In: Metarepresentations: A Multidisciplinary Perspective,
Dan Sperber, ed., NY: Oxford, 2000

Chapter 4

Consider the Source: The Evolution of Adaptations for Decoupling and Metarepresentation

Leda Cosmides and John Tooby

The Cognitive Niche and Local Information

Humans are often considered to be so distinct a species that they are placed outside of the natural order entirely, to be approached and analyzed independently of the rest of the living world. However, all species have unusual or differentiating characteristics and it is the task of an evolutionarily informed natural science to provide a causal account of the nature, organization, origin, and function, if any, of such characteristics without exaggerating, mystifying, or minimizing them.

Yet, even when placed within the context of the extraordinary diversity of the living world, humans continue to stand out, exhibiting a remarkable array of strange and unprecedented behaviors – from space travel to theology – that are not found in other species. What is at the core of these differences? Arguably, one central and distinguishing innovation in human evolution has been the dramatic increase in the use of contingent information for the regulation of improvised behavior that is successfully tailored to local conditions – an adaptive mode that has been labeled the *cognitive niche* (Tooby & DeVore, 1987). If you contrast, for example, the food acquisition practices of a Thompson's gazelle with that of a !Kung San hunter, you will immediately note a marked difference. To the gazelle, what looks to you like relatively undifferentiated grasslands is undoubtedly a rich tapestry of differentiated food patches and cues; nevertheless, the gazelle's decisions are made for it by evolved, neural specializations designed for grass and forage identification and evaluation – adaptations that are universal to the species, and that operate with relative uniformity across the species range. In contrast, the !Kung hunter uses, among many other means and methods that are not species-typical, arrows that are tipped with a poison found on only one

local species of chrysomelid beetle, toxic only during the larval stage (Lee, 1993). Whatever the neural adaptations that underlie this behavior, they were not designed specifically for beetles and arrows, but exploit these local, contingent facts as part of a computational structure that treats them as instances of a more general class.

Indeed, most species are locked in co-evolutionary, antagonistic relationships with prey, rivals, parasites, and predators, in which move and countermove take place slowly, over evolutionary time. Improvisation puts humans at a great advantage: instead of being constrained to innovate only in phylogenetic time, they engage in ontogenetic ambushes[1] against their antagonists – innovations that are too rapid with respect to evolutionary time for their antagonists to evolve defenses by natural selection. Armed with this advantage, hominids have exploded into new habitats, developed an astonishing diversity of subsistence and resource extraction methods, caused the extinctions of many prey species in whatever environments they have penetrated, and generated an array of social systems far more extensive than that found in any other single species.

This contrast – between local, contingent facts and relationships that hold over the species' range – is at the heart of what makes humans so different. To evolve, species-typical behavioral rules must correspond to features of the species' ancestral world that were both globally true (i.e., that held statistically across a preponderance of the species' range) and stably true (i.e., that remained in effect over enough generations that they selected for adaptations in the species). These constraints narrowly limit the kinds of information that such adaptations can be designed to use: the set of properties that had a predictable relationship to features of the species' world that held widely in space and time is a very restricted one. In contrast, for situation-specific, appropriately tailored improvisation, the organism only needs information to be applicable or "true" temporarily, locally, or contingently. If information only needs to be true temporarily, locally, and situationally to be useful, then a vastly enlarged universe of context-dependent information becomes potentially available to be employed in the successful regulation of behavior. This tremendously enlarged universe of information can be used to fuel the identification of an immensely more varied set of advantageous behaviors than other species employ, giving human life its distinctive complexity, variety, and relative success. Hominids entered the cognitive niche, with all its attendant benefits and dangers, by evolving a new suite of cognitive adaptations that are evolutionarily designed to exploit this broadened universe of information, as well as the older universe of species-extensive true relationships.

The hominid occupation of the cognitive niche is characterized by a constellation of interrelated behaviors that depend on intensive information manipulation and that are supported by a series of novel or greatly elaborated cognitive adaptations. This zoologically unique constellation of behaviors includes locally improvised subsistence practices; extensive context-sensitive manipulation of the physical and social environment; "culture," defined as the serial reconstruction and adoption of representations and regulatory variables found in others' minds through inferential specializations evolved for the task; language as a system for dramatically lowering the cost of communicating propositional information; tool use adapted to a diverse range of local problems; context-specific skill acquisition; multi-individual coordinated action; and other information-intensive and information-dependent activities (Tooby & Cosmides, 1992). Although social interactions may have played a role, we do not believe that social competition was the sole driving force behind the evolution of human intelligence (as in the Machiavellian hypothesis; Humphrey, 1992; Whitten & Byrne,1997). We certainly do believe that humans have evolved sophisticated adaptations specialized for social life and social cognition (e.g., Cosmides, 1989; Cosmides & Tooby, 1989; 1992), but what is truly distinctive about human life encompasses far more than the social. For example, the causal intelligence expressed in hunter-gatherer subsistence practices appears to be as divergent from other species as human social intelligence.

The benefits of successful improvisation are clear: the ability to realize goals through exploiting the unique opportunities that are inherent in a singular local situation yields an advantage over a system that is limited to applying only those solutions that work across a more general class of situation. What ten years of ordinary battle on the plains of Troy could not accomplish, one Trojan Horse could. The improvisational exploitation of unique opportunities also fits our folk intuitions about what counts as intelligence. As members of the human species, instances of intelligence excite our admiration precisely to the extent that the behavior (or insight) involved is novel and not the result of the "mindless" application of fixed rules. Indeed, it would seem that every organism would be benefitted by having a faculty that caused it to perform behaviors adapted to each individual situation. This raises a question: Why haven't all organisms evolved this form of intelligence? Indeed, how is this form of intelligence possible at all?

Elsewhere, we have written at length about the trade-offs between problem-solving power and specialization: general-purpose problem-solving architectures are very weak but broad in application, whereas special-purpose problem-solving designs are very efficient and inferentially powerful but limited in their domain of application (Cosmides & Tooby, 1987; Tooby & Cosmides, 1992). Thus, on first inspection, there appear to be only two biologically possible choices for evolved minds: either general ineptitude or narrow competences. This choice

rules out general intelligence. Traditionally, many scholars have assumed that because human intelligence appears unprecedentedly broad in application, the human cognitive architecture's core problem-solving engines must themselves be general-purpose. This has led to a fruitless insistence that viable candidate models of this architecture be largely free of special-purpose machinery. This insistence has, in our view, obstructed progress toward an accurate model of the human psychological architecture. Because general-purpose problem-solvers are too weak to supply the problem-solving power evolved organisms need to carry out the array of complex and arduous tasks they routinely face, human intelligence cannot consist primarily of domain-general computational engines. Instead of achieving general intelligence through general-purpose mechanisms, there is another alternative: Cognitive specializations, each narrow in its domain of application, can be bundled together in a way that widens the range of inputs or domains that can be successfully handled. More general-purpose engines can be embedded within this basic design (because their defects when operating in isolation can be offset by implanting them in a guiding matrix of specializations). Moreover, other architectural features are required to solve the problems raised by the interactions of these heterogeneous systems, as discussed below (Tooby & Cosmides, 1990a; 1992; in press). This is the only solution that we can see to the question of how human intelligence can be broad in its range of application but also sufficiently powerful when applied (Sperber, 1996; Tooby & Cosmides, 1990a; 1992).

Even so, the costs and difficulties of the cognitive niche are so stringent that only one lineage, in four billion years, has wandered into the preconditions that favored the evolution of this form of intelligence. Natural computational systems that begin to relax their functional specificity run into, and are inescapably shaped by, savagely intense selection pressures. One of the greatest problems faced by natural computational systems is combinatorial explosion (for discussion, see Cosmides & Tooby, 1987; Tooby & Cosmides, 1992). Combinatorial explosion is the term for the fact that alternatives multiply with devastating rapidity in computational systems, and the less constrained the representational and procedural possibilities, the faster this process mushrooms. When this happens, the system is choked with too many possibilities to search among or too many processing steps to perform. Marginally increasing the generality of a system exponentially increases the cost, greatly limiting the types of architectures that can evolve, and favoring, for example, the evolution of modules only in domains in which an economical set of procedures can generate a sufficiently large and valuable set of outputs. This means that domain-specificity will be the rule rather than the exception in natural computational systems.

A second difficulty is that, from evolutionary and computational perspectives, it is far from clear how local improvisation could evolve, operate, or even be a non-magical, genuine cognitive possibility. A computational system, by its nature, can only apply rules or procedures to problems, and must do so based on its rule-based categorization of individual problems into more general classes, so that it knows which procedures to activate in a given situation.[2] Moreover, natural selection is a statistical process that tested alternative computational designs against each other, summing over billions of individual lives and test runs, taking place over thousands of generations. A gene (and its associated design feature) could only have been selected to the extent that it operated well against the statistical regularities that recurred across vast ancestral populations of events. That is, the iterated conditions that the adaptation evolved to deal with must have extended over enough of the species range, and for an evolutionary period that was long enough to spread the underlying genes from their initial appearance as mutations to near universality.[3] In consequence, adaptations can only see individual events in the life of the organism as instances of the large-scale evolutionarily recurrent categories of events that built them (Tooby & Cosmides, 1990a). So, if computational systems can only respond to situations as members of classes to which computational rules apply, and if evolution only builds computational adaptations that "see" individual situations as members of large scale, evolutionarily recurrent classes of events, how can there be a brain whose principles of operation commonly lead it to improvise behaviors that exploit the distinctive features of a situation? How could species-typical computational rules evolve that allow situation-specific improvisation at all – or at sufficiently low cost? We will address several of these questions elsewhere (Cosmides & Tooby, in press). In this chapter, we shall concentrate on only one set of engineering problems associated with the exploitation of contingent information – what we call the *scope problem*.

The Scope Problem

When hominids evolved or elaborated adaptations that could use information based on relationships that were only "true" temporarily, locally, or contingently rather than universally and stably, this opened up a new and far larger world of potential information than was available previously. Context-dependent information could now be used to guide behavior to a far greater extent than had been possible formerly. This advance, however, was purchased at a cost: The exploitation of this exploding universe of potentially representable information creates a vastly expanded risk of possible misapplications, in which information

that may be usefully descriptive in a narrow arena of conditions is false, misleading, or harmful outside the scope of those conditions.[4] Exactly because information that is only applicable temporarily or locally begins to be used, the success of this computational strategy depends on continually monitoring and re-establishing the boundaries within which each representation remains useful. Are the beetle larvae that are used to poison arrows toxic at all times of the year? Once harvested and applied, how long does the poisoned arrow tip remain poisonous? If it is poisonous to humans, gazelles, and duikers, is it also poisonous to lions, cape buffalo, and ostriches? If these relationships are true here, are they true on foraging territories on the other side of the Okavango? If the first several statements from my father in answer to these questions turned out to be true, will the remainder be true also? Information only gives an advantage when it is relied on inside the envelope of conditions within which it is applicable. Hence, when considering the evolution of adaptations to use information, the costs of overextension and misapplication have to be factored in, as do the costs and nature of the defenses against such misapplication. Expanding the body of information used to make decisions is harmful or dangerous if the architecture does not and cannot detect and keep track of which information is applicable where, and how the boundaries of applicability shift (Tooby & Cosmides, in press).

Moreover, the problem is not simply that information that is usefully descriptive only within a limited envelope of conditions will (by definition) be false or harmful outside the scope of those conditions. The scope problem is aggravated by the fact that information is integrated and transformed through inferences. Information is useful to the extent that it can be inferentially applied to derive conclusions that can then be used to regulate behavior. Inferences routinely combine multiple inputs through a procedure to produce new information, and the value of the resulting inferences depends sensitively on the accuracy of the information that is fed into them. For example, the truth of the conclusion that it will be better to move to an area where there is more game is dependent on the proposition that there is more game in the new location, and on the implicit or explicit assumption that the necessary poisons for hunting can be obtained there as well.

Not only does inference combinatorially propagate errors present in the source inputs, but the resulting outputs are then available to be fed in as erroneous inputs into other inferences, multiplying the errors in successive chains and spreading waves. For example, if one wrong entry is made in a running total, all subsequent totals – and the decisions based on them – become wrong. This process has the potential to corrupt any downstream data-set interacted with, in a spreading network of compounding error. The more the human cognitive architecture is networked together by systems of intelligent inference, and the more it is enhanced by the ability to integrate information from many sources,[5] the greater the risk that valid existing information sets will be transformed into unreconstructable tangles of error and confusion. In short, the heavily inference-dependent nature of human behavior regulation is gravely threatened by erroneous, unreliable, obsolete, out-of-context, deceptive, or scope-violating representations.

For these reasons, the evolution of intelligence will depend critically on the economics of information management and on the tools of information management – that is, the nature of the adaptations that evolve to handle these problems. The net benefit of evolving to use certain classes of information will depend on the cost of its acquisition, the utility of the information when used, the damage of acting on the information mistakenly outside its area of applicability, and the cost of its management and maintenance. Because humans are the only species that has evolved this kind of intelligence, humans must be equipped with adaptations that evolved to solve the problems that are special to this form of intelligence.

Scope Syntax, Truth, and Naïve Realism

For these reasons, issues involving not only the accuracy but also the scope of applicability of the information that the individual human acquires and represents became paramount in the design and evolution of the human cognitive architecture. We believe that there are a large number of design innovations that have evolved to solve the specialized programming problems posed by using local and contingent information, including a specialized scope syntax, metarepresentational adaptations, and decoupling systems. Indeed, we think that the human cognitive architecture is full of interlocking design features whose function is to solve problems of scope and accuracy. Examples include truth-value tags, source-tags (self versus other; vision versus memory, etc.), scope-tags, time-and-place tags, reference-tags, credal values, operators embodying propositional attitudes, content-based routing of information to targeted inference engines, dissociations, systems of information encapsulation and interaction, independent representational formats for different ontologies, and the architecture and differential volatility of different memory systems. One critical feature is the capacity to carry out inferential operations on sets of inferences that incorporate suppositions or propositions of conditionally unevaluated truth value, while keeping their computational products isolated from other knowledge stores until the truth or utility of the suppositions is decided, and the outputs are either integrated or discarded. This capacity is essential to planning,

interpreting communication, employing the information communication brings, evaluating others' claims, mind-reading, pretense, detecting or perpetrating deception, using inference to triangulate information about past or hidden causal relations, and much else that makes the human mind so distinctive. In what follows, we will try to sketch out some of the basic elements of a scope syntax designed to defuse problems intrinsic to the human mode of intelligence.

By a scope syntax, we mean a system of procedures, operators, relationships, and data-handling formats that regulate the migration of information among subcomponents of the human cognitive architecture. To clarify what we mean, consider a simple cognitive system that we suspect is the ancestral condition for all animal minds, and the default condition for the human mind as well: naïve realism. For the naïve realist, the world as it is mentally represented is taken for the world as it really is, and no distinction is drawn between the two. Indeed, only a subset of possible architectures is even capable of representing this distinction and, in the origin and initial evolution of representational systems, such a distinction would be functionless. From our external perspective, we can say of such basic architectures that all information found inside the system is assumed to be true, or is treated as true. However, from the point of view of the architecture itself, that would not be correct, for it would imply that the system is capable of drawing the distinction between true and false, and is categorizing the information as true. Instead, mechanisms in the architecture simply use the information found inside the system to regulate behavior and to carry out further computations. Whatever information is present in the system simply is "reality" for the architecture. Instead of tagging information as true or false – as seems so obvious to us – such basic architectures would not be designed to store false information. When new information is produced that renders old information obsolete, the old information is updated, overwritten, forgotten, or discarded. None of these operations require the tagging of information as true or false. They only involve the rule-governed replacement of some data by other data, just like overwriting a memory register in a personal computer does not require the data previously in that register be categorized as false. For most of the behavior-regulatory operations that representational systems evolved to orchestrate, there would be no point in storing false information, or information tagged as false. For this reason, there is no need in such an architecture to be able to represent that some information is true: Its presence, or the decision to store it or remember, it is the cue to its reliability. In such a design, true equals accessible.

With this as background, and leaving aside the many controversies in epistemology over how to conceptualize what truth "really" is, we can define what we will call *architectural truth:* information is treated by an architecture as true when it is allowed to migrate (or be reproduced) in

an unrestricted or scope-free fashion throughout an architecture, interacting with any other data in the system with which it is capable of interacting. All data in semantic memory, for example, is architecturally true. The simplest and most economical way to engineer data use is for "true" information to be unmarked, and for unmarked information to be given whatever freedom of movement is possible by the computational architecture. Indeed, any system that acquires, stores, and uses information is a design of this kind. The alternative design, in which each piece of information intended for use must be paired with another piece of information indicating that the first piece is true, seems unnecessarily costly and cumbersome. Because the true-is-unmarked system is the natural way for an evolved computational system to originate, and because there are many reasons to maintain this system for most uses, we might expect that this is also the reason why humans – and undoubtedly other organisms – are naïve realists. Naïve realism seems to be the most likely starting point phylogenetically and ontogenetically, as well as the default mode for most systems, even in adulthood.

The next step, necessary only for some uses, is to have representations described by other data structures: metarepresentations (in a relaxed rather than narrow sense). For example, a cognitive architecture might contain the structure, *The statement that "anthropology is a science" is true.* This particular data structure includes a proposition (or data element) and an evaluation of the truth of the proposition (or data element).[6] However, such structures need not be limited to describing single propositions. Although it is common in talking about metarepresentations and propositional attitudes to depict a single representation embedded in an encompassing proposition, a single proposition is only a limiting case. A set of propositions or any other kind of data element can be bundled into a single unit that is taken, as a data packet, as an argument by a scope operator to form a metarepresentation. For example, the metarepresentation, *Every sentence in this chapter is false,* describes the truth value of a set of propositions[7] as easily as *The first sentence in this chapter is false* describes the truth value of a single proposition. Indeed, sometimes integrated sets of propositions governed by a superordinate scope operator might become so elaborated, and relatively independent from other data structures, that they might conveniently be called worlds. We think large amounts of human knowledge inside individuals exist inside data structures of this kind.

A sketch of the kind of cognitive architecture and operators we have in mind begins with a primary workspace that operates in a way that is similar, in some respects, to natural deduction systems (see Gentzen, 1969/1935; Rips, 1994; Cosmides & Tooby, 1996b), although it may include axiom-like elements, and many other differences, as well. Its general features are familiar: There is a workspace containing active data elements;

procedures or operators act on the data structures, transforming them into new data structures. Data structures are maintained in the workspace until they are overwritten, or if not used or primed after a given period of time, until they fade and are discarded. Products may be permanently stored in appropriate subsystems if they meet various criteria indicating they merit long-term storage, or warrant being treated as architecturally true. Otherwise, the contents and intermediate work products of the workspace are volatile, and are purged, as one adaptation for protecting the integrity of the reliable data stores elsewhere in the architecture. Data structures may be introduced from perception, memory, supposition, or from various other system components and modules. Some of the procedures and tags available in the workspace correspond to familiar logical operators and elements such as variable binding, instantiation, 'if' introduction and 'if' elimination, the recognition and tagging of contradictions, *modus ponens*, and so on. Some of the procedures are ecologically rational (Tooby & Cosmides, in press; Cosmides & Tooby, 1996a), that is, they correspond to licensed transformations in various adaptive logics (which may diverge substantially from licensed inferences in the content-independent formal logics developed so far by logicians). Indeed, many procedures consist of routing data structures through adaptive specializations such as cheater detection or hazard management algorithms (Cosmides, 1989; Cosmides and Tooby, 1997), with outputs placed back into the workspace: a process that resembles either calling subroutines or applying logical transformations, depending on one's taste in formalisms.[8] Deliberative reasoning is carried out in this workspace: while many other types of inference are carried out automatically as part of the heterogeneous array of specializations available in the architecture. Some areas of this workspace are usually part of conscious awareness, and most are consciously accessible.

The data sets in this system exist in structured, hierarchical relations,[9] which we will represent as indented levels. Data elements in the left-most position are in what might be thought of as the ground state, which means they are licensed to migrate anywhere in the architecture they can be represented. They may enter into inferences in combination with any other ground state data-structure, and (usually) may be permitted to interact with subordinate levels as well: They are architecturally true, or scope-free. Other elements are subordinated under ground state elements through scope operators. So, we might represent an architecturally true statement in the leftmost position, such as:

(1) Anthropology is a science.

It is unmarked by the architecture, and is free to be stored or to be introduced into any other nondecoupled process in the architecture. A subordinated statement may be scope-limited, such as:

(2) The statement is false that:

(3) Anthropology is a science.

In this case, the scope operator (2) binds the scope within which the information of the data structure (3) can be accessed, so that (3) is not free to be promoted to the ground state or to be used elsewhere in the system. In contrast, the function of an explicit true tag in a statement description operator (i.e., *The statement is true that p)* would be to release the statement from previous scope-restriction, promoting it to the next leftmost level or, if it was originally only one level down, changing its status to unmarked or architecturally true.[10] Time and location operators operate similarly:

(4) In ≠Tobe (!Kung for "autumn"),

(5) the mongongo nuts become edible and plentiful.

Or,

(6) At Nyae Nyae,

(7) there are chrysomelid beetles suitable for making arrow poison.

Scope operators define, regulate, or modify the relationships between sets of information, and the migration of information between levels. They involve a minimum of two levels, a superordinate (or ground) and subordinate level. In these cases, the subordinate propositions cannot be reproduced without their respective scope-tags, which describe the boundary conditions under which the information is known to be accurate, and hence which license their use in certain inferences but not others. As with classical conditioning, we expect that additional mechanisms are designed to keep track of the reality of the scope boundaries; e.g., observing a lack of contingency outside the boundaries may eventually release the restriction. Thus, (6)–(7) may be transformed into (7) for an individual whose travels from camp to camp are typically inside the beetle species' range. Conversely, architecturally true statements like (1) can be transformed by a scope operation into something scope-limited, as new information about its boundary conditions is learned. A time-based scope transformation would be:

(8) It is no longer true that

(9) anthropology is a science.

Scope operators regulate the migration of information into and out of subordinated data sets, coupling (allowing data to flow) and decoupling them according to the nature of the operator and the arguments it

is fed. They bind propositions into internally transparent but externally regulated sets. In so doing, they provide many of the tools necessary to solve the problems posed by contingent information. By imposing bounds on where scope-limited information can travel (or what can access it), it allows information to be retained by the system and used under well-specified conditions, without allowing it to damage other reliable data-sets through inferential interaction. We will call representations that are bound or interrelated by scope operators *scope-representations* or *S-representations*.

Since computational features evolve because they enhance behavioral regulation, it is worth noting that these innovations markedly increase the range of possible behaviors open to the organism. In particular, one major change involves *acting as if*. The organism would be highly handicapped if it could only act on the basis of information known to be true or have its conduct regulated by architecturally true propositions, although this was likely to be the ancestral state of the organism. With the ability *to act as if p*, or *to act on the basis of p*, the organism can use information to regulate its behavior without losing any scope-represented restrictions on the nature of the information, or without necessarily losing a continuing awareness that the information acted on is not, or might not be, true. Conditions where such a behavioral-representational subsystem are useful include the many categories of actions undertaken under conditions of uncertainty (e.g., we will assume they got the message about the restaurant; or we will act as if there is a leopard hiding in the shadows of the tree), actions with respect to social conventions or deontic commitments (which are by themselves incapable of being either true or not true, at least in an ordinary sense; e.g., *Elizabeth is the rightful Queen of England; it is praiseworthy to make the correct temple sacrifices*), adapting oneself to the wishes of others, hypothesis testing, and so on.[11] Pretense (Leslie 1987) and deception (Whiten & Byrne, 1997) are simply extensions of this same competence, in which the agent knows the representations on which she is acting are false. These are simply the limiting cases rather than the defining cases. In order to get coordinated behavior among many individuals, and the benefits that arise from it, it is necessary to agree on a set of representations that will be jointly acted upon – a reason why social interaction so often involves the manufacture of socially constructed but unwarranted shared beliefs. Structures of representations can be built up that can be permanently consulted for actions without their contents unrestrictedly contaminating other knowledge stores.

Credal values and modals (*it is likely that p; it is possible that p; it is certain that p*) allow the maintenance and transformation of scope-marked information bound to information about likelihood and possibility – regulatory information that often changes while the underlying

propositions are conserved. Propositional attitude verbs (e.g., think, believe, want, hope, deny) are obviously also a key category of scope-operator, as we will discuss.

Supposition, Counterfactuals and Natural Deduction Systems

What makes such a system resemble, to a certain extent, natural deduction systems is the presence of scope-operators such as supposition, and the fact that these operators create subdomains or subordinate levels of representation, which may themselves have further subordinate levels, growing into multilevel, tree-like structures. Supposition involves the introduction of propositions of unevaluated or suspended truth value, which are treated as true within a bound scope, and then used as additional content from which to combinatorially generate inferential products. The operator "if," for example, opens up a suppositional world: for instance, *I am in my office this afternoon. If students believe I am not in my office this afternoon, then they won't bother me. If I close my door, and leave my light off, they will believe I am not here.* The contents of this suppositional world are kept isolated from other proposition-sets, so that true propositions are not intermixed and hence confused with false ones (e.g., *I am not in my office*) or potentially false ones (e.g., *they won't bother me*). Any number of subordinate levels can be introduced, with additional subordinate suppositions or other scope operations. A key feature of such a deduction system is the restricted application of inferences. Inferences are applied in a rule-governed but unrestricted fashion within a level – e.g., *students believe I am not in my office this afternoon*, therefore, *they won't bother me* – but not across levels – e.g., there is no contradiction to be recognized between *I am in my office this afternoon*, and the proposition *I am not in my office this afternoon*, because they are at different levels in the structure. Contents are architecturally true with respect to the level they are in and may enter into inferences at that level, while remaining false or unevaluated with respect to both the ground state of the architecture and other intermediate superordinate levels. Certain propositional attitudes (e.g., "believe" as opposed to "know") also decouple the truth value of the propositions ("I am not in my office") that are embedded in encompassing statements, a process that can be dissected computationally. Paradoxically, an architecture that only processes true information is highly limited in what it can infer, and most forms of human discovery by reasoning involve supposition. While some cases are famous (10), normal cases of suppositions are so numerous that they permeate our thoughts in carrying out routine actions in our daily lives (11).

(10) Suppose I threw this rock hard enough that the earth fell away in its curvature faster than the rock's downward ballistic took it?

(11) What if I hold my airline ticket in my teeth while I pick up the baby with my right arm and our bags with my left arm?

Supposition is a scope operation that suspends truth-values for all successive computations that result from taking the supposition as a premise. For example, (12) suspends the truth-value of (13):

(12) Suppose my wife, Desdemona, was unfaithful with Cassio.

(13) Then Cassio, whom I thought was my friend, has betrayed me.

Suppositions and their entailments remain internally interrelated and generative but isolated from the rest of the data in the architecture. If (13) were allowed to escape its scope-restriction to enter into ground-state originating inferences, the effects would be disastrous. Othello would have (13) as part of his uncritically accepted semantic store of propositions, without it being warranted (or "true" within the decoupled world of Shakespeare's *Othello*).[12] Nevertheless, S-representations like (12)–(13) allow many types of useful and revelatory reasoning to proceed – everything from proof by contradiction to the construction of contingency plans. Additionally, suppositions contain specifications of when subordinate deductions can be discharged. This occurs when other processes produce a true proposition that duplicates that supposition. Evidence establishing (12) as true discharges the supposition, promoting (13) to architectural truth and stripping it of its scope restrictions.

Actions can also discharge suppositions – a key point. Consider a hominid considering how to capture a colobus monkey in a tree. An architecture that cannot consider decoupled states of affairs is limited in the behaviors it can take (e.g., close distance with monkey). This may often fail because of the nature of the situation. For example, consider the situation in which there is a branch from the tree close to the branch of a neighboring tree. In this situation, the hominid confronts the following contingencies: If he climbs the trunk, then the monkey escapes by the branch. If he climbs across the branches, then the monkey escapes by the trunk. If, before taking action, the hominid suppositionally explores the alternative hunt scenarios, then it will detect the prospective failure. Moreover, given alternative inferential pathways leading to failure, the hominid, armed with the inferential power of supposition (and various other inferential tools, such as a model of the prey mind and a theory of mechanics), may then begin to consider additional courses of action suppositionally, reasoning about the likely consequences of each alternative.

Suppose there were no branch on the neighboring tree; then it could not be used as an escape route. Suppose, before I initiate the hunt by climbing up the trunk, I break that branch; then it could not be used as an escape route. If I then go up the trunk, the monkey cannot escape. The hunt will be a success. End search for successful outcome. Transform suppositional structure into a plan.

Conveniently for planning and action, the conditions for discharging a supposition specify the actions that need to be taken to put that aspect of the plan into effect, and the tree structure of suppositions provides the information about the order of the causal steps to be taken. Hominids armed with suppositional reasoning can undertake new types of successful behaviors that would be impossible for those whose cognitive architectures lacked such design features. It allows them to explore the properties of situations computationally, in order to identify sequences of improvised behaviors that may lead to novel, successful outcomes. The restricted application of inferences to a level, until suppositions (or other scope-limitations) are discharged is a crucial element of such an architecture. The states of affairs under the scope of a specific supposition are not mistaken for states of affairs outside that supposition: superordinate and subordinate relationships are kept clear until their preconditions can be discharged (as when an action is taken).

Like a clutch in an automobile, supposition and other scope operators allow the controlled engagement or disengagement of powerful sets of representations that can contain rich descriptions and acquired, domain-specific inference engines that can be applied when their preconditions are met. These operators provide vehicles whereby information that may or may not be counterfactual can be processed without the output being tagged as *true* and stored as such. Because contingent information can change its status at any time, with any new change in the world, it is important to have tools available that can take architecturally true information and scrutinize it. For example, the workspace that contains proposition *p* may benefit from demoting *p* into the scope-representation, *It appears that p*. Proposition *p* can still provide the basis for action, but can now be subjected to inferential processes not possible when it was simply a free representation at ground state. Demotion into a scope-representation brings a representation out of architectural truth and into a new relationship with the primary workspace. Because of this feature of the human cognitive architecture, humans can contingently refrain from being naïve realists about any specific data structure, although presumably we will always be naïve realists about whatever happens to be in the ground state in the workspace at any given time.[13]

Some operators are recursive, and some types of subordinated data structures can serve as the ground for further subordinated structures,

leading potentially to a tree structure of subordinated and parallel relations whose length and branching contingencies are restricted only by performance limitations of the system. For example:

(14) Chagnon was under the impression that

(15) Clifford has claimed that

(16) most anthropologists believe that

(17) the statement is false that:

(18) anthropology is a science. [and]

(19) quantum physicists have demonstrated that:

(20) science is only an observer-dependent set of arbitrary subjective opinions.

Extensive thinking about a topic can produce structures too elaborate to be placed, in their entirety, into the workspace, and which are therefore considered in pieces. The cultural development of memory aids such as writing have allowed an explosion of conceptual structures that are larger than what our ancestors would have routinely used.

Scope operators greatly augment the computational power of the human cognitive architecture compared to ancestral systems lacking such features. One advantage of an architecture equipped with scope operators is that it can carry out inferential operations on systems of inferences of unevaluated or suspended truth value, while keeping their computational products isolated from other knowledge stores until the truth or utility of the elements can be decided. If they were not kept isolated, their contents would enter into inferences with other data-structures in the architecture, often producing dangerously false but unmarked conclusions (e.g., *science is only a set of arbitrary subjective opinions* would be disastrous guidance for someone who has to choose a medical strategy to arrest an epidemic in a developing country). Fortunately, (14) decouples the uncertain information in (15)–(20) from the rest of the architecture, but allows the information to be maintained, and reasoned about, within various lawful and useful restrictions specified in the scope operators. The structure (14)–(20) is free to migrate through the system as a bound unit, entering into whatever licensed inferences it can be related to, but its subordinate elements are not.

Within subordinate levels (15)–(20), similar scope operations structure the inferences that are possible. The operator "demonstrate" assigns the value "true" to the subordinate element (20: *science is only* . . .), allowing its contents to be promoted to the next level. Within that level, it is treated as true, although it is not true above that level or outside of its scope-circumscription. The operator that governs that level – "claim" –

prevents it from migrating independently of the metarepresentation it is bound to (*Clifford has claimed that* . . .). Both (16) plus entailments and (19) plus entailments are true within the world of Clifford's claims, and are free to inferentially interact with each other, along with (20), as well as with any other of Clifford's claims that turn up. Indeed, one can say that a representation is true with respect to a particular level in a particular data-structure; any level can function as a ground level to subordinate levels. A data-structure is scope-conditionally true when it is permitted by the architecture to interact with any other information held within the same or subordinate levels of that data-structure.

Source, Error Correction, and the Evolution of Communication

Different scope operators obviously have different regulatory properties, and hence different functions. *Claim*, *believe*, and *demonstrate*, for example, require source tags as arguments, as well as conveying additional information – i.e., publicly assert as true that *p*; privately treat as architecturally true that *p*; publicly establish the truth that *p*, respectively. Source tags are very useful, because often, with contingent information, one may not have direct evidence about its truth, but may acquire information about the reliability of a source. If the sources of pieces of information are maintained with the information, then subsequent information about the source can be used to change the assigned truth-status of the information either upwards or downwards. For example, one may not assign much credal value to what most anthropologists believe (16), or one may discover that Clifford in particular is highly unreliable (15), while having a solid set of precedents in which Chagnon's impressions (such as 14–20) have proven highly reliable, despite the fact that he himself is unwilling to evaluate his impressions as trustworthy. Sources may include not only people but also sources internal to the architecture, such as vision, episodic memory, a supposition, previous inference, and so on. Thus, humans can have the thought "My eyes are telling me one thing, while my reason is telling me another."

In general, our minds are full of conclusions without our having maintained the grounds or evidence that led us to think of them as true. For a massively inferential architecture like the human mind, each item can serve as input to many other inferential processes, whose outputs are inputs to others. To the extent that the s information is sourced, or its grounds and derivation are preserved in association with the data, then new data about the grounds can be used to correct or update its inferential descendants. To the extent that the information is not sourced or its process of inferential derivation is not preserved in association

with it, then it cannot be automatically corrected when the grounds for belief are corrected. Indeed, our minds are undoubtedly full of erroneous inferential products that were not corrected when their parent source information was updated, because they could no longer be connected with their derivation. Because source tags, and especially derivations, are costly to maintain, mechanisms should monitor for sufficient corroboration, consistency with architecturally true information, or certification by a trusted source: If or when a threshold is reached, the system should no longer expend resources to maintain source information, and it should fade. This is what makes trust so useful (one does not need to keep the cognitive overhead of scope-processing communication) but so dangerous (one cannot recover and correct all of the implanted misinformation). After all, what is important about an encyclopedia of (accurate) knowledge about the world is the facts themselves: not who told them to you, what their attitude towards them was, or when you learned them. Typically, once a fact is established to a sufficient degree of certainty, source, attitude, and time tags are lost (Sperber, 1985; Tulving, 1983; Shimamura, 1995). For example, most people cannot remember who told them that apples are edible or that plants photosynthesize.[14] Moreover, an encyclopedia is most useful when the facts can cross-reference one another, so that each can support inferences that may apply to others, thereby adding further, inferred facts to the body of knowledge (e.g., "Mercury is a poison"; "Tuna has high levels of mercury"; therefore "people who eat tuna are ingesting poison"). This means that truth conditions must not be suspended for facts in semantic memory, and the scope of application for any truth-preserving inference procedures must be relatively unrestricted within the encyclopedia, such that facts can "mate" promiscuously to produce new, inferred facts.

Source tagging, source monitoring, and the scope-limitation of information by person must have played a critical role in the evolution of human communication and culture. Evolutionary biologists have long been aware that different organisms will have conflicting fitness interests and that this poses problems for the evolution of communication (Krebs & Dawkins, 1984). Information transmitted from other organisms will only be designed to be transmitted if it is in their interests, which opens up the possibility that each signal may be either deceptive or simply erroneous. The capacity to receive and process communication could not evolve if the interpretive process simply treated the communicated information as architecturally true, or unmarked, because deceptive exploitation would reduce the signal value to zero in most cases (see Sperber, this volume, for an analysis of how this adaptive problem may have led to the emergence of logical abilities deployed in the context of communication). The same argument holds true for culture-learning adaptations as well. Culture could not have evolved without the co-evolution

of a representational immune system to keep the acquirer from adopting too many false or exploitive cultural elements (Sperber, 1985; Tooby & Cosmides, 1989). Source tagging and scope syntax are crucial to this process. Take, for example:

(21) Fagles argues that

(22) Homer says that

(23) Odysseus told Achilles that

(24) he ought to be happy among the dead.

This structure uses communicative terms that attribute representations to sources, and that in so doing, clearly suspends their truth relations. This is just what one would expect of a scope-syntactic system that is well-designed for processing communication, while not being at the mercy of erroneous or deceptive messages.

Gerrig & Prentice (1991) have provided some empirical support for the notion that representations that are inconsistent with present knowledge are decoupled from representations that are consistent with it. After having read a story that contained statements like "Most forms of mental illness are contagious" subjects were asked to judge the truth "in the real world" of certain target statements. Regardless of retrieval context, they were faster at judging the inconsistent statements than the consistent ones, indicating that inconsistent ideas were stored separately from semantic memory. Judgments were even faster when the retrieval context suggested that the questions asked would be drawn from the story they had heard, lending some support to the idea that inconsistent information retains a source tag (in this case, the story-telling experimenter) that can be used for rapid retrieval.

Even more basically, Sperber has persuasively argued that the inferential nature of communication itself requires the on-line metarepresentational processing of language in order for interpretation to be successful (Sperber & Wilson, 1986; Sperber, 1985; 1996; this volume). Sperber (1985) has also proposed that metarepresentations as a data-format may be an adaptation to pedagogy, to deal with the problems posed by the long-term maintenance of information that is only partially comprehended.

Since Frege, philosophers have been aware that propositional attitudes suspend semantic relations such as truth, reference, and existence (Frege, 1892; Kripke, 1979; Richard, 1990). Frege noticed, for example, that the principle of substitution of co-referring terms breaks down when they are embedded in propositional attitudes (i.e., one can believe that Batman fights crime without believing that Bruce Wayne fights crime). Or, consider the statement:

(25) Shirley MacLaine believes that

(26) she is the reincarnation of an Egyptian princess named Nefu.

This can be true without Nefu ever having existed and without it being true that Shirley is her reincarnation. The propositional attitude verb *believe* suspends truth, reference, and existence in (26), fortunately decoupling (26) from the semantic memory of those who entertain this statement. However, rather than being quirks, problems, and puzzles, as philosophers have often regarded them, it seems clear that such suspensions are instead adaptations – design features of a computational architecture designed to solve the problems posed by the many varieties of contingent information exploited by our ancestors and by the interrelationships among sets of contingent information. Humans perform these operations effortlessly and easily acquire words and grammatical forms that correspond to various operators implementing these procedures. Indeed, it seems likely that these features are species-typical, reliably developing features of the human cognitive architecture, because it seems very difficult to conceive how they could plausibly be learned (in the domain-general, general-purpose sense).[15]

Development, Decoupling, and the Organizational Domain of Adaptations

Decoupling and scope syntax also offers insight into some aspects of how cognitive adaptations develop. Genes underlying adaptations are selected so that, in development, genes and specific, stable aspects of the world interact to cause the reliable development of a well-designed adaptation (Tooby & Cosmides, 1992). This means that information and structure necessary for the proper development of an adaptation may be stored in the world as well as in the genome, and that selection will shape developmental programs to exploit enduring features of the world. This allows adaptations to be far more elaborate than could be managed if all of the necessary information had to be supplied by the genome. What is likely to be genetically specified in adaptations is an economical kernel of elements that guides the construction and initialization of the machinery through targeted interactions with specific structures, situations, or stimuli in the world. This means that aesthetic motivations may be a necessary guidance system for the development of each adaptation – that is, motivations to detect, seek, and experience certain aspects of the world may be evolved design features, present to help adaptations become organized into their mature form. Consequently, a computational system may operate not just to perform its proper function on-line (e.g., the visual system performing useful scene analysis, the language system generating utterances for communicative purposes), but may operate in an organizational mode as well, designed to develop a better organization for carrying out its function (e.g., looking at sunsets to calibrate the visual system; babbling or speaking in order to develop a more effective language system). Thus, one might want to distinguish, in addition to the proper, actual, and cultural domains of an adaptation, what one might call its organizational domain, which consists of the conditions of operation for the adaptation that serve to organize the adaptation. Thus, a hunter-gatherer child might throw rocks at randomly chosen targets, developing her projectile skills outside of the context of hunting. On this view, aesthetics are aspects of the evolved components of the adaptation, designed to organize the adaptation in preparation for the performance of its function.

Now, much of the time, an adaptation may be improving its efficacy while it is performing its function in the actual situation for which the adaptation was designed, but the presence of scope and decoupling syntax offers the possibility of broadening the contexts of organization. Through scope syntax and other design features, activities that organize an adaptation can be liberated from the constraints of having to encounter the actual task, which may be very limited, dangerous, or simply not contain the informative feedback or revelatory data necessary by the time the organism needs the adaptation to be functioning well. For example, playing tag may develop flight skills that could not be advantageously developed purely in the context of actual instances of escape from a predator. The emancipation of the organizational domain from the proper domain of an adaptation can take place, if there is an abstract isomorphism between elements in the organizing experience and elements in the adaptive task, and if there are adaptations that can

(a) detect activities embodying this isomorphism;

(b) extract the organizing information present in them, and

(c) decouple the aspects of the organizational domain that are irrelevant or noncongruent from being processed by the adaptation as true or relevant for its development (e.g., although my father chases me, my father is not a predator with respect to me).

This last element is crucial: Not all parts of the experience are registered or stored, and the ability to decouple the processing of some inputs while preserving others is essential to the functioning of such a system.

It is important to recognize that this isomorphism can be very abstract and decontextualized, making some aesthetically driven activities seem very bizarre and nonfunctional when, in fact, they may have evolved to promote computational development. Because humans have

many more computational adaptations, which require data-based elaborations from the world to fuel them, one might expect aesthetics to play a far larger role in human life than it does in the life of other species. Humans, being social and communicative organisms, can greatly increase their rate of computational development because individuals are no longer limited by the flow of actual experience, which is slow and erratic in comparison with the rapid rate of vicarious, contrived, or imagined experience. So, vicarious experience, communicated from others, should be aesthetically rewarding. But what could possibly be useful about fictive, counterfactual, or imagined worlds – that is, about false or indeterminate information? We will return to the case of fiction at the end of the paper.

Theory of Mind and the Prediction of Behavior

One domain of critical importance to the success of organisms is understanding the minds of other organisms, such as conspecifics, predators, and prey, and it is plausible that humans have evolved computational specializations for this purpose. There is now considerable evidence that the human cognitive architecture contains computational machinery that is designed to infer the mental states of other people – their beliefs, desires, emotions, and intentions – and to use these to predict and explain their behavior (for a review, see Baron-Cohen, 1995). This machinery produces the *intentional stance* (Dennett, 1987), a mode of causal explanation based on mental states. For example, in answer to the question, "Why did Nike open the box?" most people over the age of three would consider "Because she *wanted* chocolate and *believed* there was chocolate in the box" a full and adequate response, even though Nike's mental states – her beliefs and desires – are the only causes mentioned.[16]

Designing a computational device that can predict behavior on the basis of beliefs presents certain problems: Not only does the machine need to infer the content of propositions in another person's head, but it needs to remember which person's head the proposition is in, what that person's attitude toward the proposition is (does the person *believe* X, *doubt* X, *imagine* X?), and when the person had that attitude. At the same time, it is important that the organism's *own* behavior be based on true beliefs about the world. This will not happen if other people's beliefs (a mixture of true and false propositions) are stored as "true." So the architecture needs to file memories specifying the content of other people's beliefs separately from its own mental encyclopedia of facts about the world.

Carefully noting these computational requirements, Leslie and his colleagues have proposed that the content of other people's beliefs – that is, the content of a (potentially counterfactual) proposition – is embedded

in a special kind of data format, the M-representation (a kind of metarepresentation) (Leslie, 1987; Frith, Morton, & Leslie, 1991; Leslie & Frith, 1990). M-representations are a particular type of scope-representation that evolved specifically for the purpose of modeling other minds. The M-representation has a number of design features that solve the problems listed above, thereby making it particularly useful for understanding and predicting an agent's behavior. These features are as follows.

(a) An *agent slot*. This is a variable that represents *who* it is that believes (doubts, imagines, etc.) that X. In Leslie's example "Mommy is pretending that the banana is a telephone," "Mommy" fills the agent slot. In locating this in the broader landscape of scope-representations, we would say that the *agent slot* is one form of *source tag*. The specific arguments required for a scope representation obviously depend on the specific kind of scope representation (some require a source tag; some require that the source tag be an agent, etc.).

(b) An *attitude slot*. This variable specifies the attitude that the source (the agent) has to the information represented in X: whether the agent *is pretending* that X, *believes* X, *doubts* X, *imagines* X, and so on. For scope-representations, this corresponds to the *relationship slot*, which defines the relationship (and scope-restrictions) between two or more sets of representations.

(c) An *anchor*. In the case of pretense and beliefs (and perhaps other attitudes), there is an anchor: a primary representation (i.e., a representation of a real entity or state of the world) to which the embedded proposition refers. A fuller version of an M-representation in which Mommy's act of pretense could be stored would be [Mommy]–[is pretending (of the banana)]–[that it is a telephone]. The anchor is "the banana": It is the primary representation to which the decoupled proposition, "it is a telephone," refers. Different scope-operators take different numbers of arguments: In this case of pretense, there are two ground state representations, "Mommy" and the "banana," related to the decoupled proposition.

(d) A *proposition slot*. This is where the content of the belief or desire is stored ("It is a telephone".) For scope-representations, the proposition slot can include any number of propositions, and potentially any number of levels.

To this, we would also add

(e) A *time tag*. There must be a tag specifying *when* the agent held the attitude toward the proposition. After all, "Nike believes X," "Nike used to believe X," and "(after she sees Y) Nike will believe X" all specify different mental states.

In addition to these formatting properties, an M-representation has several other closely related features, which are also necessary if an organism is to represent (potentially false) beliefs yet still behave adaptively.

(a) *Suspending semantic relations.* Propositions stand in certain relationships to one another, such as contradiction, equivalence, or mutual consistency. For example, "the chocolate is in the box" implies certain other propositions, such as (1) "there is a chocolate" (*existence*); (2) "the chocolate is not outside the box" (*truth*); and (3) (if the chocolate being referred to is a Toblerone), "a Toblerone is in the box" (*reference*). Embedding the same proposition within the agent-attitude-proposition format of an M-representation takes that proposition out of circulation by suspending its normal truth relations. For example, "Nike *believes* the chocolate is in the box" can be a true statement, even if, unbeknownst to Nike, (1) someone has already eaten the chocolate (i.e., it no longer exists; *existence* relations suspended); and (2) the chocolate is not in the box (*truth* relations suspended). Moreover, if Nike does not realize the chocolate at issue is a Toblerone, she could simultaneously believe "the chocolate is in the box" and "there is no Toblerone in the box" (i.e., substituting "Toblerone" for "chocolate" is no longer truth preserving: *reference* relations suspended). As these situations show, to make adaptive choices, the system needs simultaneously to represent two parallel tracks: the actual state of the world versus Nike's beliefs about the world. Suspending truth relations for beliefs is necessary if both tracks are to be represented accurately.

(b) *Decoupling.* By virtue of being embedded in an M-representation, a proposition is "decoupled" from semantic memory. That is, it is not stored as "true." For example, the child who represents her mother as *pretending* that the banana is a telephone does not store as true "the banana is a telephone." As a result, she does not become confused about the properties of bananas or telephones.

(c) *Restricted application of inferences.* As Leslie and Frith note in the case of pretense, "Decoupling creates an extra level within the representation ... [Inference mechanisms] respect the levels and apply to them one at a time" (Leslie & Frith, 1990, p. 129). For example, they point out that (27) "The cup is full" and (28) "I pretend the cup is full" are both sensible propositions, whereas (29) *"The empty cup is full" involves a contradiction and (30) *"I pretend the cup is both empty and full" is strange. This is because the M-representation has a superordinate level and a subordinate level, which they call the upstairs and downstairs levels. So for (28), which translates into mentalese as "I pretend (of the empty cup) [it is full]," logical inference mechanisms cannot detect a contradiction at either the upstairs level – "I pretend (of the empty cup) [X]" or at the downstairs level – [it is full]. For sentence (30) – "I pretend (of the cup) [it is both empty and full]" – no contradiction is detected at the superordinate level ("I pretend (of the cup) [X]"), but a contradiction is detected at the subordinate level ([it is both empty and full]).

Note that none of these design features are necessary for propositions stored in semantic memory. However, as discussed earlier, all three of these properties are widespread, basic features of scope syntax, appearing in many system components, including, but not limited to, theory of mind contexts. We wish to particularly emphasize the importance of the restricted application of inferences, which is a crucial property of scope representations, as in the supposition processing outlined above. We want to underline that it is not an oddity or byproduct of either pretense or of ToMM, but is a core set of computational adaptations essential to modeling the minds of others accurately. When applied solely at the subordinate level, valid inferences can be made about other beliefs the agent holds at that level. For example, if Nike *believes* "the chocolate is in the box," then she also believes "the chocolate is not outside the box" and "there is a chocolate." These inferences about Nike's *beliefs* hold even if the chocolate is gone, that is, even if the premise ("the chocolate is in the box") is false. When applied solely to the superordinate or ground level, valid inferences can be made about the agent because semantic relations (reference, existence, truth) are suspended only for the embedded proposition, not for the scope-representation as a whole. For example, because an M-representation is itself a proposition, reference/identity relations allow substitution inferences, such as "Nike believes something," or (pointing to Nike) "That girl believes there is a chocolate in the box," or "Leda and John's daughter believes there is a chocolate in the box." In other words, the full power of whatever parts of propositional logic are implemented in the human mind can be brought to bear as long as the levels are kept separate for the purposes of inference making.

Beyond Modeling Other Minds

Predicting the behavior of other people is a critical adaptive problem for humans, and some scholars have proposed that mind-reading was the adaptive problem that drove the emergence of the distinctively human form of intelligence. We think this is very plausible, but far from certain, because mind-reading is not the only adaptive problem that poses computational requirements involving scope syntax. Many abilities critical to

the cognitive niche require representations with scope-processing properties. We think that M-representations are one particular and important form of scope-representation, built out of various elements of scope-syntax. However, scope-representations of various permutations, often sharing many properties with M-representations, appear to play a key role in a wide variety of cognitive processes that create the distinctive form of intelligence one finds in our species. This includes our ability to engage in long chains of suppositional reasoning; our practical ability to craft tools that take advantage of facts that are true only contingently, rather than universally; and our ability to remember a personal past.

On this view, there is a close relationship – both conceptual and empirical – between decoupling, source monitoring, specifying an agent's attitude, and memory tags specifying source, time, and place. They are all key features of a scope syntax, required by many different cognitive niche abilities. To illustrate how they cluster – and some of their permutations – we shall consider a variety of different types of representation, and ask the following about each in turn:

(a) Does the representation need to be decoupled from semantic memory to prevent the corruption of data structures (i.e., to prevent what Leslie (1987) calls *representational abuse*)?

(b) Is it necessary to monitor where the representation originated (its source)?

(c) Is it necessary to store an agent's attitude toward the representation?

(d) When stored in memory, does the representation need a source tag? a time tag? a place tag?

We think that the answer to these questions is "yes" for a number of adaptive information-processing problems beyond modeling other people's beliefs. Some other adaptive information-processing problems that require the same kind of computational solution include certain kinds of goals and plans (Frith, 1992); simulations of the physical world; pedagogy (Sperber, 1985); episodic memories;[17] simulations of social interactions that have not yet happened; understanding story-telling; representing one's own beliefs when these are not yet confirmed; and representing one's own beliefs when their truth is in question. (Dreams pose similar, though different, problems, and therefore present an interesting contrast in which decoupling is accomplished via the volatility and purging of the memory trace.)

We propose that scope-representations (S-representations) are involved in each of these activities, and briefly review evidence that implicates them. There are obviously many different species of S-representation – e.g., S-representations designed for simulating the physical

world may differ in certain respects from those designed for simulating social interactions. But evolution is a conservative process. Once a design that satisfies a particular set of computational requirements exists, natural selection can engineer solutions to new adaptive information-processing problems that pose similar requirements more quickly by modifying the existing design than by creating totally new designs from scratch. Consequently, even if there are different species of scope-representation, we expect that they will share certain important properties, and perhaps even share certain computational components. Indeed, Christopher Frith has proposed that there is computational machinery common to all metarepresentations, and that this common machinery is selectively damaged in schizophrenia, explaining many of this disorder's otherwise puzzling symptoms and signs (Frith, 1992; Frith and Frith 1991). Parts of our argument were inspired by his observations and theoretical analyses. One way that we will test the adequacy of our own view of the role of scope-representations is by applying it to episodic memory. This application yields testable predictions about which memory systems should be impaired in schizophrenia.

Representations of Goals

In this section, we consider representations of goals. Obviously, not all behavior that looks goal-directed involves representations of goals. For example, ticks have a circuit directly linking chemoreceptors to motor neurons, so that the smell of butyric acid causes the tick to drop from a tree. Because butyric acid is emitted only by mammals, this circuit usually results in the tick landing on a mammalian host, whose blood it then drinks. The design of this circuit makes the tick's behavior functionally goal-directed. Yet it involves no explicit representation of a goal state.

In addition to embodying circuits that only appear, by virtue of their design, to be goal-driven, the human cognitive architecture is also capable of representing goal states – such as "I want to have dinner at Downey's, on State Street" – and then devising plans to achieve these goals. The adaptive function of such representations is to regulate one's own behavior – an adaptive function different from Baron-Cohen's Intentionality Detector (1995) or Leslie's ToMM System 1 (1994), which are designed to infer other people's goals for the purpose of predicting and explaining their behavior. As a result, there are many differences in design. For example, whereas the Intentionality Detector infers goals on the basis of external cues, such as self-propelled motion or eye direction, individuals can formulate goals of their own without having to infer them on the basis of observations of their own behavior. Nevertheless, the ability to represent one's own goals – and remember them – while still

engaging in adaptive behavior poses a number of computational requirements that are similar to those for representing other people's beliefs.

What Are the Computational Requirements?

(1) Decoupling

The goal represents a state of the world that is not yet true of the world. Without decoupling, goals would be stored as true states of the world. Indeed, we sometimes find ourselves confused as to whether we did something or only entertained it as a goal. These are cases when the decoupling of the representation has failed.

(2) Source Tag

When I look out the window and see the ocean, the source of that representation is assigned by my mind to "the outer world" (see *Simulations*, below). But a goal representation cannot have the outer world as a source: Goals cannot be observed in the environment because they are not (by definition) states of the world that have already occurred. More importantly, only agents are capable of having goals, and the agent – the source of the goal representation – needs to be specified. The source of a goal representation is either (a) my own mind; or (b) someone else's mind. Moreover, if the source is someone else's mind, (c) was it the mind of Person A or Person B? Only goal representations whose ground level has a "self" source tag ($goal_{self}$) should be readable by mechanisms for planning and producing one's own motor responses. (Obviously, we are capable of taking other people's goals into account in formulating our own; hence goal representations with an "other" source tag – $goal_{other}$ – must be readable by those systems that formulate own-goal representations). We expect, however, that there is an important distinction to be made between an implicit source tag of self, computationally present because of the representation's location in a motivational specialization, and an explicit representation in a format common to other representations of social agency.

If the source of the goal were not specified, delusions would ensue (Frith, 1992). If the "self" source tag were lost, the content of the goal would escape its normal scope-processing tag, perhaps being experienced as an order. "I want to [kiss that girl]" would be experienced as "Kiss that girl". If the "other" source tag were lost, the same thing would happen: "He wants to [kiss that girl]" would be experienced as a voiceless order, "Kiss that girl" or might reacquire an implicit source tag. If source tags were switched, *He wants to kiss that girl* might be remembered as *I want to kiss that girl*. As Frith points out, all of these things can happen in schizophrenia (Frith, 1992, p. 127). In schizophrenia, source monitoring is impaired (Frith, 1992), although it is not clear whether the malfunction involves the machinery that reads source tags or the adhesion of the source tags themselves. Aberrations in source monitoring also appear to occur under hypnosis: the hypnotist suggests that certain goals originate with the subject.[18]

(3) Attitude Slot

For the representation to be useful, the agent's attitude toward the goal needs to be specified. The agent may *consider* (the goal), *want* (the goal to be realized), *intend* (to cause the goal to be realized), *decide to drop* (the goal), and so on.

(4) Memory Requirements

Storing a goal that has not yet been realized may be thought of as remembering the future (Ingvar, 1985) or, more precisely, remembering a possible (or subjunctive) future. A goal representation needs a time tag (e.g., "I would like *X* to happen"), so that any system that reads this representation can tell whether the goal has been realized yet, and modify its own functioning accordingly. For example, a planning system should only take $goals_{self}$ with a "future" tag on them as input; if it registered a past tag on the goal, presumably the planning system would abort operations based on that goal. Naturally, the source and attitude must also be remembered, in addition to the content. As time passes, one would expect goals that have already been realized to eventually lose their source tags and be stored in semantic memory as states of the world (unless the source of the goal is somehow relevant to one's social situation; see section on episodic memory).

(5) Restricted Scope of Inferences

Suspension of truth relations is necessary for the content of a $goal_{self}$ or a $goal_{other}$. The reason is related to that for beliefs but differs somewhat. In the case of beliefs, the suspension of truth relations is necessary because it is possible that the content of the belief is false: you need to be able to make accurate inferences about what other beliefs the person might hold, even if these are premised upon a belief you know to be false. Unlike a belief, however, a goal cannot, strictly speaking, be false. In the case of a goal – whether it is a $goal_{self}$ or a $goal_{other}$ – suspension of truth relations is necessary because the content of the goal specifies a state of affairs that *does not yet exist*. It specifies a possible world. Goal representations need a subordinate and superordinate level as well, so that inferences are restricted in their scope of application to only one level at a time. When applied subordinately, to the content of the goal itself, this allows one to reason suppositionally about possible worlds, and to make inferences about other goals, plans, and intentions that one might formulate. When applied superordinately, to a representation of the agent

who has an attitude toward the goal, this allows one to make inferences about the real world (e.g., "Nike wants *X*" implies "John and Leda's daughter wants *X*").

(6) Relationship to Other Representation Systems

One would expect goal representations to be read by (1) simulation systems, (2) planning systems, and (3) a self-monitoring system. (The latter is a system designed to detect certain inner states, creating representations of these states that can then be acted on by other inferential systems, as when one reflects on one's own goals, plans, intentions, and thoughts; see Johnson, Hashtroudi, & Lindsay, 1993; Frith, 1992.) The machinery that formulates $goal_{self}$ representations must take input from motivational systems and from the planning system (formulating plans often requires the formulation of sub-goals); it must also be able to access semantic memory (because facts about the world are relevant to deciding on goals).

What Is the Solution?

One way to satisfy these requirements is to store the content of a goal in a scope-representation similar to an M-representation. The goal S-representation is decoupled from semantic memory, tagged with a source, attitude, and time scope, suspends truth relations for the goal content (but not reference), and has a ground and subordinate level such that inferences are applied to only one level at a time.

Representations of Plans

What Are the Computational Requirements?

(1) Decoupling

A plan represents a sequence of actions that can achieve a goal (or, more precisely, were chosen because it is believed that they can achieve a goal). The goal is not yet true of the world. The sequence of actions has not (yet been) carried out. Without decoupling, a plan could be stored as actions already carried out. Again, a prospect would be stored in semantic memory as a reality.

The issues relating to the source tag, the attitude slot, the memory requirements, and the restricted scope requirements closely parallel those for goal representations. Of course, suspension of truth relations is necessary for the content of a plan because it is a sequence of actions that has not yet occurred, and may not occur. Like goals, plan representations need superordinate and subordinate levels, so that inferences are

restricted in their scope of application to only one level at a time. When applied downstairs, to the content of the plan itself, this allows one to reason suppositionally about possible chains of actions, and to make inferences about other goals, plans, and intentions that the agent might have. When applied upstairs, to a representation of the agent who has an attitude toward the plan, this allows one to make inferences about the real world (again, "Nike plans to *X*" implies "John and Leda's daughter plans to *X*"). Moreover, each necessary step in the plan creates another suppositional subdomain (*if x, then y; if y, then z*), whose execution must be completed before its contents are discharged, and the next step of the plan can be promoted and executed.

(2) Relationship to Other Representation Systems

A plan representation must be linked to the motor system in two ways. First, the plan needs linkages that generate the requisite sequence of motor actions – plans cause willed actions. Second, the plan must suppress the stimulus-driven action system (Frith, 1992), as discussed below. Because plan representations are formed in order to realize goals, they must also be linked to goal representations, and to the factual database of semantic memory. They should, in addition, be linked to simulation systems (see below).

What Is the Solution?

These computational requirements are almost identical to those for a goal representation. They can be satisfied by storing the content of the plan in a scope-representation specialized for the task: a representation that is decoupled from semantic memory, tagged with a source, attitude, and time scope, suspends truth relations for the plan's content, and has hierarchical levels such that inferences are applied to only one level at a time, or actions are taken in proper sequence.

The fact that plans must be linked to the motor system creates additional functional requirements that have interesting consequences. It is obvious that there must be ways to transform plans into actions. What is less obvious is that there needs to be a system whereby a plan scope-representation can inhibit stimulus-driven actions.

Based on neuropsychological evidence from a number of different disorders, Frith argues that the human cognitive system is constructed such that "there are two major sources of action. Some actions are carried out directly in response to environmental stimuli. Others are seemingly spontaneous and self-initiated" (Frith, 1992, p. 43). A stimulus driven action originates in perception: "perception → stimulus intention → action → response." (Frith, 1992, p. 46). A willed action originates in goal and plan representations: "goals/plans → willed action → responses"

(Frith, 1992, p. 46). This is supported by various disconnection syndromes. In Parkinson's disease, for example, plans and willed intentions are formed, but they do not generate action representations. In the negative cycle of schizophrenia, a person may be able to engage in stimulus-driven actions, but has difficulty translating a plan representation into a willed intention, resulting in poverty of action or perseveration.[19] In the positive cycle, a person may have difficulty inhibiting stimulus-driven actions, resulting in incoherence of action.

The system that allows plan representations to *inhibit* stimulus-driven actions can be neurologically compromised by any number of disorders and conditions: (1) frontal lobe damage (see Duncan, 1995, on "goal neglect," in which the goal is remembered but behavior is driven by external stimuli rather than by a representation of a plan that would achieve the goal; see also Shimamura, 1995, on inhibitory gating); (2) damage to anterior cingulate gyrus (a frontal structure; Posner & Raichle, 1994; Devinsky, Morrell, & Vogt, 1995; for supporting Position Emission Tomography (PET) results, see Pardo, Pardo, Janer, & Raichle, 1990); (3) conditions in which the stimulus-driven action system is intact, yet difficult to override, such as schizophrenia (Frith, 1992); and, possibly, (4) Tourrette's syndrome (Baron-Cohen, Robertson, & Moriarty, 1994; Baron-Cohen, Cross, Crowson, & Robertson, 1994). The ability to *construct* plans can be impaired by (1) frontal lobe damage (Frith, 1992); (2) anterior cingulate damage (Devinsky, Morrell, & Vogt, 1995), and (3) schizophrenia (Frith, 1992). The ability to *carry out* plans can be impaired by (1) frontal lobe damage, particularly when there is cingulate damage, as in akinesis and mutism (Damasio & Van Hoesen, 1983; Devinsky, Morrell, & Vogt, 1995; Duncan, 1995; Frith, 1992); (2) Parkinson's disease (Frith, 1992; Goldberg, 1985); (3) hypnosis, which creates a dissociation between plans and actions (Bowers, 1977; Hilgard, 1977); and (4) depression. It is interesting that a number of these conditions involve improper levels of dopamine (Parkinson's, schizophrenia, Tourrette's and, sometimes, depression).[20]

Information about how a plan representation can be neurologically compromised is relevant to our argument that scope-representations are involved in a number of cognitive processes. As you will see, other cognitive-niche abilities can be compromised by damage to the same brain regions and by the same syndromes. This is what one would expect if scope-representations were involved not only in beliefs, but in goals, plans, simulations, episodic memory, and so on. It suggests that some of these brain regions may be evolutionarily more recent adaptations to the cognitive niche, and that they are damaged more easily than other cognitive systems because the shorter evolutionary time-depth means less time in which selection could operate to debug them.

Representations of *Simulations* of the Physical World

In his William James Lectures, Roger Shepard posed the question: Why have thought experiments been so fruitful in physics? Why should our ability to imagine the world ever generate knowledge that corresponds to reality? (Shepard, 1994). Through experiments on apparent motion, "mental rotation," and related phenomena, Shepard and his colleagues have shown that representations of the movement of objects are constrained by procedures that reflect evolutionarily long-enduring properties of the world – even when these representations occur in the absence of an external stimulus. Consequently, this system represents translations and rotations that are, in many ways, functionally isomorphic to the translations and rotations of rigid objects through three-dimensional space (e.g., Shepard, 1984; 1987).

In other words, the mental models it produces reflect the world with some accuracy. That is why thought experiments can be useful: We have an analog representational system for simulating the movements of real world objects and the "motion" of these imagined objects is constrained in the same ways as the motion of real objects (Shepard 1984, 1987). Shepard calls these simulations "internally driven hallucinations," to contrast them with perceptual representations, which he calls "externally driven hallucinations." Both are constructed using a great deal of inferential machinery (hence "hallucinations," to remind one that the world is never perceived absent an inferential construction). The main difference between them is that perceptions are prompted by the world external to a person's mind whereas simulations are prompted by other internal representations.

Other researchers have focused on how infants (and adults) represent perceived objects and their movement – externally driven hallucinations – and have shown that the ways in which people conceive of objects and model their interactions is governed by a rich set of interlocking principles, which Leslie has dubbed a "theory of bodies" (ToBy) (e.g., Baillergeon, 1986; Leslie, 1988; 1994; Shepard, 1984; Spelke, 1988; 1990; Talmy, 1988). The conclusions of this work dovetail quite closely with Shepard's work on mental simulations, suggesting that ToBy provides constraints on both perceptual representations and mental simulations of the physical world (Brase, Cosmides, & Tooby, 1996).

In our view, this simulation system did indeed evolve to do physical thought experiments: those that might help a tool-using and environment-manipulating primate to imagine ways in which new tools can be developed, existing tools can be applied to specific situations, and environments can be physically modified. Simulations provide a way

of forecasting how physical objects will interact before they actually do. Time may be lacking for a series of dry runs and, in any case, time, materials, and energy can be saved by doing mental experiments prior to physical experiments. Simulations may also allow one to avoid disastrous situations (such as being in the path of an impending rock slide) or take advantage of fortuitous ones (such as cleaning something by leaving it outside during a rainstorm).

From this standpoint, the fact that simulations of objects and their movement are constrained in the same way as real objects is critical: the thought experiments would be useless otherwise. Nevertheless, one would expect simulation systems to have certain properties that perceptual representations lack.

What Are the Computational Requirements?

(1) Decoupling

A simulation represents the ways in which objects can interact physically. These physical interactions have not happened in the real world. Without decoupling, a simulation could be stored as something that happened.

(2) Source Tag

Simulations are not externally derived through perception: They do not represent actual states of the world. They have an internal source, the mind of the agent who is doing the simulation. It is an internally driven "hallucination" (Shepard, 1984) and, therefore, needs a source tag to keep it identified once it is output from the simulation system.

(3) Credal Value and Memory Requirements

Storing a simulation is equivalent to remembering a potentiality. Simulations may be associated with different levels of certainty: a simple interaction among objects (e.g., that one billiard ball will launch another after hitting it) might be tagged with a higher degree of certainty than a complex one (e.g., that hitting the corner of one billiard ball will put enough spin on it to make it bounce off a side wall and hit two balls at the other end of the table, knocking one into the corner pocket). A simulation tagged with a high level of certainty might be marked as "timeless," rather than having a past, present, or future tag. After all, the laws of physics do not change over time; something that is true of the physical world now – e.g., that a sharp edge can be struck from flint – will always be true.

(4) Restricted Scope of Inferences

Simulation representations depict hypothetical transformations of objects in space and time and, thus, sequential transformations are suppositions with ordered hierarchical relations, describing states of the system at various points in time.

(5) Relationship to Other Representation Systems

For a simulation to occur, stored object representations must be retrieved – presumably from the Perceptual-Representational System (PRS) (Schacter, 1995) and perhaps from semantic memory – and placed into the simulation buffer. This buffer would be a form of working memory. The system that retrieves the object representations would control what interacts with what (and when) during the simulation, but not how the interaction proceeds. That would be governed by ToBy. The output of the simulation system would inform the planning system (e.g., on how to make a tool) as well as the goal system (creating, e.g., the realization that the goal of making a particular tool is feasible). It can also inform the motor system, allowing one to anticipate the future (e.g., to jump out of the way or to strike a stone from one angle rather than another.)

What Is the Solution?

The simulation must be conducted in a buffer that is decoupled from semantic memory. Something resembling an M-representation – [self]-[wants to know]-(about objects *X, Y, Z*)-[how they will interact] – might govern what gets simulated – i.e., it would retrieve appropriate object representations from the PRS and semantic memory systems and deposit them in the decoupled buffer. ToBy would conduct the simulation. The output would also be tagged with a degree of certainty.

Working memory, in the sense that Baddeley (1995) uses the term, would appear to meet these requirements. The *visuospatial sketchpad*, which Baddeley describes as a "slave system" of working memory (the other slave system being a phonological loop), is the decoupled buffer in which simulations depicting objects moving in space and time are conducted. The executive controls the contents of the sketchpad, i.e., determines what objects are placed in the visuospatial sketchpad. Simulations in the sketchpad – such as mental rotation – are governed by ToBy, which causes the simulations to reflect long-enduring properties of the world (e.g., Shepard, 1984; Leslie, 1988; 1994). (Hegarty has also suggested that there is a kinesthetic sketchpad for simulating body motions; personal communication).

In this view, there are a number of ways in which the simulation system could be damaged.

(a) The integrity of the buffer itself could be compromised. Insofar as the buffer uses parts of the visual system to represent space, damage to the visual system will cause damage to the buffer (Farah,

Soso, & Dasheiff, 1992; Kosslyn, Alpert, Thompson, et al., 1993; Kosslyn, 1994). Processes of visual attention appear to be involved in the "inspection" of mental images; patients with right parietal lesions who show neglect of the left half of their visual field, also neglect the left half-space when they form visual images (Bisiach & Luzzatti, 1978).

(b) The operations of ToBy or related specializations could be compromised. A suggestive, though not definitive, case is that of L.H., an agnosic who is impaired in his ability to form visual images of living things but not of non-living things (Farah, 1990).

(c) The executive, which is hypothesized to be responsible for providing input to the simulation buffer, could be impaired. Frontal-lobe damage causes impairments to executive functions and can impair one's ability to make plans. It is possible that impairments in the executive's ability to place object representations into the simulation buffer may be responsible for some of these cases. Left hemisphere damage is also known to affect imagery-control processes (Farah, 1984).

(d) The decoupler – the system that keeps the contents of the simulation buffer separate from semantic memory – could be compromised. For example, patients who have had a cingulectomy (to control obsessive compulsive disorder [OCD]) confabulate: They have unusually vivid mental experiences, which are mistaken for perceptions (in other words, there is a misattribution of the source, which is experienced as external rather than internal). These patients use their powers of inference to correct these misattributions. For example, one such patient thought he had had tea with his wife, when in fact he had only imagined it: "I have been having tea with my wife ... Oh, I haven't really. She's not been here today ... The scene occurs so vividly, I can see the cups and saucers and hear her pouring out" (Whitty & Lewin, 1957, p. 73; see Johnson, Hashtroudi, & Lindsay (1993) for discussion).

There are, of course, social, biological, and psychological simulations as well as physical simulations and admixtures. Reports from patients who have had a cingulectomy, as well as normal experience, suggest that although various simulation processes may differ in what specializations they invoke to carry out specific inferential steps (ToBy, ToMM, or a theory of human nature, ToHN, etc.), they may share some common buffers and representational machinery. Nevertheless, the principles should be similar to what we have sketched above.

Representations of Integrated Simulations and Fiction

The emergence of pretend play in children at 18 months of age, and a consideration of its computational requirements, was the basis of Leslie's initial proposals about the existence, format, and properties of M-representations (Leslie 1987). Moreover, the virtual absence of imaginative activities and pretend play is one of three impairments used to diagnose autism (Frith, 1989). In spite of the universality of pretend play, the apparent presence of a mechanism designed to cause it, and the existence of a disorder that selectively impairs it, little attention has been paid to what its adaptive function might be. The same is true of general imaginative simulations. If these are written down or shared, we call them fiction. There are two distinct functional questions raised by these activities. First, what is the function of generating imagined sets of propositions? Second, what (if any) is the function of attending to such representations, when others express them to you? Of course, many have advocated the idea that simulations are useful models developed to guide behavior (Shepard, 1994), and this seems highly plausible and likely to be a major explanatory factor. However, people often imagine obviously and blatantly false situations, which could play no realistic role in planning. Pretense is also often extravagantly at variance with ordinary reality. Of course, these could be functionless byproducts or susceptibilities of a system for practical simulation and planning.

However, we think there is another hypothesis worth considering. As previously discussed, imaginative activities may have the evolved function of organizing and elaborating computational adaptations that require exposure to certain inputs in order to develop properly, and the aesthetic motivations governing which inputs are sought, the way they are processed, what elements are decoupled, and where the outputs are routed may all be governed by a suite of evolved design features selected to implement this function. If this is true, one can supply inputs for oneself through imagination or one can seek out social and cultural products, such as narratives, that provide rich inputs that help to organize adaptations such as ToMM, adaptations for social interaction, and so on, as well as motivational and emotional adaptations. Still, how can imaginary, false, or counterfactual inputs possibly be useful?

One way is in fleshing out the motivational system, in which imagery or stories become the occasions for releasing information bound in some formats so that it becomes available to other subsystems (Tooby & Cosmides, 1990a). The human mind contains decision and evaluation rules that initially evolved to be triggered by actual exposure to biologically meaningful situations, and which can therefore be elicited by cues indicating the presence of the situation (e.g., fangs, smiles, running

sores, sexual possibility, enemy ambush, death of a child). We have argued that imagery allows these ancestral systems to be tapped without the need to wait until one encounters the real situation, greatly augmenting the power of the motivation and planning functions (Tooby & Cosmides, 1990a). This would allow prospective courses of action to be evaluated using the same circuits as would be activated if the event were actually to occur. Using imagery and vicarious experience to evoke these systems (with appropriate decoupling) would provide motivational adaptations with a rich array of weightings for events. For example, imagining the death of your child can evoke something of the emotional state you would experience had this actually happened, activating previously dormant algorithms and making new information available to many different mechanisms. Even though you have never actually experienced the death of a child, for example, an imagined death may activate an image-based representation of extremely negative proprioceptive cues that instruct the planning function on the proper valence and intensity of motivational weightings. Instead of Aristotle's notion of catharsis, an alternate view would be that fictionally triggered emotion is a re-weighting process of the motivational system.

Story-telling is ubiquitous across cultures, and people are interested in stories even when they are told in advance that they are not true. Indeed, people are relatively indifferent to the truth of a narrative, compared to its other aesthetic qualities. This is consistent with the hypothesis that the evolved aesthetics guiding human attention are looking for input useful in organizing adaptations, but that they can recognize cues of useful inputs independent of the truth value of the total set of propositions involved. With socially supplied narratives, one is no longer limited by the flow of actual experience, slow and erratic compared to the rapid rate of vicarious, contrived, or imagined experience. "False" inputs incorporate many elements that are true or informative, and also provide occasions that activate procedures that build valuable higher-level structures (such as forms of social manipulation that one has not encountered first hand). "False" accounts may add to one's store of knowledge about possible social strategies, physical actions, and types of people, in a way that is better than true, accurate, but boring accounts of daily life. This does not mean that falsehoods are, other things being equal, preferred. True narratives about relevant people and situations – "urgent news" – will displace stories, until their information is assimilated, and the organizational demand of adaptations that need to be computationally fed in a decoupled fashion resumes.

Whether it is an adaptation or a byproduct of adaptations, our ability to think about fictional worlds poses a familiar set of computational problems.

What are the Computational Requirements?

(1) Decoupling

Stories may not be true – either in full or in part – even when the teller claims to be recounting events that actually happened. People tell stories for many different reasons other than to impart true information – including, sometimes, in an attempt to manipulate the hearer (Sugiyama, 1996). Thus, even stories that purport to be true ought to be treated with caution. But when a story is explicitly labeled as fictional, the propositions therein are false (by stipulation). Without decoupling, a fiction would be stored as a reality.

(2) Source Tag

Only an agent can be the source of a story, and the source of a representation of a fictional world needs to be specified if one is to avoid confusion and manipulation. It matters whether the source of a fiction is (a) my own mind, or (b) someone else's mind. Given that people have a wide variety of reasons for telling stories, a mechanism designed to treat information originating with other people with caution should also specify whether the source of the fiction is the mind of Person A or Person B.

(3) Attitude Slot

There is a difference between false beliefs and fiction. In the first case, the agent *believes* that the propositions recounted are true. In the second case, the agent *is providing* propositions regardless of their truth value. And, in fiction explicitly labeled as such, the teller *intends* that the hearer *believe* that the teller is providing false representations, but that the teller intends them to form a coherent narrative.

(4) Memory Requirements

Storing a story is similar to storing a false belief, but with different arguments in the attitude slot. In the case of a person who purported to be telling the truth but instead was knowingly telling a fiction, the time tags might be multiply embedded: "Kurt *was pretending* that [Story *X*] was true but he *believed* it was false and, now that we caught him, he *is no longer pretending* that it is true." Stories that turn out to be true may lose their scope-restrictions and their contents may be allowed to interact freely with encyclopedic knowledge. But, those explicitly labeled as fiction (e.g., *Little Red Riding Hood*) ought never to be retired without a source tag. The specificity of the source tag may be degraded, from (say) "*Mother* told me [story *X*]" to "*Someone* told me [story *X*]," but one would expect a fiction to remain decoupled and, insofar as source tags are an important part of a self-monitoring system (Frith, 1992; Johnson,

Hashtroudi, & Lindsay, 1993; Kunzendorf, 1985–1986), one would expect them to be retained in some form as well.

(5) Restricted Scope of Inferences

Suspension of truth relations is necessary for fictional worlds. This fact is directly reflected in the phrase "suspending disbelief," a state people recognize as essential to entering a fictional world. Representations of stories require hierarchical levels so that inferences are restricted in their scope of application to one level at a time. When applied to the content of the story itself, we are able to make all the ordinary implicit inferences necessary to understand the goals, intentions, beliefs, and motivations of the characters in the story. When applied superordinately to a representation of the agent who is telling the story, one can make inferences about the real world and the agent (e.g., "Mother dislikes wolves.")

(6) Relationship to Other Representation Systems

The falsity of a fictional world does not extend to all of the elements in it, and useful elements (e.g., Odysseus' exploits suggest that one can prevail against a stronger opponent by cultivating false beliefs) should be identified and routed to various adaptations and knowledge systems. One thing does seem likely: The fact that fiction can move people means that it can serve as input to whatever systems generate human emotions and motivation.

What Is the Solution?

These computational requirements are similar to those required for pretend play, which was Leslie's basis for developing his model of the M-representation. Consequently, one would expect the content of stories to be stored in an M-representational system: a representation that is decoupled from semantic memory, tagged with a source, attitude, and time scope, suspends truth relations for the story's content, and has ground and subordinate levels such that inferences are applied to only one level at a time. If the hypothesis that these activities function to organize adaptations in development is true, then there should be a set of associated aesthetic systems, isomorphism detectors, and output-routing systems as well (as described earlier in the paper, pp. 73–74).

Potts, St. John, & Kirson (1989) report memory-retrieval experiments suggesting that representations of fiction are indeed decoupled from representations in semantic memory. Each subject read a story about unfamiliar wildlife. Half of the group was told that the information in the story was all true and had been verified from reference books; the other half was told that most of the information in the story was fictional and had been made up for the purposes of conducting the experiment. When the retrieval context cued subjects that they were

being asked to retrieve information from a particular *source* – the story they had been told – subjects who had been told the story was fictional verified statements faster than those who had been told it was true. This makes sense if information one believes to be fictional is stored in a decoupled scope-representation with a source tag, whereas information one believes to be true is stored in semantic memory without a source tag. Similarly, when the retrieval context cued subjects that they were being asked about real-world knowledge, those who had been told the story was true verified sentences slightly faster than those who had been told it was fictional. On our account, retrieving the fictional information would take longer in this condition because the context cued subjects to search semantic memory; to find information that is stored in a fiction scope-representation, they would have to switch their search to a different, decoupled memory system (see Gerrig & Prentice, 1991, for similar results). In other words, the results suggest that information from fiction is stored separately from true information ("compartmentalized", to use Potts, St. John, & Kirson's [1989] preferred term) and that it is clearly marked with a source tag, which can be accessed in retrieval.

It is not clear whether there are neuropsychological deficits that selectively knock out a person's ability to understand fiction without simultaneously knocking out other imaginative activities. But autism does seem to be an example of a developmental disorder that selectively impairs the imagination. A great deal of research is currently going on to track down the brain systems involved (Baron-Cohen, Ring, Moriarty, et al., 1994; Fletcher, Happe, Frith, et al., 1995; Stone, Baron-Cohen, & Knight, in press). Most evidence so far points to the frontal lobes, although there is still controversy over exactly which locations (dorsolateral, orbitofrontal, etc.) are involved.

Representations of a Personal Past: Episodic Memory

Tulving (1995) argues that there are at least five functionally distinct memory systems. Semantic memory is one of the five; episodic memory is another (Tulving, 1993a; 1995). Perhaps the best way to convey the difference between the two is to quote Tulving himself.

> Semantic memory registers, stores, and makes available for retrieval knowledge about the world in the broadest sense: If a person knows something that is in principle describable in propositional form, that something belongs to the domain of semantic memory. Semantic memory enables individuals to represent states, objects, and relations in the world that are not present to the senses ... Episodic memory is memory for

> personally experienced events ... It transcends semantic memory by being ego-centered: its contents include a reference to the self in subjective space/time. Its operations are subserved by a neurocognitive system *specialized* for that purpose. The owner of an episodic memory system is capable not only of mental space travel but also mental time travel: it can transport itself at will into the past, as well as into the future, a feat not possible for those who do not possess episodic memory. (Tulving 1993a, p. 67; italics in the original)

Note that episodic memory is not equivalent to autobiographical knowledge. Autobiographical knowledge can be stored in either episodic or semantic memory. This is captured by the distinction between remembering and knowing: "I recall seeing the Grand Canyon" (episodic) versus "I know that I saw the Grand Canyon" (semantic). Tulving further points out that the nature of conscious awareness (*qualia*) that accompanies retrieval of information differs for episodic and semantic memory. Episodic retrieval is accompanied by autonoetic awareness, "a distinctive, unique awareness of re-experiencing here and now something that happened before, at another time and in another place. The awareness and its feeling-tone is intimately familiar to every normal human being" (Tulving, 1993a, p. 68). Semantic retrieval is accompanied by noetic awareness, "the kind of awareness that characterizes thinking about other facts of the world" (Tulving, 1993a, p. 68).

In this view, the episodic system is different from the semantic system because the computational requirements of a system that can remember a personal past are very different from those of a system for storing and retrieving general knowledge. General knowledge (a) need not come from one's own perceptual experience, (b) does not require a source tag, and (c) if it initially has a source tag, that usually fades with time, as the proposition is increasingly well-validated by converging sources of evidence. Moreover, general knowledge (d) need not refer to the past (it can be about time-invariant properties of the world, such as that the sun rises every day), (e) needs to be retrievable for use at any time and in a wide variety of circumstances, and (f) its retrieval should be relatively independent of the context in which the information was learned (e.g., it would be inefficient if we could only retrieve the fact that plants need sun to grow by first remembering who it was who first taught us about plants.)

At this point, a considerable body of data from experimental psychology and cognitive neuroscience supports the hypothesis that the design of episodic memory differs from that of semantic memory: that they form two functionally distinct systems (for reviews, see Nyberg & Tulving, 1996; Schacter & Tulving, 1994; Tulving, 1995; 1998). Far less attention has been paid, however, to *why* humans should have a functionally

distinct, episodic memory system: that is, to what its adaptive function is. Why should this system have evolved in our species? What can a person with episodic memory do that would be impossible for a person without episodic memory?

This question has only begun to be explored but there are a number of possibilities. For example, episodic memory may have evolved to handle the social world. Many social interactions have game-like properties, and to act in an adaptive manner, many strategies require that histories of interaction be stored, as well as the present state of play. We think three of the more interesting functions of episodic memory involve (1) re-evaluating conclusions in light of new information, (2) judging the credal value of information, especially that derived from other people, and (3) bounding the scope of generalizations. To illustrate, we give examples involving judgments of other people's character (see Klein, Cosmides, Tooby, & Chance, submitted) but the argument applies equally to nonsocial judgments.

It is known that people form generalizations about their own personality (e.g., "I am usually friendly") and that of other people (e.g., "Donna is usually shy"), which are stored in semantic memory (Klein & Loftus, 1993). After such a trait summary has been made, what can be gained by retaining the database of episodes – in quasi-perceptual form and with source tags – on which the summary was based?

(1) Re-evaluating Conclusions

New information may cause previous episodes to be re-interpreted, drastically changing one's judgments of a person or a situation. Fred's friendly willingness to help you with household repairs may take on different significance if you learn that he is attracted to your wife. If episodes were lost after they had been analyzed to form a summary judgment of Fred's character, re-evaluating his past actions in light of new information about his intentions and values would be impossible.

Keeping a database of episodes is helpful even when a drastic reinterpretation of previous events is not called for. Judgments can be revised in light of new information. If a judgment that "Fred is usually friendly" was based on 30 episodes, an unfriendly act by Fred should have less impact on it than if it had been based on three episodes (Cosmides & Tooby, 1996a; see also Sherman & Klein, 1994). Without the original database, it is difficult to know whether new, inconsistent information should change one's summary judgment and, if so, by how much. Moreover, new reference classes can be formed to answer new questions. Suppose you need to decide whether your best friend would make a good employee – something you had never considered before. If a database of richly encoded episodes exists, it can be sifted for events relevant to making such a judgment.

(2) Evaluating Credal Value

Maintaining source information allows one to evaluate the credal value of stored information. Noting that (a) confidence ratings are positively correlated with judgments that one has "remembered" a fact (as opposed to having simply "known" it) and (b) amnesic patients lack a subjective sense of certainty about knowledge that they do, in fact, possess, Tulving suggested that the adaptive value of the autonoetic consciousness associated with episodic memory "lies in the heightened subjective certainty with which organisms endowed with such memory and consciousness believe, and are willing to act upon, information retrieved from memory ... [leading] to more decisive action in the present and more effective planning for the future" (Tulving, 1985, p. 10). Information derived from perception should be assigned a higher credal value than information derived from other people. This would explain why we might have a memory system that allows retrieval of engrams with a quasi-perceptual format: Preserving perceptual information allows one to "re-experience" – to retrieve a broad-band encoding of the original event in which a piece of information was encountered or from which it was inferred. The benefit of this is even stronger for humans, who get such a large proportion of their information from others. Most organisms acquire all of their (non-innate) knowledge through their own senses and, so, have less need for a system that discriminates between perceptually derived information and information from other sources.

For a species that subsists on information, much of it supplied by other people, judging how reliable that information is can be a matter of life and death. For example, the !Kung San, a hunter-gatherer group in the Kalahari desert of Botswana, distinguish sharply between the following four kinds of evidence: (1) "I saw it with my own eyes"; (2) "I didn't see it with my own eyes but I saw the tracks. Here is how I inferred it from the tracks"; (3) "I didn't see it with my own eyes or see the tracks but I heard it from many people (or a few people or one person) who saw it"; (4) "It's not certain because I didn't see it with my eyes or talk directly with people who saw it" (Tulkin & Konner, 1973, p. 35).

Assessing credal value is particularly important in navigating the social world. Often informants have agendas that bias or distort the information they communicate to you. Imagine that Eve, who immediately befriended you when you started your new job, told you many terrible things about another co-worker, Adam. Much later, you find out that she has been stalking him ever since he broke up with her a year ago. As a result, you realize that the stories you heard from Eve might well be untrue. In forming an impression of Adam, you integrated information from many sources. But one of these sources turned out to be unreliable: Eve has sowed the seeds of data corruption.

How can you update your judgments about Adam? Which of your trait summaries for Adam are still reliable, and which are not? A database of episodic memories would allow you to re-evaluate your judgments about Adam. Your "Adam database" would include episodes in which you had interacted with Adam yourself, episodes in which other people told you things about Adam, and episodes in which Eve told you stories about Adam. Because all these episodes have source tags, you can "consider the source": you can sort through your database and decide which judgments were based on sound information and which were colored by Eve's distortions. Had the episodes on which your judgments of Adam's character were based been lost, there would be no way to repair the corrupted segments of your semantic store. The ability to judge and re-evaluate the credal value of other people's communications is essential in an organism with language.

(3) Bounding the Scope of Generalizations

For quick decisions, it can be convenient to have summary judgments stored in semantic memory (Klein & Loftus, 1993; Klein, Cosmides, Tooby, & Chance, under review). But, there is a trade-off between speed and accuracy because information about particularities is inevitably lost in any generalization. Keeping an independent store of episodes allows the scope of a summary judgment – the circumstances under which it does, and does not, apply – to be specified. A trait summary such as "He is rarely honest" or "I am usually friendly" gives information about behavior under "average" circumstances, but it does not tell you under what circumstances the person's behavior deviates from average. In deciding how to behave, one is always facing a *particular* situation.

Imagine your semantic memory has an entry on Vanessa: "Vanessa is usually calm." You are planning what you hope will be a relaxed dinner party with some friends who are political activists of a different persuasion than Vanessa. Access to appropriate episodic memories can bound the scope of your semantic summary. Recalling that "Vanessa is usually calm – except those times we talked about abortion" may alter your decision about whom to invite. (Indeed, if there is a pattern to the exceptions, a summary of the exceptions might eventually be made as well and stored as an if-then proposition about the conditions under which Vanessa can be expected to become tense (Wright & Mischel, 1988).

Note that each of these three adaptive functions requires representations that are held separately from semantic memory, and that specify both the source of the information and the source's attitude toward it.

What are the Computational Requirements?

(1) Decoupling

The information in episodes is regulated and may be either isolated from other information structures, repressed, or activated and employed. It is regulated by time and space tags and does not necessarily have scope-unlimited implications: "At the time his wife became sick, we were not friends," rather than "We are not friends."

(2) Source Tag

The source of episodic memory representations is always the self, but also includes time and place tags. *When encoded*, the information was externally derived through perception and proprioception (the embedded knowledge represents a state of world *as perceived by the self*), so the encoding source tag should be $self_{external\ world}$. *When retrieved*, the memory has an internal source (one's own mind), so the source tag should be $self_{own\ mind}$. Without the $self_{external\ world}$ source tag, an episodic memory would not be a memory of having actually experienced something through perception. Even if it retained quasi-perceptual detail, there would be no way to tell whether it originated in perception or through someone else's account prompting mental imagery in the simulation system. The time and place tags allow the reconstruction of the state of play at the time of the event: for example, "Was he cool after you treated him unkindly, or before"? In fact, the scope tags identify when his change in attitude occurred by virtue of its position order of relevant social events rather than by reference to some other measure of time.

The information embedded in the episodic representation might itself have a source tag: "*I* recall that, at the water cooler, Eve told me that [Adam stole from the company]." Without such source tags, one could not distinguish which information was derived from one's own perceptual experiences and which was told to one by other people. Evaluating credal value and re-interpreting the meaning of past events would be impossible.

(3) Attitude Slot

According to Tulving, one has specific propositional attitudes toward the content of an episodic memory, which can be either "I *experienced* [the episode depicted]" or "I *am re-experiencing* [the episode depicted] *right now.*" This is a defining element distinguishing an episodic memory from a semantic memory. Events stored in episodic memory are *remembered, recalled,* or *recollected;* those stored in semantic memory are merely known.

(4) Memory Requirements

To have a personal past, one must (a) store the episode, which is equivalent to remembering the past (i.e., there is a time tag); (b) store the source of the experience (i.e., have a self-source tag attached to the engram), and (c) store the source of a linguistically transmitted proposition, in the case of an episode in which someone told you something.

(5) Relationship to Other Representation Systems

Episodic memories can be a source of input to many different kinds of decision rules. They may sometimes be retrieved in tandem with representations from semantic memory, for example, to bound the scope of a generalization (Klein, Cosmides, Tooby, & Chance, under review). They can be used to assess the credal value of propositions originating with others (e.g., re-evaluations after betrayal) or originating from the self (e.g., Did I see him with my own eyes? Did I just hear the rustling in the leaves and assume that it was him?).

What Is the Solution?

These computational requirements can be met by storing an episode in a specific kind of scope-representation, i.e., a representation that is regulated or decoupled from semantic memory, has a source tag (own experience, time *X*, location *Y*), has an attitude slot (= *experienced* or *am re-experiencing*), has a time tag (a place in a chronology), and has a place tag. Moreover, for any propositional content originating with another person, the episodic M-representation can include an embedded source tag indicating that person and his attitude toward the proposition. In retrieval, dissociations between the episodic memory system and semantic memory system have been documented many times.

Moreover, episodic memory is impaired by some of the same brain areas and syndromes as other functions that we have argued involve scope-representations. For example:

(a) In classic amnesic syndrome, semantic information is largely intact but the person cannot recall any personal episodes. Sometimes the person cannot recall episodes from before the accident that caused the amnesia (retrograde amnesia) and sometimes the person cannot recall episodes that occurred after the accident (anteriograde amnesia). This underscores the fact that episodic memories come with time tags that place them in a chronology or sequence. Some amnesiacs mistake things they imagined for things they actually experienced, creating confabulated, pseudomemories (e.g., Wilson & Wearing, 1995) – exactly what one would expect if decoupling were compromised.

(b) Frontal lobe damage selectively impairs episodic memory. Free recall is most impaired, then cued recall, then recognition memory (Wheeler, Stuss, & Tulving, 1995). In other words, frontal lobe damage particularly impairs retrieval that depends on having intact and/or accessible source, time, and place tags. (In free recall, these are the *only* basis on which the memory can be retrieved; cued recall and recognition memory provide "external" perceptual prompts, such that the memory could, in principle, be accessed through its embedded content, circumventing the need to retrieve via source, time, and place tags.) Frontal lobe damage is known to cause source amnesia (Janowsky, Shimamura, & Squire, 1989), and to impair memory for temporal order of events (Shimamura, 1995).

(c) Dissociations occur in new learning as well. K.C., the amnesic studied by Tulving, can learn new semantic information but he cannot remember any of the episodes in which he learned it (Hayman, Macdonald, & Tulving, 1993).

(d) This extends to learning about one's own personality. After the accident that rendered him amnesic, K.C.'s personality changed. But it turns out that he has trait summaries of the new personality, even though he has no access to the episodes on which the summaries were (presumably) based (Tulving, 1993b).

(e) Knowlton, Mangels, & Squire (1996) demonstrated what is arguably a double dissociation between episodic and semantic memory. Patients with Parkinson's disease were not able to learn a probabilistic rule, but they were able to recall the episodes that were the basis for learning in other subjects. Amnesics were able to learn the rule, but were not able to recall any episodes.

(f) PET studies suggest that episodic retrieval differentially engages the right hemisphere (including the right prefrontal cortex) whereas semantic retrieval differentially engages the left hemisphere (Nyberg, Cabeza, & Tulving, 1996; Tulving, 1998).

(g) In normal, brain-intact subjects, one can create functional dissociations between episodic and semantic memory. Retrieving trait summaries primes episodes that are inconsistent with the summary, but not those that are consistent with it (Klein, Cosmides, Tooby, & Chance, under review). This is what one would expect if one function of keeping a database of episodic memories was to allow one to bound the scope of generalizations.

If we are correct in positing that episodic memories are stored in scope-representations resembling M-representations, then three predictions follow: (1) Episodic memory should be impaired in individuals with autism (because individuals with autism cannot form M-representations; e.g., Baron-Cohen, 1995; Leslie, 1987; Leslie & Thaiss, 1992). (2) Episodic memory should not emerge in children until they are capable of forming M-representations. (3) Episodic memory should be impaired in individuals with any condition that damages the machinery that produces M-representations.

Studies of episodic memory in autism are just beginning, but preliminary results by Klein, Chan, & Loftus (under review) support the first prediction. Regarding the second prediction, we note that so-called "childhood amnesia" lasts until one is 3 to 4 years of age (Sheingold & Tenney, 1982; White & Pillemer, 1979; Perner, 1991) – approximately the time that children start to pass the false-belief task (a standard test of a mature ability to form M-representations; see Baron-Cohen, 1995). Moreover, the lack in preschool age children of a fully mature system for source tagging and forming multiply-embedded M-representations would explain an otherwise curious fact about their memories: They come to believe that they actually experienced events that never happened, if they are asked about these (fictitious) events repeatedly (Bruck & Ceci, 1999). Evidence for the third prediction will be presented below, in the context of schizophrenia.

Schizophrenia: A Test Case

If goals, plans, simulations, episodic memories, other people's beliefs, fiction, and so on are stored in, or regulated by, scope representations resembling M-representations, then an impairment to the M-representational system should disrupt these functions. For example, any condition that interferes with decoupling and source monitoring, and that impairs one's ability to make inferences about the attitude slot or contents of an M-representation, should lead to the corruption of semantic memory files. Semantic memory would store as true: fiction, false beliefs (originating in the self or others), unrealized goals and plans, and so on. A dysfunction in the machinery that produces or reads M-representations should impair episodic memory retrieval. Impaired links between M-representations and other representation systems – e.g., the ability of metarepresentations to suppress stimulus-driven actions – should lead to difficulties in communicating, controlling simulations, planning, and in executing actions specified by a plan, such as shifting from one M-represented goal to another.

The question is: Is there any syndrome or disease process that involves a breakdown of machinery necessary for producing, reading, or maintaining the integrity of metarepresentations?

Christopher Frith has argued – compellingly, in our view – that schizophrenia is a late-onset breakdown of a metarepresentational system (Frith, 1992). We cannot do justice to his argument and data in this chapter, but Table 1 gives a sense of how it accounts for some of schizophrenia's most distinctive symptoms. In Frith's view, goals and plans are metarepresented, and his book presents lucid accounts of how schizophrenia would cause disruptions in goals, plans, and inferences about other minds.

If our prior claims about what gets metarepresented are true *and* schizophrenia is caused by an impairment of a metarepresentational system, then what would this predict about memory in schizophrenia?

(a) If episodic memories are stored in metarepresentations, then schizophrenics should have episodic memory impairments.

(b) If intact metarepresentational ability is necessary to prevent data corruption, then one should see symptoms of impaired semantic memory in schizophrenics.

(c) If the executive component of working memory uses metarepresentations, then it should be disrupted in schizophrenia.

(d) Memory systems that do not depend on metarepresentations should be unaffected.

After making these predictions, we found that the literature on schizophrenia contained data that bears on them. Schizophrenia does indeed cause episodic memory impairment. According to McKenna, Mortimer, & Hodges, for example, "the existence of episodic memory impairment in schizophrenia is well established. [It is] selective and disproportionate to the overall level of intellectual impairment . . . [In schizophrenia] episodic memory is not only impaired but seems to be emerging as the leading neuropsychological deficit associated with the disorder" (McKenna, Mortimer, & Hodges, 1994, pp. 163, 169).

There is also evidence that semantic memory becomes corrupted in schizophrenia. In sentence verification tasks, which require retrieval of general knowledge from semantic memory, schizophrenics (a) are slower than normals (two-thirds fall outside normal range), (b) make more classification errors than normals, and (c) usually (but not always) misclassify false statements as true (McKenna, Mortimer, & Hodges, 1994). The last feature is perhaps the most interesting as it is exactly the kind of representational corruption that one would expect from a dysfunction of a metarepresentational system. The purpose of decoupling is to allow one to represent false beliefs, fictions, and not-yet-existing states of affairs separately from the database of true propositions stored in semantic memory. Damage to a decoupling system would therefore

Table 1: Symptoms of Schizophrenia Related to Impairment of the Metarepresentation System

Impaired source monitoring

- *thought insertion*: experience internally generated thoughts as originating from external agent (*"auditory" hallucinations*)
- *delusions of control*: experience own actions as having been caused by external agent rather than by self

Impaired ability to infer intentions, attitudes, and/or content of beliefs – other people's and one's own

- *delusions of reference*: (falsely) believe other people's communications are aimed at self
- *paranoia*: false beliefs about other people's beliefs and intentions
- *difficulty in communicating*: cannot infer relevance

Impaired ability to plan and/or execute plans in action

- incoherent speech
- lack of volition
- psychomotor slowness

(based on Frith, 1992)

cause false propositions to be stored as true, resulting in more misclassifications of this kind. (Alternatively, a breakdown in the system that allows one to understand that beliefs can be false might cause a response bias towards accepting sentences as true.)

The following phenomena are also plausibly interpreted as resulting from corruption of semantic memory: (a) formal thought disorders (ideas that appear disordered to an outside observer can result from false information; to see this, consider what the child in Leslie's example might say if she really did think that telephones were edible and yellow, or that fruit could serve as a transmitter of voices); (b) a "tendency for concepts to become pathologically large, their boundaries loose and blurred, and their content accordingly broad, vague, and overlapping" (McKenna, Mortimer, & Hodges, 1994, p. 176); and (c) exaggerated semantic priming (McKenna, Mortimer, & Hodges, 1994.[21]

Damage to a metarepresentation system should also have sequelae for working memory. We posited that the executive component of work-

ing memory uses metarepresentations to select content for input into other systems (such as the visuospatial sketchpad). If we are correct, then schizophrenia should impair the executive component of working memory, but not its slave systems (except insofar as their functions depend on an intact executive). There is, indeed, evidence of deficits in the executive (McKenna, Clare, & Baddeley, 1995), while the functioning of the articulatory loop is normal as well as verbal and non-verbal short-term and primary memory.

More generally, damage to the metarepresentational components of the system should leave any memory system that is not scope-regulated relatively intact. This does seem to be the pattern in schizophrenia. The procedural memory system (responsible for conditioning, storage of automated motor sequences, and habits) and the perceptual-representational system (responsible for object recognition and implicit priming) both appear to be intact in people with schizophrenia. In fact, McKenna, Clare, & Baddeley (1995) found that the pattern of memory impairment in schizophrenia is similar to the pattern in classic amnesic syndrome – except that there is evidence of some semantic memory corruption in schizophrenia. In the classic amnesic syndrome, there is (a) impaired episodic memory, (b) impairments in executive functions of working memory, (c) intact working-memory slave systems (articulatory loop, visuospatial sketchpad), (d) intact procedural memory, (e) intact PRS memory (implicit priming), and (f) intact semantic memory. This is the same pattern as found in schizophrenia, with the exception of the semantic memory. In amnesia due to head injury, there is no reason to think that inferences about metarepresentations would be impaired. Hence, there is no reason to expect corruption of semantic memory.

Conclusions

Behaviorally, humans are the strangest species that we have encountered so far. How did we get this way? The hypothesis that the ability to form metarepresentations initially evolved to handle the problems of modeling other minds (Leslie, 1987; Baron-Cohen, 1995) or the inferential tasks attendant to communication (Sperber, 1996; this volume; Sperber & Wilson, 1986) is very plausible, and our thinking is heavily indebted to this body of work. Still, the problems handled by metarepresentations, scope syntax, and decoupling are so widespread, and participate in so many distinct cognitive processes, that it is worth considering whether they were also shaped by selection to serve a broader array of functions – functions deeply and profoundly connected to what is novel about hominid evolution.

The central engine that has driven humanity down its unique evolutionary path may have been selection for computational machinery that allowed our species to enter what we have called the cognitive niche: that is, machinery that radically increased our ability to extract and exploit information that is local, transient, and contingent, wringing inferences from it that permit us to devise plans of action and behavioral routines that are successfully tailored to local conditions. For humans to enter, survive in, and take advantage of this strange new world of uncertain representations and the inferences that can be drawn from them, the human cognitive architecture had to evolve cognitive adaptations that solve the special problems that it posed. Because this new type of information is only applicable temporarily, locally, or contingently, the success of this computational strategy depends on the existence of machinery that ceaselessly locates, monitors, updates, and represents the conditional and mutable boundaries within which each set of representations remains useful. The problem of tracking the applicable scope of information is magnified by the fact that inference propagates errors, given that contingent information is often wrong outside its envelope of valid conditions. An error in the information that serves as input to an inference program will often lead to errors in the output, which may then be fed as input into yet other inference programs. As a result, a defective representation has the power to infect any data set with which it subsequently interacts, damaging useful information in contagious waves of compounding error. Inference is more powerful to the extent that information can be integrated from many sources, but this multiplies the risk that valid existing information sets will be progressively corrupted. Hence, the novel evolutionary strategy of using contingent information and densely networked inferential processing to regulate behavior could only evolve if natural selection could devise computational methods for managing the threat posed by false, unreliable, obsolete, out-of-context, deceptive, or scope-violating representations. Cognitive firewalls – systems of representational quarantine and error correction – have evolved for this purpose. They are, no doubt, far from perfect. But without them, our form of mentality would not be possible.

In this chapter, we have attempted to sketch out a few elements of the large series of specialized computational adaptations that we believe evolved to handle these problems. These include elements of a scope syntax, the regulated decoupling and recoupling of data structures, and metarepresentations. The basic elements of scope syntax must be built into the evolved architecture of our species because (i) there is a combinatorially infinite array of possible scope systems (e.g., ways of dividing up information into subsets, and procedures for regulating their permitted interactions), (ii) there are no observable models to which one can compare the output of a scope syntax for the purpose of modifying it so

that it will perform more adaptively, and (iii) the problem of attributing computational success or failure to the scope-regulating design features responsible appears to be intractable, given that inferential networks are complex and that there is an open-ended set of variations that could be introduced ontogenetically. It remains, however, very likely that evolved developmental programs (as opposed to machinery invented *de novo* during ontogeny) can establish new boundaries and patterns of connection and dissociation over the course of the lifespan (as when, e.g., the representations produced by a wandering eye are disconnected from, and therefore cease to influence, or interfere with, higher levels of visual processing).

We are agnostic about whether the evolution of metarepresentational, scope-syntax, and decoupling machinery that subserves mindreading and social interaction was a precondition for entering the cognitive niche, or whether the mind-reading machinery evolved after, or in tandem with, the machinery that accomplishes these functions in the other domains we discussed. That question can only be answered by a combination of (1) comparative studies of mindreading, planning, and other scope-regulated abilities in species that vary in the extent to which their evolutionary history involved complex social interaction and tool use and (2) close analysis in humans of the design features that accomplish scope-regulation in different domains, to see exactly how computationally similar they really are.

Many questions about the architecture that accomplishes scope-regulation are wide open. It is not clear, for example, whether the same neural system implements source tags, decoupling, and scope regulation for disparate cognitive activities, or whether different circuits with similar functional properties have been duplicated (and, perhaps, modified by selection) in different parts of the brain. Demonstrations by Leslie & Thaiss (1992) and by Charman & Baron-Cohen (1993) that one can lose the ability to reason about mental representations while retaining quite parallel abilities to reason about nonmental representations (such as photographs, models, and maps) suggests neural parallelism. In contrast, Christopher Frith's (1992) analysis of a patterned breakdown of metarepresentational abilities in schizophrenia (and some of our additions to his analysis) suggest that at least some of the requisite neural circuitry might be shared across functions. Another architectural question that remains open is the extent to which decoupling and scope-regulation are handled by explicitly syntactic features of cognitive operations (e.g., by source tags and operators within a deliberative reasoning system). In some cases, the same decoupling functions might be handled by neural independence, that is, by an architecture in which the outputs of certain imaginative, planning, or memory functions are quarantined from semantic memory or other representational systems by virtue of their being located in physically separate subsystems, without machinery that allows their outputs to become inputs to the systems that they could corrupt.

The exploration of the properties of scope management is just beginning and it would be premature to claim that any such proposals about the architecture have yet been established. Still, we believe that much that is so distinctive and otherwise puzzling about the human mind – from art, fiction, morality, and suppositional reasoning to dissociative states of consciousness, imaginary worlds, and philosophical puzzles over the semantic properties of propositional attitudes to the function of aesthetic sensibilities and the improvisational powers of human intelligence – are attributable to the operation of these adaptations. Further investigation of these issues seems to hold substantial promise.

Acknowledgments

This paper owes a deep intellectual debt to Alan Leslie for his work on pretense, propositional attitudes, and decoupling; to Dan Sperber for his work on metarepresentations and communication; and to Christopher Frith for his work on schizophrenia. We also wish to warmly thank Paul Hernadi, Michelle Scalise Sugiyama, and Francis Steen for their many illuminating thoughts about the nature of fiction. We thank the James S. McDonnell Foundation, the National Science Foundation (NSF Grant BNS9157-449 to John Tooby), and the UCSB Office of Research (through a Research across Disciplines grant: Evolution and the Social Mind) for their financial support.

Notes

1 Although some successful improvisations may be conserved across multiple lifespans and spread across many individuals, they still are very rapid with respect to the time it takes selection to operate.
2 By rules or procedures, we only mean the information-processing principles of the computational system, without distinguishing subfeatural or parallel architectures from others.
3 or stable frequency-dependent equilibria.
4 Indeed, the world outside the local conditions may be commonly encountered and, depending on how narrow the envelope of conditions within which the information is true, scope-violating conditions are likely to be far more common than the valid conditions.
5 i.e., to be de-encapsulated
6 There is no need, in particular, for the data-structure to be a sentence-like or quasi-linguistic proposition. For most purposes, when we use the term

"proposition" throughout this chapter, we are not commiting ourselves to quasi-linguistic data-structures – we will simply be using it as a convenient short-hand term for a data-element of some kind.

7 While not everyone would accept this as a metarepresentation, we think that such a rejection was a convenient rather than an accurate way of dealing with such problems as referential opacity.

8 Various operators and features of the workspace provide the intuitions that logicians have elaborated into various formal logics – the elaboration taking place through the addition of various elements not found in the workspace, the attempt simultaneously to impose self-consistency and conformity to intuition, and the removal of many content-specific scope-operators. For the human architecture itself, there is no requirement that the various procedures available to the workspace be mutually consistent, only that the trouble caused by inconsistency be less than the inferential benefits gained under normal conditions. Task-switching and scope-limiting mechanisms also prevent the emergence of contradictions during ordinary functioning, which makes the mutual consistency of the architecture as an abstract formal system not relevant. Mental-logic hypotheses for human reasoning have been rejected empirically by many on the assumption that the only licensed inferences are logical. We believe that the content-sensitivity of human reasoning is driven by the existence of domain-specific inference engines, which coexist beside operators that parallel more traditional logical elements.

9 There are, as well, heterarchical relations, governed by rules for data incorporation from other sources.

10 Promotion is equivalent to Tarskian disquotation with respect to the next level in the architecture.

11 Indeed, this kind of architecture offers a computational explanation of what kind of thing deontic ascriptions are: decoupled descriptions of possible actions and states of affairs, of suspended truth value, connected to value assignments of the possible actions.

12 Such an architecture explains how humans process fictional worlds without confusing their environments and inhabitants with the real world.

13 We think that ground state representations are present in consciousness, but are not automatically the objects of consciousness – that is, we are not automatically reflectively conscious of these data structures, although they can easily be made so. Data-structures in the ground state must be demoted to become the object of inferential scrutiny. Indeed, we think that the function of the architectural component that corresponds to one referent of the word consciousness is to be a buffer to hold isolated from the rest of the architecture the intermediate computational work products during the period when their truth-value and other merits are unevaluated. This explains why consciousness is so notoriously volatile.

14 A ubiquitous phenomenon, familiar to professors, is that when students deeply assimilate the knowledge being taught, they often forget who taught it to them, and feel compelled to excitedly share what they have learned from their teachers with their teachers.

15 We are not claiming that every propositional attitude term, for example, is reliably developing or "innate." We consider it more plausible that there is an evolved set of information-regulatory primitives that can be combined to produce a large set of scope-operators and scope-representations.

16 What other causes could there be? One taking a *physical stance* might mention muscle contractions and force; one taking a *design stance* might mention the evolution of food seeking mechanisms; a behaviorist taking a *contingency stance* might mention a history of reinforcement; an astronomer might mention the Big Bang as a necessary (though not sufficient) cause; and so on.

17 Perner (1991) states that episodic traces are engrams with a metarepresentational comment regarding how the information was obtained. This is not quite an M-representation in Leslie's sense (see Perner, 1991, p. 35). However, Perner does not argue that episodic traces are metarepresentational because this is the only way that certain computational requirements can be met.

18 It is not clear why this is possible. The framework of Johnson, Hashtroudi, & Lindsay (1993) emphasizes inference in source monitoring; in this view, proprioceptive feedback may be critical to source monitoring, and the deep relaxation of hypnosis may interfere with proprioception (see also Kunzendorf (1985–1986) for a view more closely related to source tagging). It should also be noted that individuals differ in their hypnotic susceptibility – in their ability to enter "dissociative" states. It would be interesting to find out whether hypnotic susceptibility were related to individual differences in source monitoring or in decoupling. Two of the few things that correlates with hypnotic susceptibility is the tendency to become engrossed in movies or books, and vividness of imagery – both of which are plausibly related to scope-representational abilities (see sections on Fiction and Simulations).

19 Frith argues that perseveration occurs when the person knows a response is required of him but has trouble generating willed actions. Because the person either cannot form plans or cannot transform them into willed intentions, he simply repeats the last thing.

20 It is also interesting to note that dopamine is an inhibitory neurotransmitter. It is reasonable to assume that stimulus-driven action systems are evolutionarily more ancient than systems that allow the formation of plans and willed intentions; moreover, excitatory neurotransmitters, which open ion gates, are far more common than inhibitory ones. Plan S-representations would be part of an evolutionarily more recent system, which is designed to inhibit the more ancient stimulus-driven action system when a plan is to be enacted. A straightforward way of doing so would be through an inhibitory neurotransmitter, that is, one that operates by closing ion gates.

21 Seeing a semantically related word speeds time to classify a string of letters as word or non-word; this is known as semantic priming. Having pathologically large concepts means a wider variety of words will be seen as semantically related. This would lead to "exaggerated semantic priming" in schizophrenics. Indeed, schizophrenics with other evidence of formal thought disorder show exaggerated priming compared to controls.

References

Baddeley, A. (1995). Working memory. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences* (pp. 755–764). Cambridge, MA: MIT Press.

Baillergeon, R. (1986). Representing the existence and the location of hidden objects: Object permanence in 6- and 8-month old infants. *Cognition 23*, 21–41.

Baron-Cohen, S. (1995). *Mindblindness: An essay on autism and theory of mind.* Cambridge, MA: MIT Press.

Baron-Cohen, S., Cross, P., Crowson, M., & Robertson, M. (1994). Can children with Tourette's Syndrome edit their intentions? *Psychological Medicine 24*, 29–40.

Baron-Cohen, S., Leslie, A., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition 21*, 37–46.

Baron-Cohen, S., Ring, H. Moriarty, J., Schmitz, B., Costa, D., & Ell, P. (1994). Recognition of mental state terms: Clinical findings in children with autism and a functional neuroimaging study of normal adults. *British Journal of Psychiatry 165*, 640–649.

Baron-Cohen, S., Robertson, M., & Moriarty, J. (1994a). The development of the will: A neuropsychological analysis of Gilles de la Tourette's Syndrome. In D. Cicchetti and S. Toth (Eds.), *The self and its dysfunction: Proceedings of the 4th Rochester symposium.* Rochester, NY: University of Rochester Press.

Bisiach, E., & Luzzatti, C. (1978). Unilateral neglect of representational space. *Cortex 14*, 129–133.

Bowers, K. S. (1977). *Hypnosis for the seriously curious. New York: Jason Aronson.*

Brase, G., Cosmides, L., & Tooby, J. (1998). Individuation, counting, and statistical inference: The role of frequency and whole object representations in judgment under uncertainty. *Journal of Experimental Psychology: General 127* (1), 1–19.

Bruck, M., & Ceci, S. (1999). The suggestibility of children's memory. *Annual Review of Psychology 50* 419–439.

Charman, T., & Baron-Cohen, S. (1993). Understanding photos, models, and beliefs: A test of the modularity thesis of theory of mind. *Cognitive Development 10*, 287–298.

Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. *Cognition 31*, 187–276.

Cosmides, L., and Tooby, J. (1987). From evolution to behavior: Evolutionary psychology as the missing link. In J. Dupré (Ed.), *The latest on the best: Essays on evolution and optimality.* Cambridge, MA: MIT Press.

Cosmides, L., & Tooby, J. (1989). Evolutionary psychology and the generation of culture, Part II. Case study: A computational theory of social exchange. *Ethology and Sociobiology 10*, 51–97.

Cosmides, L., &Tooby, J. (1992). Cognitive adaptations for social exchange. In J. Barkow, L. Cosmides, and J. Tooby (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture.* New York: Oxford University Press.

Cosmides, L., & Tooby, J. (1996a). Are humans good intuitive statisticians after all? Rethinking some conclusions of the literature on judgment under uncertainty. *Cognition 58*, 1–73.

Cosmides, L., & Tooby, J. (1996b). A logical design for the mind? Review of *The psychology of proof*, by Lance J. Rips (1994, MIT Press). *Contemporary Psychology 41* (5), 448–450.

Cosmides, L., & Tooby, J. (1997). Dissecting the computational architecture of social inference mechanisms. In *Characterizing human psychological adaptations* (Ciba Foundation Symposium Volume #208). Chichester: Wiley.

Cosmides, L., & Tooby, J. (in press). Unraveling the enigma of human intelligence: Evolutionary psychology and the multimodular mind. In R. J. Sternberg & J. C. Kaufman (Eds.). *The evolution of intelligence.* Hillsdale, NJ: Erlbaum.

Ceci, S. (1995). False beliefs: Some developmental and clinical observations. In D. Schacter (Ed.), *Memory distortions* (pp. 91–125). Cambridge, MA: Harvard University Press.

Damasio, A., & Van Hoesen, G. (1983). Focal lesions of the limbic frontal lobe. In K. Heilmman & P. Satz (Eds.), *Neuropsychology of human emotion* (pp. 85–110). NY: Guilford Press.

Dennett, D. (1987). *The intentional stance.* Cambridge, MA: MIT Press.

Devinsky, O., Morrell, M., & Vogt, B. (1995). Contributions of anterior cingulate cortex to behaviour. *Brain 118*, 279–306.

Duncan, J. (1995). Attention, intelligence, and the frontal lobes. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences* (pp. 721–733). Cambridge, MA: MIT Press.

Farah, M. (1984). The neurological basis of mental imagery: A componential analysis. *Cognition 18*, 245–272.

Farah, M. (1990). *Visual agnosia: Disorders of object recognition and what they tell us about normal vision.* Cambridge, MA: MIT Press.

Farah, M., Soso, M., & Dasheiff, R. (1992). Visual angle of the mind's eye before and after unilateral occipital lobectomy. *Journal of Experimental Psychology: Human Perception and Performance 19*, 241–246.

Fletcher, P., Happe, F., Frith, U., Baker, S., Dolan, R., Frackowiak, R., & Frith, C. (1995). Other minds in the brain: A functional imaging study of "theory of mind" in story comprehension. *Cognition 57*, 109–128.

Frege, G. 1892. On sense and reference. In P. Geach & M. Black (Eds.), *Translations of the philosophical writings of Gottlob Frege.* Oxford: Blackwell.

Frith, C. (1992). *The cognitive neuropsychology of schizophrenia.* Hillsdale, NJ: Erlbaum.

Frith, C., & Frith, U. (1991). Elective affinities in schizophrenia and childhood autism. In P. E. Bebbington (Ed.), *Social psychiatry: Theory, methodology, and practice.* London: Transaction.

Frith, U. (1989). *Autism: Explaining the enigma.* Oxford: Basil Blackwell.

Frith, U., Morton, J., & Leslie, A. (1991). The cognitive basis of a biological disorder: Autism. *Trends in Neuroscience 14*, 433–438.

Gentzen, G. (1969). Investigations into logical deduction. In M. E. Szabo (Ed.), *The collected papers of Gerhard Gentzen* (pp. 405–431). (Original work published 1935.)

Gerrig, R., & Prentice, D. (1991). The representation of fictional information. *Psychological Science 2*, 336–340.

Goldberg, G. (1985). Supplementary motor area structure and function: Review and hypotheses. *Behavioral and Brain Sciences 8*, 567–616.

Hayman, C., Macdonald, C., & Tulving, E. (1993). The role of repetition and associative interference in new semantic learning in amnesia: A case experiment. *Journal of Cognitive Neuroscience 5*, 375–389.

Hilgard, E. (1977). *Divided consciousness: Multiple controls in human thought.* New York: Wiley.

Humphrey, Nicholas (1992). *A history of the mind.* New York: Simon and Schuster.

Ingvar, D. (1985). Memory of the future: An essay on the temporal organization of conscious awareness. *Human Neurobiology 4*, 127–136.

Janowsky, J., Shimamura, A., & Squire, L. (1989). Source memory impairment in patients with frontal lobe lesions. *Neuropsychologia 27*, 1043–1056.

Johnson, M., Hashtroudi, S., & Lindsay, D. (1993). Source monitoring. *Psychological Bulletin 114*, 3–28.

Klein, S., Chan, R., & Loftus, J. (under review). Independence of episodic and semantic self-knowledge: The case from autism.

Klein, S., Cosmides, L., Tooby, J., & Chance, S. (under review, 1999). Decisions and the evolution of memory: multiple systems, multiple functions. *Psychological Review.*

Klein, S., & Loftus, J. (1993). The mental representation of trait and autobiographical knowledge about the self. In T. Srull & R. Wyer (Eds.), *The mental representation of trait and autobiographical knowledge about the self:* Vol. 5. *Advances in Social Cognition.* Hillsdale, NJ: Erlbaum.

Knowlton, B., Mangels, F., & Squire, L. (1996). A neostriatal habit learning system in humans. *Science 273*, 1399–1402.

Kosslyn, S. (1994). *Image and brain: The resolution of the imagery debate.* Cambridge, MA: MIT Press.

Kosslyn, S., Alpert, N., Thompson, W., Maljkovic, V., Weise, S., Chabris, C., Hamilton, S., Rach, S., & Buonanno, F. (1993). Visual mental imagery activates topographically organized visual cortex: PET investigations. *Journal of Cognitive Neuroscience 5*, 263–287.

Krebs, J. R., & Dawkins, R. (1984). Animal signals: Mind reading and manipulation. In J. R. Krebs & N. B. Davies, *Behavioural ecology: An evolutionary approach* (2nd ed.) (pp. 380–402). Oxford: Blackwell.

Kripke, S. (1979). A puzzle about belief. In A. Margalit (Ed.), *Meaning and Use.* Dordrecht: Reidel.

Kunzendorf, R. (1985–1986). Hypnotic hallucinations as "unmonitored" images: An empirical study. *Imagination, Cognition and Personality 5*, 255–270.

Lee, R. B. (1993). *The Dobe Ju/'hoansi.* (2nd ed.). New York: Holt, Reinhart, & Winston.

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

Leslie, A. (1988). The necessity of illusion: Perception and thought in infancy. In L. Weiskrantz (Ed.), *Thought without language* (pp. 185–210). Oxford: Clarendon Press.

Leslie, A. (1994). ToMM, ToBy, and Agency: Core architecture and domain specificity. In L. Hirschfeld & S. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture.* New York: Cambridge University Press.

Leslie, A., & Frith, U. (1990). Prospects for a cognitive neuropsychology of autism: Hobson's choice. *Psychological Review 97*, 122–131.

Leslie, A., & Thaiss, L. (1992). Domain specificity in conceptual development: Neuropsychological evidence from autism. *Cognition 43*, 225–251.

McKenna, P., Clare, L., & Baddeley, A. (1995). Schizophrenia. In A. Baddeley, B. Wilson, & F. Watts (Eds.), *Handbook of memory disorders.* New York: Wiley.

McKenna, P., Mortimer, A., & Hodges, J. (1994). Semantic memory and schizophrenia. In A. David & J. Cutting (Eds.), *The neuropsychology of schizophrenia.* Hove, Sussex: Erlbaum.

Nyberg, L., & Tulving, E. (1996). Classifying human long-term memory: Evidence from converging dissociations. *European Journal of Cognitive Psychology 8*, 163–183.

Nyberg, L., Cabeza, R., & Tulving, E. (1996). PET studies of encoding and retrieval: The HERA model. *Psychonomic Bulletin and Review 3*, 135–148.

Pardo, P., Pardo, K., Janer, W., & Raichle, M. (1990). The anterior cingulate cortex mediates processing selection in the Stroop attentional conflict paradigm. *Proceedings of the National Academy of Sciences (USA) 87*, 256–259.

Perner, J. (1991). *Understanding the representational mind.* Cambridge: MIT Press.

Posner, M., & Raichle, M. (1994). *Images of mind.* New York: Freeman.

Potts, G., St. John, M., & Kirson, D. (1989). Incorporating new information into existing world knowledge. *Cognitive Psychology 21*, 303–333.

Richard, M. (1990). *Propositional attitudes: An essay on thoughts and how we ascribe them.* Cambridge: Cambridge University Press.

Rips, Lance J. (1994). *The psychology of proof: deductive reasoning in human thinking.* Cambridge, MA: MIT Press.

Schacter, D. (1995). Implicit memory: A new frontier for cognitive neuroscience. In M. Gazzaniga (Ed.), *The cognitive neurosciences* (pp. 815–824). Cambridge, MA: MIT Press.

Schacter, D., & Tulving, E., Eds. (1994). *Memory systems 1994.* Cambridge, MA: MIT Press.

Sheingold, K., & Tenney, Y. (1982). Memory for a salient childhood event. In U. Neisser (Ed.), *Memory observed* (pp. 201–212). San Francisco: W. H. Freeman.

Shepard, R. N. (1984). Ecological constraints on internal representation: Resonant kinematics of perceiving, imagining, thinking, and dreaming. *Psychological Review 91*, 417–447.

Shepard, R. N. (1987). Evolution of a mesh between principles of the mind and regularities of the world. In J. Dupré (Ed.), *The latest on the best: Essays on evolution and optimality* (pp. 251–275). Cambridge, MA: MIT Press.

Shepard, R. (1994). The mesh between mind and world. The William James Lectures, Harvard University, Cambridge, MA.

Sherman, J., & Klein, S. (1994). Development and representation of personality impressions. *Journal of Personality and Social Psychology 67*, 972–983.

Shimamura, A. (1995). Memory and frontal lobe function. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences* (pp. 803–813). Cambridge, MA: MIT Press. .

Spelke, E. (1988). The origins of physical knowledge. In L. Weiskrantz (Ed.), *Thought without language* (pp. 168–184). Oxford: Clarendon Press.

Spelke, E. (1990). Principles of object perception. *Cognitive Science 14*, 29–56.

Sperber, D. (1985). Anthropology and psychology: Towards an epidemiology of representations. The Malinowski Memorial Lecture, 1984. *Man (N.S.) 20*, 73–89.

Sperber, Dan. (1996). *Explaining culture: A naturalistic approach.* Oxford and Cambridge, MA: Blackwell.

Sperber, D. (this volume). Culture and the epidemiology of meta-representations. *Tenth Annual Vancouver Cognitive Science Conference,* Simon Fraser University, Vancouver, Canada.

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

Steen, F. (personal communication). Dept. of English, University of California, Santa Barbara.

Stone, V., Baron-Cohen, S., & Knight, R. (in press). Does frontal lobe damage produce theory of mind impairment? *Journal of Cognitive Neuroscience.*

Sugiyama, M. Scalise (1996). On the origins of narrative: Storyteller bias as a fitness-enhancing strategy. *Human Nature 7,* 403–425.

Symons, D. (1993). The stuff that dreams aren't made of: Why wake-state and dream-state sensory experiences differ. *Cognition 47,* 181–217.

Talmy, L. (1988). Force dynamics in language and cognition. *Cognitive Science 12,* 49–100.

Tooby, J., & Cosmides, L. (1989). Evolutionary psychology and the generation of culture, Part I. Theoretical considerations. *Ethology & Sociobiology 10,* 29–49.

Tooby, J., & Cosmides, L. (1990). The past explains the present: emotional adaptations and the structure of ancestral environments. *Ethology and Sociobiology 11,* 375–424.

Tooby, J., & Cosmides, L. (1992). The psychological foundations of culture. In J. Barkow, L. Cosmides, & J. Tooby (Eds.), *The adapted mind: Evolutionary psychology and the generation of culture.* New York: Oxford University Press.

Tooby, J., & Cosmides, L. (in press). Ecological rationality and the multimodular mind: Grounding normative theories in adaptive problems. In J. Tooby & L. Cosmides, *Evolutionary psychology: Foundational papers.* Foreword by Steven Pinker. Cambridge, MA: MIT Press.

Tooby J., & DeVore, I. (1987). The reconstruction of hominid behavioral evolution through strategic modeling. In W. Kinzey (Ed.), *Primate Models of Hominid Behavior.* New York: SUNY Press.

Tulkin, S., & Konner, M. (1973). Alternative conceptions of intellectual functioning. *Human Development 16,* 33–52.

Tulving, E. (1983). *Elements of episodic memory.* Oxford: Oxford University Press.

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology/ Psychologie Canadienne 26,* 1–12.

Tulving, E. (1993a). What is episodic memory? *Current Perspectives in Psychological Science 2,* 67–70.

Tulving, E. (1993b). Self-knowledge of an amnesic individual is represented abstractly. In T. Srull & R. Wyer (Eds.), *The mental representation of trait and autobiographical knowledge about the self:* Vol. 5. *Advances in social cognition.* Hillsdale, NJ: Erlbaum.

Tulving, E. (1995). Organization of memory: Quo vadis? In M. S. Gazzaniga (Ed.) *The cognitive neurosciences* (pp. 839–847). Cambridge, MA: MIT Press.

Tulving, E. (1998). Brain/mind correlates of human memory. In M. Sabourin & F. Craik (Eds.), *Advances in psychological science:* Vol. 2. *Biological and cognitive aspects.* Hove, Sussex: Erlbaum.

Wheeler, M., Stuss, D., & Tulving, E. (1995). Frontal lobe damage produces episodic memory impairment. *Journal of the International Neuropsychological Society 1,* 525–536.

White, S., & Pillemer, D. (1979). Childhood amnesia and the development of a socially accessible memory system. In J. Kihlstrom & F. Evans (Eds.), *Functional disorders of memory* (pp. 29–73). Hillsdale, NJ: Erlbaum.

Whiten, Andew W., & Byrne, Richard W. (1997). *Machiavellian intelligence II: Extensions and evaluations.* Cambridge: Cambridge University Press

Whitty, C., & Lewin, W. (1957). Vivid day-dreaming: An unusual form of confusion following anterior cingulectomy. *Brain 80,* 72–76.

Wilson, B., & Wearing, D. (1995). Prisoner of consciousness: A state of just awakening following herpes simplex encephalitis. In R. Campbell & M. Conway (Eds.), *Broken memories: Case studies in memory impairment* (pp. 14–30). Cambridge, MA: Blackwell.

Wright, J. C., & Mischel, W. (1988). Conditional hedges and the intuitive psychology of traits. *Journal of Personality and Social Psychology 55,* 454–469.