical research is to reconstruct the timing and order in which the pieces of this profoundly new psychology emerged.

The possible outcomes of such an endeavor run in all possible phylogenetic directions. Our methodologies are still too young, our experiments far too sparse, and our subjects far too few, to believe that we have come to closure on which nonhuman species possess various components of a theory of mind. I am not yet convinced that even chimpanzees possess a true awareness of the private mental world, despite the growing body of suggestive evidence (for detailed reviews of this evidence, see Cheney & Seyfarth, 1990b; Povinelli, 1991; Povinelli & Godfrey, in press; Premack, 1988; White, 1993). Chimpanzees and the other great apes may either affirm or disappoint those who wish to find other minds as replete with social intelligence as our own. On the other hand, the diversity of minds in the animal kingdom may still surprise those who see the psychological capacities of the great apes and humans as fundamentally derived. Resolving these questions will ultimately allow us to appreciate the diversity of ways of knowing about the mental world that evolution has unknowingly produced.

REFERENCES


Gallup, G. G. Jr., & Povinelli, D. J. (in press). Mirror, mirror, on the wall, which is the most heuristic theory of them all? New Ideas in Psychology.


judged to recognize themselves. Paper presented at the 13th annual meeting of the American Society of Primatologists, Davis, CA.


Pratt, C., & Bryant, P. (1990). Young children understand that looking leads to knowing (so long as they are looking into a single barrel). *Child Development*, 61, 973–982.


