18

JOHN TOOBY, LEDA COSMIDES, & H. CLARK BARRETT

Resolving the Debate on Innate Ideas

Learnability Constraints and the Evolved Interpenetration of Motivational and Conceptual Functions

1 On the Sociological Need to Find Arguments That Are Effective as Well as True

> Plato says . . . that our "necessary ideas" arise from the preexistence of the soul, are not derivable from experience—read monkeys for preexistence.
>
> Charles Darwin, M Notebooks (entry 128)

In order for the study of the human mind and brain to become a successful natural science, a sufficiently large number of researchers must organize their research on the basis of theoretical commitments and methodologies that reflect, in broad outline, the realities of their object of study. Yet there has been, for over a century, enormous resistance to incorporating into the human sciences the most fundamental truth about the species they study: our functional, species-typical design is the organized product of ancestral natural selection (for discussion, see Pinker, 2002; Tooby & Cosmides, 1992; for opposing views, see Fodor, 2000; Gould, 1997a, b). The brain came into existence and acquired a functional organization to the extent that its arrangements acted as a computational system whose operations regulated the organism's behavior to promote propagation. Studying psychology and neuroscience without the analytical tools offered by evolutionary theory is like attempting to do physics without using mathematics. It may be possible, but the rationale for inflicting needless damage on our ability to understand the world is obscure.

We warmly thank Pascal Boyer, Peter Carruthers, Martin Daly, Steve Pinker, Dan Sperber, Steve Stich, Don Symons, and Margo Wilson and the participants in the Innateness Workshops for many illuminating conversations on these issues.

Why treat natural selection as central to psychology, neuroscience, and the human sciences? Why does it have a privileged organizational and explanatory role? Why is the neglect, peripheralization, or dismissal of natural selection in these sciences necessarily misguided? The reason inheres in what makes organisms (self-replicating physical systems) different from all other natural phenomena: organisms differ from other natural phenomena in that they manifest a profusion of thermodynamically improbable arrays of extraordinarily attuned interrelationships—states that are simultaneously highly ordered and highly functional (Dawkins, 1986; Schrödinger, 1944; Tooby & Cosmides, 1992; Tooby et al., 2003). This physically unspontaneous order would collapse in a fraction of a second were it not for the ceaseless operation of complexly engineered chemical and computational arrangements designed to combat the ubiquitous encroachments of entropy, in service of bringing about those narrowly targeted outcomes that facilitate propagation. To put it more simply, the second law of thermodynamics is the first law of psychology: functional order in organisms requires explanation (Tooby et al., 2003). This high level of functional organization is not a brute fact of the world, produced randomly or inexplicably. Instead, this functional organization has a known explanation, an explanation that is unique, well established, and beautifully principled. Physics and biology, considered together, inform us that natural selection is the only known natural process that pushes populations of organisms thermodynamically uphill into higher degrees of functional order, or counterbalances the otherwise inevitable increases in disorder that plague ordered systems. In other words, all complex (i.e., significantly better than random) functional organization in the designs of organisms traces back to the prior operation of natural selection, and must necessarily be explained in terms of it. Natural selection builds developmental adaptations into the designs of organisms, and the operation of these adaptations assembles each organism's functional machinery, and calibrates it to its circumstances. To use a nineteenth-century scientific idiom, it might be said that the second law of psychology is that ancestral natural selection is the cause of the functional order in brains and allied regulatory systems.

Psychology and neuroscience, if they are to be successful as sciences, must recognize, describe, and explain the functional order[1] to be found in minds and brains. Since this functional order derives uniquely from the evolutionary process, any accurate, theoretically principled psychology that humans might eventually build must inevitably become an evolutionarily centered science. The essential elements of this argument were clear in 1859 to Darwin, and are not hard to follow. Yet the 145 years since the publication of *The Origin of Species* have not seen the steady, linear growth of a reasoned appreciation of the Darwinian framework,

1. Of course, there are other characteristics of the evolved architectures of organisms in addition to their largely species-typical functional organization. These include transient and idiosyncratic features, as well as by-products of adaptations, which emerge as concomitants of the aspects of the architecture that have been selected (Tooby & Cosmides, 1992). There are indefinitely many by-products, because there are indefinitely many ways of describing organisms without making reference to their evolved adaptations.

especially in the psychological and behavioral sciences. In contrast, the value of far more conceptually taxing advances in quantum mechanics and relativity (not to mention the Newtonian revolution, or electrodynamics) were rapidly recognized, accepted, and disseminated throughout the relevant disciplines. Although there have been a few efforts to integrate Darwinism, these were generally followed by periods when evolutionary research went into near eclipse. Even at the high-water marks of these Darwinian infiltrations, evolutionarily informed psychology always remained a minority enterprise. To this day, evolutionary biology is not taught routinely, along with statistics and mathematics, as an indispensable element of professional training. Many researchers in the neural, psychological, and social sciences have only the vaguest idea about what is known in the evolutionary sciences and are often prey to lay mythology about Darwinism. Generally speaking, the biologists to whom nonbiological audiences are exposed are seen as both representative of biological thought and authoritative in proportion to their tendency to reassure their audiences of the fundamental irrelevance of Darwinism to the human sciences (Gould 1997a, b; Lewontin, 1998; Rose & Rose, 2000).

As a result of this strangely endemic resistance to Darwinism, the presuppositions of most of the research enterprises in the psychological and social sciences clash with the core nature of the phenomena they investigate. In consequence, over the last century efforts have been misdirected, results confused, and progress (where there is any)[2] made painfully slow (Tooby & Cosmides, 1992; Pinker, 2002). Otherwise gifted people advance and laboriously defend arguments whose obvious weakness they themselves would readily detect in other contexts (e.g., Chomsky, 1987; Fodor, 2000).[3] The rationalizations for peripheralizing Darwinism have impeded the emergence of a critical mass of researchers who appreciate its analytic centrality and inferential power. The institutional entrenchment of these rationalizations interferes with ordinary research and training, requiring responses that

2. Sociocultural anthropology, for example, has been moving backward for decades, and large segments of it are dead as a science.

3. Two striking examples are Fodor's argument that Darwinian conceptions of function are superfluous to building a functionalist cognitive science (Fodor, 2000) and Chomsky's argument that the understanding of language will be better elucidated "in molecular biology, in the study of what kinds of physical systems can develop under the conditions of life on earth and why, ultimately, because of physical principles," than through the analysis of the organizing effects of natural selection on cognitive architectures (Chomsky, 1987, p. 167).

That such arguments are advanced and defended by such typically strong thinkers is evidence of how unpalatable natural selection is even to leaders of the cognitive science community—a fact of considerable sociological importance.

Fodor (echoing many others) justifies his claim that natural selection is superfluous in the analysis of cognitive function by pointing out that the identification of the function of the heart preceded Darwin. That is, one can employ functionalist reasoning without being forced to traffic in unholy knowledge of evolutionary biology. In what other science would one find large numbers of people defending the use of a folk concept (common-sense function) in order to avoid the use of a well-established, technically rigorous, formally derived scientific concept (evolved function)—a concept, indeed, that connects cognitive science logically and empirically to the rest of the natural sciences?

siphon off much of the effort that would otherwise go into the progressive mapping of our evolved architecture.

Of greatest concern, the intellectual history of the last century makes it clear that a consensus that lacks good scientific justification can maintain itself through sociological processes for long periods of time, and perhaps indefinitely (Pinker, 2002; Tooby & Cosmides, 1992). This brings us to the heart of the issue to be addressed. The fact that resistance to Darwinism in the human sciences has been profound and enduring and yet not supported by an adequate scientific justification is significant. It makes it clear that we are not dealing with purely scientific objections that can be surmounted solely by addressing issues of logic and evidence. Instead we are confronted with a formidable practical problem in the sociology of science. If the intellectual ecology of the psychological, neural, and social sciences is to change for the better, it will be necessary to do more than come up with arguments of scientific merit. We need to find valid arguments that in addition have the potential to be sociologically successful. Revising one's set of scientific beliefs by getting rid of propositions that are inconsistent with facts from the evolutionary sciences is painful. We need to identify arguments that make the effort of adhering to poorly founded positions greater than the effort of correcting them. It is this problem that must be addressed and solved if our sciences are to move ahead. Where might such arguments be found?

We have folk notions of heat and temperature from boiling water, but through the use of concepts derived from thermodynamics, engineers can build (for example) power stations with tens of thousands of intricate, efficiency-promoting features that could not have been designed, manufactured, managed, or understood without the scientific concept of heat. Why have a kitchen science of psychology using folk function when we can have a vastly larger, far more rigorous genuine science? The architectures of animals are far more complexly engineered than any human-built system, so the correct idea of functionality (together with the long list of functions known to biologists) will be even more necessary for the understanding of humans. This is particularly true because the biological definitions of functionality that predict the principles of our construction often depart radically from folk notions, which often lead psychologists astray (see, e.g., the theory of intragenomic conflict, Cosmides & Tooby, 1981). Are incestuous desires evolutionarily functional or dysfunctional? What about jealousy, guilt, aggression, in-group favoritism, infanticide in langurs, within-family conflict, the perception of beauty, pregnancy sickness, mitochondrially induced pollen sterility, fever, avian siblingicide, play, and gestational diabetes? Biological theories of function provide clear and often quantitative criteria in these and hundreds of other cases that folk functionality has not and cannot.

In respect to Chomsky's argument, brief reflection reminds us that the number of designs for physically possible systems is vast beyond all possible analysis but is known to include circuits for honeybee dancing, web spinning, spotting nests suitable for brood parasitism by cowbirds, killing male rivals' still-nursing offspring in langurs, echolocation, bat detection in moths, throat targeting by wolves, reverse peristalsis, reciprocal blood regurgitation in vampire bats, nuptial gift analysis in insects, alarm call discrimination in vervets, copulation continuation after decapitation in the praying mantis, upstream salmon homing, sex change upon receipt of dominance information in the coral reef living wrasse, as well as every known human and nonhuman neural syndrome, impairment, developmental anomaly, and embryological experiment. In the absence of natural selection, physical principles are not a very plausible or significant source of information about why the language system has one set of computational properties rather than another. Again, it seems the kind of argument that is advanced more to deny the relevance of Darwinism than because there is any compelling affirmative case for it.

2 The Debate over Innate Ideas Is a Possible Turning Point in the Integration of Darwinism with the Human Sciences

One arena in which such progress might be made is over the fiercely contested claim that our reliably developing, species-typical neurocomputational architecture includes what would once have been called innate ideas (see, e.g., the attack on representational nativism in Elman et al., 1996). This is an argument that matters: if it became recognized that human minds are infused with content many of whose specifics are the downstream consequence of natural selection, this would require revision throughout psychology, neuroscience, and the social sciences. Of course, for most of the last century, the default position of most learning theorists, cognitive scientists, and neuroscientists has been that the neurocomputational mechanisms and developmental programs that operate on experience to produce mental content are primarily content independent and general purpose. On this view, such mechanisms have no content-like organization built into their structure nor do they introduce evolved content of their own into the mind. That is, they lack any neurocomputational implementations of innate ideas, such as evolved, reliably developing conceptual primitives, content-specialized inferential procedures, representational formats that impose contentful features on different inputs, domain-specific skeletal principles, or anything else that was designed by evolution to process inputs, throughputs, or outputs differently by virtue of their content. We will call any perspective that makes content-specificity exceptional or peripheral to the mind's evolved architecture a *blank slate view* (for discussion, see Cosmides & Tooby, 1987; Pinker, 2002; Tooby & Cosmides, 1992).[4]

For evolutionary psychologists, the blank slate view is both theoretically implausible (because a blank slate architecture would pointlessly and fatally handicap any animal so designed), and inconsistent with the comparative evidence (Cosmides & Tooby, 1987; Tooby & Cosmides, 1992). Darwin and subsequent evolutionary

4. Fodor uses some terms differently from the way we do, leading to some considerable confusion in the literature. For example, he writes in his critique of us that "poverty of the stimulus arguments militate for *innateness*, not for *modularity*. The domain-specificity and encapsulation of a cognitive mechanism on the one hand, and its innateness on the other, are orthogonal properties"; and "[y]ou can have perfectly general learning mechanisms that are born knowing a lot, and you can have fully encapsulated mechanisms (e.g., reflexes) that are literally present at birth" (2000, pp. 68–9). We certainly agree that innateness is a different dimension from information encapsulation (e.g., driving may be encapsulated but is not innate). To us, however, information encapsulation is also distinct from domain specificity (in context, we are almost always talking about evolved domain specificity). We have also used the term *modularity* to mean the tendency of biological systems to evolve functional specializations and the term *module* to refer to an evolved specialization, regardless of the degree to which it exists in a heavily policed informational quarantine or operates on information available to other procedures in the architecture. In this usage, we did not mean to invoke Fodor's particular and narrow concept of modularity, which appears to make information encapsulation a defining feature rather than (in our view) an occasional concomitant. In particular, we are suspicious of the encapsulation spatial metaphor of cognitive mechanisms being *containers* that act on the informational *objects* they hold *inside* of them. This produces spurious problems of how information trapped inside one container could manage to touch and so interact with information walled off inside another container (see, e.g., Barrett, in press).

researchers have investigated numerous species in which organisms display knowledge and competences that they did not acquire ontogenetically from any general-purpose, content-independent neurocomputational procedure (Cosmides & Tooby, 1987; Tooby & Cosmides, 1992; for specific examples, see e.g., Gallistel, 1990, 1995; Garcia & Koelling, 1966; Gaulin, 1995; Johnson & Morton, 1991; Mineka et al., 1984; see also Darwin, 1859, 1871). That is, many species develop knowledge that is either absent from the stimuli they have access to or is not uniquely entailed by it.

Natural selection provides an elegant, naturalistic explanation for the origin of such innate ideas, a point that Darwin himself realized shortly after developing the theory of natural selection (Darwin, 1974). In modern terms, mutations that cause neural machinery to reliably develop useful, world-reflecting mental contents (or organizing principles, categories, etc.) give their possessors a propagative advantage over blank slate designs that must consider an unconstrained set of possibilities, and are limited to applying the same procedures to all contents. Natural selection constitutes a second route, independent of the specific characteristics of individual experience, by which the mind might become endowed with knowledge, and endowed with the Kantian conceptual tools that shape and make use of experience in an evolutionarily functional way (for discussion, see e.g., Cosmides & Tooby, 1987; Tooby & Cosmides, 1992).

Hence, from an evolutionary perspective, a primary roadblock to progress has been the persisting consensus that general-purpose, content-independent mechanisms are the null hypothesis: that hypotheses about the existence of content-specific mechanisms are (in the words of Farah et al., 1996) "a priori implausible."

Although we cannot specifically remember using the phrase *massive modularity* ourselves, we are happy to endorse it, provided it is taken to be a claim that the number of evolved functional specializations in the brain (regardless of whether they are encapsulated) is substantially greater than has been traditionally believed—and not that there are no content-independent operations whatsoever, or that all mechanisms are informationally encapsulated with respect to all others. Finally, *general* has been used by scholars in a diversity of ways, but in the nativism debate we have used it to refer to evolved mechanisms that lack attributes that were added by natural selection because they work for some specialized domains but fail for others. Such systems require design features that activate them for the contents or inputs they evolved to work on, and deactivate them outside of their specific functional domain. In particular, general learning mechanisms will include content-independent computational procedures. Antinativists doubt (while we conclude) that there exist different evolved, content-dependent procedures specialized (in some way) for computing about mothers (Lieberman et al., 2003), predators (Barrett, 2005; Barrett et al., under review), coalitions (Kurzban et al., 2001), social exchanges (Cosmides, 1989; Cosmides & Tooby, 1989, 2005), and so on. For this reason, we find Fodor's statement that "[y]ou can have perfectly general learning mechanisms that are born knowing a lot" baffling. If the learning system knows a lot (e.g., that there are two sexes, that it had a mother, to avoid open running sores on others), then it cannot be as prepared to face one environment (where things that it knows are not true) as another environment (where everything it knows is true). The system is not general with respect to the set of possible environments and inputs it might receive, nor with respect to the kinds of environments it might have to act in. In our experience, antinativists express their antinativism through their belief in the explanatory adequacy for humans of mechanisms that are innately equipped with general-purpose, content-indendent procedures, arriving into the world free of any preexisting innate knowledge.

For the majority who hold this view, the only scientific problem worth addressing is choosing among models of content-independent processes, perhaps occasionally noting some content-sensitive exceptions of no general significance. This extreme Bayesian a priori skepticism deployed against reported architectural content specificity is the scientifically respectable face for its obverse—an extreme credulity extended to the sociologically preferred, blank slate alternative. If for many, as experience suggests, this theoretical precommitment floats free of the evidence, then no amount of contrary evidence by itself may be able to displace it. The remedy for this sociologically rooted epistemological problem must therefore be to change the scientific culture so that both kinds of explanations are put on an equal footing, subject to the same burdens of evidence, consistency, testability, economy, and predictive power. How might this be accomplished?

3 The Role of Learnability Analyses in Testing the Computational Sufficiency of Content-Independent Cognitive Architectures to Account for the Development of Competences

Chomsky, influenced in part by this independent Darwinian tradition, was the most prominent cognitive scientist of the modern era to attempt to relegitimize nativism, at least within the domain of language (Chomsky, 1957, 1959, 1965). Indeed, the history of the Chomskian enterprise is illuminating with respect to the problem at hand. Citing biological principles, Chomsky (1959, 1965) famously made poverty of the stimulus arguments about the acquisition of language—arguments that were modern applications of Darwin's reasoning about the emergence in individual development of knowledge and competences that were not wholly extracted from individual experience. These arguments gained substantial formal weight from subsequent learnability analyses (see, e.g., Pinker, 1984). A competence is *learnable* by a given computational architecture in a given environment if the architecture's procedures, in interaction with the structure of the environment, cause the development within the architecture of the competence in question (whether knowledge, skill, or regulatory structure). If the proposed procedures are not computationally sufficient to construct the competence with the given set of inputs, then they cannot be a correct model of the computational design. This form of analysis requires one to fully and explicitly characterize the procedures that constitute the set of acquisition mechanisms of the relevant features of the developmental environment, and the competence (or behavioral output) that develops. Candidate models for human learning and other cognitive mechanisms can be evaluated as computationally sufficient or not based on the following criteria:

1. *They should produce the set of competences humans actually acquire*. Examples include the ability to speak one's local language grammatically (Chomsky, 1965); the observed distribution of aversion intensities at the prospect of sex with family members (Lieberman et al., 2003, in press, under review); the ability to make correct inferences about predator-prey interactions (Barrett, 2005; Barrett et al., under review);

the observed, complementary patterns of insensitivity to and use of local social categories such as "race" in person representation (Hirschfeld, 1996; Kurzban et al., 2001); the scaling of punitive sentiment directed at free riders according to the magnitude of the individual's anticipated contribution to a collective action (Price et al., 2002); the ability to detect possible violations of social contracts in contexts of social exchange (Cosmides, 1989; Cosmides & Tooby, 1989, 1992, in press).

2. *They should refrain from producing those competences that humans fail to acquire*. For example, pattern associator architectures unguided by specializations predict the acquisition of large bodies of strange knowledge that real organisms, including humans, do not acquire (see, e.g., Marcus, 2002). More simply, content-free acquisition mechanisms should cause children in urban America to develop fears to local causes of injury and mortality, such as cars, stoves, and stairs. But these fears rarely develop, whereas fears concerning snakes, spiders, the dark, wild animals, and skeletons often do—even though they do not reflect local dangers (Maurer, 1965).
3. *Their success should not depend on the presence in the environment of properties that do not, in fact, exist* (e.g., specific forms of social instruction, reinforcement or feedback; the direct observation of unobservable things such as mental states; signals of the objective value of a goal-state).
4. *They should produce the patterns of individual and cultural uniformity and variation that are actually observed, using the observed distributions of environmental conditions as inputs.* For example, despite enormous cultural differences in rates of exposure to predator-prey interactions, the predator-prey inference system develops precociously and in parallel in different cultures—a fact that any acquisition mechanism must account for (Barrett, 2005; Barrett et al., under review).

These are very stringent requirements. Indeed, well-specified domain-general models reliably fail learnability tests for language (e.g., Pinker, 1979, 1984; Pinker & Prince, 1988; Wexler & Culicover, 1980). The scope and informativeness of this kind of argument can be greatly expanded, however, by considering tests of learnability and computational sufficiency for an entire range of problem-types that we know ancestral (and modern) foragers had to be able to solve in order to exist, survive, reproduce, and take advantage of their fitness-promoting opportunities (Cosmides & Tooby, 1987, 1989; Tooby & Cosmides, 1992). Learnability analyses for this broader set of competences can play a pivotal role in demonstrating that our species-typical cognitive architecture manifests an evolved, pervasive content sensitivity in its operation. There exist large sets of formally definable computational problems that humans routinely solve (and evolved to solve) that no content-independent architecture can solve, even in principle. To be worth considering as a viable candidate model for the human cognitive architecture, a domain-general model must generate the entire set of ancestrally necessary competences that human foragers (and

humans in general) manifest, without also generating nonexistent competences (Cosmides & Tooby, 1987; Tooby & Cosmides, 1992).

In the case of language, leaving aside the specific claims associated with particular models of language acquisition, we believe that Chomsky's arguments and Pinker's and others' learnability analyses logically demonstrated the need for positing some implementation of innate ideas that make possible the acquisition of language and, in particular, grammar. That is, the human neurocomputational architecture contains a language acquisition device in the form of a set of procedures at least some of which are language specific and whose embodied inferential strategies reflect structural or statistical regularities in the set of languages humans spoke ancestrally (as well as the contexts of meaning within which utterances were made). In our view, these Chomskian arguments should have established a scientific consensus that the blank slate viewpoint was mistaken, at least in the case of language.

However, despite the intellectual force of these arguments, and as influential as they have been in cognitive science, they have failed to bring about a consensus among psychologists, neuroscientists, and behavioral scientists of the kind one regularly sees in the other natural sciences. One possible reason for this failure is that the arguments over the acquisition of grammatical competence have become increasingly technical. The language system (whatever its nature) is very complex, making it difficult for researchers outside of language to arrive confidently at independent judgments of their value. This cannot, however, be the whole reason. After all, far more technical and counterintuitive theories were rapidly adopted in the quantum and relativistic revolutions. Another reason might be that Chomskian psycholinguistics, despite its various successes, has not clearly produced the step-by-step theoretical advances coupled to empirical demonstrations that aggregate into an ever-expanding circle of persuasively well-explained phenomena. Nevertheless, we think the key reason for the persistence of the debate lies in the fact that the grammatical patterns exhibited by human languages are widely believed to be objectively and publicly present in the world.

For the majority who are attracted to a blank slate view, the seemingly objective character of the learning task invites the perennial speculation that some presently unknown kind of cognitive architecture will be discovered that could detect such patterns without any assistance from computational machinery specialized for the task. Certainly substantial subcomponents of learning tasks appear to be tractable to content-independent operations such as pattern association, giving evidence of partial successes. Moreover, it takes a great deal of time and effort to explore the computational virtues and limitations of each new proposal, and for their explanatory deficiencies to become manifest (in the case of connectionism, see, e.g., Marcus et al., 1995; Marcus et al., 1992; Marcus, 2002; Pinker, 1999; Pinker & Prince, 1988). New variants on previously discredited approaches can be introduced at least as rapidly as they can be analyzed, especially if they contain large numbers of degrees of freedom that can be fitted to already gathered data (as is true of connectionist models). Most critically, it is impossible to show that unspecified models that might be developed in the indefinite future are computationally insufficient. The result in the scientific community has been a steady-state indeterminacy, where researchers continue to believe what they are disposed to believe,

and fractionate into self-reinforcing communities of belief. As the decades pass, it is difficult to escape the conclusion that whatever the virtues of the Chomskian enterprise (we think it has many), sociologically it has been ineffective in generally legitimizing proposals of functional content specificity and its evolutionary basis.

Poverty-of-the-stimulus arguments similar to Chomsky's have been outlined in cognitive development, where there is a vigorous and increasingly evolutionary subcommunity of cognitive nativists studying a larger and more diverse set of evolved functional specializations (Atran, 1990; Baillargeon, 1986; Baron-Cohen, 1995; Boyer, 2001; Hirschfeld, 1996; Hirschfeld & Gelman, 1994; Leslie, 1987; Markman, 1989; Spelke, 1990).[5] However, learnability arguments in these areas suffer from the same vulnerabilities that have made the Chomskian argument inconclusive: the knowledge that develops is widely seen as reflecting objectively true sets of relationships manifested in the world. Consequently, it is hard to convince blank slate advocates that no possible architecture of truth-discovery or relationship extraction would be able to account for the development of these competences without recourse to evolved content-dependent functional specializations (Quartz & Sejnowski, 1997). To solve our problem, we need to look elsewhere.

4 Hume's *Ought From Is* Barrier Poses a Set of Learnability Problems for Content-Independent Architectures, Which Are Insurmountable Whatever Their Implementation

If the Chomskian debate has not produced a consensus because the knowledge to be learned is (believed to be) objectively present in the world,[6] this suggests a strategy of argument that might be effective if its preconditions could be satisfied. If it can be shown that organisms need to acquire—and do develop—competences based on patterns that are not sensorily available or objectively present in the external world, then no possible blank slate learning architecture could acquire

5. We deeply admire the achievements of the cognitive development community (and consider ourselves part of it). Moreover, we appreciate the widespread understanding within this community of the need for biological constraints on induction. At least since Quine, the interaction on this issue between philosophy and cognitive development has been extraordinarily fruitful. What baffles us, however, is that when it comes time to go looking for biology to inform the investigation of biological constraints on induction, so many researchers in cognitive development go looking only in philosophy. Cognitive scientists need to mature beyond the point of regarding evolutionary biology as a stigmatizing contaminant.

6. Of course, Chomskians argue that the local grammar is not objectively present in the external world, because there are an infinite number of possible grammars that are consistent with any finite set of observed utterances. We agree with this, but we have observed that, sociologically, this argument is ineffective, and it leaves blank-slate researchers unpersuaded. It cannot be argued that evidence about the local grammar is unavailable in observable utterances. In the defense of anti-Chomskians, one could argue that whatever the architecture of general-purpose learning engines turns out to be, it could provide, incidentally as a by-product of its implementation, the necessary constraints on the hypothesis space that are, for Chomskians, supplied by the design features of the language acquisition device.

those competences or extract the requisite knowledge. Acquisition would require the presence in the evolved architecture of content-specific systems (innate ideas). The impossibility of learning things that are not objectively present in the world to be observed would demonstrate conclusively the reality of innate ideas, resolving the issue sociologically (we are optimists) as well as analytically.

Hume's argument (Hume, 1740/1978) that one cannot derive an *ought* from an *is* suggests one major class of competences fitting this precondition: motivational competences. Hume's argument generalizes to any psychological phenomenon that requires valuation to operate. From the point of view of the valuer, value is not a physical property, or a set of patterned relationships among entities in the external world, or an observer-independent property. Because the value of a behavioral outcome is not objectively present in the external world, it is absent from inputs to the sensory systems. Accordingly, mental representations of the value of a behavioral outcome cannot, even in principle, be learned through the operation of any content-independent procedures, including logical operations, pattern association, or inductive processes as traditionally conceived. If organisms have motivational systems and concepts that play an embedded role in them, then both motivational systems and the concepts they employ must be, at least in part, developmentally architecture derived. That is, regardless of what environmental features they are designed to take as inputs during development, motivational machinery and the core concepts they require must be assembled by specialized developmental programs designed by natural selection for that function.

No stimulus intrinsically mandates any response, or any value hierarchy of responses. In the tangled bank of coevolved organisms that Darwin memorably contemplated at the end of the *Origin*, naturally selected differences in the brains of different species cause them to treat the same objects in a rich and conflicting diversity of ways: the infant who is the object of caring attention by one organism is the object of predatory ambition by another, an ectoparasitic home to a third, and a barrier requiring effortful trajectory change to a fourth. It is the brains of these organisms that introduce behavior-regulatory valuation into the causal stream, and it was natural selection that introduced into brains the neural subsystems that

---

Indeed, during the initial emergence of language, prior to the evolution of any rich set of specializations to support it, the primary constraints on language learning would have to have been supplied by nonlanguage components of the cognitive architecture (whether specialized or general purpose). In the final analysis, it boils down to claims about what the evolved functions of the implicated machinery are. The choices are: (1) all aspects of the system used for language acquisition evolved for general cognition, producing language for free; (2) at least some parts of the system evolved for specialized functions, but none specifically for language; or (3) some parts of the system evolved for specialized functions, and some of these evolved specifically for language. Whatever the truth turns out to be about grammar, mechanisms for the acquisition of meaning could not be blank slate, or no one could ever learn language. Our interpretation of likely messages must be informed by a rich set of content-specialized mechanisms that tells us what someone is likely to be saying under given circumstances. If meaning were unconstrained and indeterminate, this process could not take place (see e.g., Markman, 1989; Sperber, 1996; Sperber & Wilson, 1995; Tooby & Cosmides, 1992).

accomplish valuation. The same stimulus set, by itself, cannot explain differences in the preferences and actions it provokes, or indeed, the preferences themselves. Value is not in the world, even for members of the same species. Members of the same species view the same objects differently: the very same object is one person's husband and another's father—an object of sexual preference in one case and sexual aversion in the other. Moreover, because each evolved organism is by design the center of its own unique web of valuations, evolved value by its nature cannot have an objective character (Cosmides & Tooby, 1981; Hamilton, 1964). Because of the structure of natural selection, social organisms are regularly in social conflict, so that the objective states of the world that are preferred by some are aversive or neutral to others (e.g., that this individual, and not that one, should get the contested food, mating opportunity, territory, parental effort, status, grooming, and so on). This gives value for organisms an intrinsically indexical quality. Indeed, fitness "interests"—the causal feedback conditions of gene frequency that value computation evolved to track—cannot be properly assigned to such a high-level entity as a person but are indexical to sets of genes inside the genome, defined in terms of their tendency to replicate under the same conditions (Cosmides & Tooby, 1981). Whatever else might be attainable by sense data and content-independent operations, value or its regulatory equivalents must be added by the architecture.

The architecture's evolved systems for assigning value and computing motivation were shaped by the relative fitness productivity of ancestral design variants, as matched against the set of evolutionarily recurrent choice problems. That is, content-specific value processing is done by mechanisms that ultimately were shaped according to whether their rankings and decisions were, on balance, reproduction-promoting under ancestral conditions. So value exists for animals solely because natural selection built neurocomputational circuitry into our minds to compute it as one of several kinds of representation necessary for regulating our behavior according to evolutionarily functional performance criteria.

The ramifications of integrating value into cognitive science will be far reaching because valuation is not a rare or peripheral neurocomputational activity. Valuation is cognitively *ubiquitous*. It goes on continuously, entering into the representation of almost all situations, and into the regulation of almost all behavior. Animals depend on motivational systems to assign tradeoffs, establish goal states, apportion effort, prepare plans, and trigger actions, assigning different kinds of valuation as a regular and necessary part of the generation of behavior. Valuation is intrinsically *content sensitive*. That is, valuation by its nature depends on discriminating situations from each other on the basis of their content. Predators but not prey must be avoided, substances with nutrients must be chosen over toxins or inorganic materials as food, offspring must be fed rather than eaten, fertile people as opposed to prereproductives or nonhumans courted, skills as opposed to eccentricities acquired, reliable as opposed to faithless cooperators preferred, free riding punished rather than rewarded, genetic relatives avoided rather than chosen as sex partners, injured legs favored rather than damaged further, role models attended to rather than ignored, friends cultivated, sexual rivals intimidated, coalitions formed, relatives assisted, and so on across an enormous range of ancestrally necessary and evolutionarily favored activities.

Valuation is intrinsically *content generative*: upon discriminating objects, situations, or prospects on the basis of their content, valuation intrinsically introduces its own proprietary forms of content into other representational structures. Persons, situations, objects, actions, and experiences are tagged as frightening, sexually attractive, appetizing, disgusting, dull, funny, glorious, grievous, embarrassing, beloved, horrifying, disturbing, shameful, fatiguing, irritating, fascinating, beautiful, fun, and so on (for an evolutionary-computational approach to the emotions and their relationship to motivation, see Cosmides & Tooby, 2000b; Tooby & Cosmides, 1990). Valuation processes and valuation ontologies are necessarily rich because of the large number of hetereogeneous mechanisms they need to orchestrate in preparation for action (e.g., flight, courtship, eating) and to recalibrate after action (e.g., guilt, shame, regret, satisfaction).

In short, many evolved motivational mechanisms, by virtue of the nature of the functions they serve, are necessarily functionally specialized rather than general purpose, are content dependent rather than content independent, introduce content not derived from the senses into the operation of the architecture, and do so ubiquitously.

The proprietary content introduced by the architecture constitutes a form of knowledge: the architecture must know (in some sense) that living children are better than dead children, social approval is better than disapproval, salt and sweet are better than acrid or putrefying, sex with your mother or father is to be avoided, helping siblings is (within certain tradeoffs) better than helping fungi, your mate copulating with your sexual rival is worse than his or her fidelity, spiders on your cheek are worse than in the garden, understanding is better than confusion, skill mastery is better than inept performance, and so on. Of course, the interaction of motivational systems with other cognitive activities occasioned by experience massively expands and enriches evaluative knowledge representations (e.g., from generalization along psychophysical dimensions; from the backward derivation of valuation of instrumentally useful intermediate steps to a primarily valued goal; for an analysis of various aesthetic activities as valuation processing, see Tooby & Cosmides, 2001). Nevertheless, there must be an irreducible core set of initial, evolved, architecture-derived content-specific valuation assignment procedures, or the system could not get started. The debate cannot sensibly be over the necessary existence of this core set. The real debate is over how large the core set must be, and what the proper computational description of these valuation procedures and their associated motivational circuitry is.

Valuation processes are often necessarily *domain specific* (Cosmides & Tooby, 1987): because the sets of outcomes that constitute biological success in some domains of adaptive problem are different from the sets of outcomes that are biologically successful in others, the same evolved definitions of success or valuation cannot be used to regulate action across them all. Indeed, this gives us a way of distinguishing evolved domains with respect to valuation and action regulation. The question is: can the criteria for valuation (or the criteria-deriving procedures) in two areas be developmentally derived from the same evolved core set? If the answer is "no," then two different evolved motivational domains are implicated. For example, humans do not and could not evaluate potential mates by using the

same criteria they use to evaluate foods or dangers or interactions with their children or projects for advancing their status. Nor is there any possible evolved core set from which such diverse definitions of valued outcomes or successful action in these five domains (for example) could be derived (Cosmides & Tooby, 1987). Different adaptive problems require different computational properties for their solution when reliance on the same properties would lead to functional incompatibilities and poor performance. To see this, consider designing a computational program that chooses foods based on their kindness or one that chooses friends on the basis of their flavor and the aggregate calories to be gained from consuming their flesh. This thought experiment suggests the kind of functional incompatibility issues that naturally sort motivational domains based on their incommensurability. Hence, by evolved design, different content domains activate different evolved criteria sets and evaluation procedures.

For those unused to thinking about the computational requirements for action, particularly as seen within an evolutionary framework, this argument will not seem as powerful as it is. After all, maybe humans do not solve motivational problems, or do so only very poorly. What sort of justification could there be in the endless parade of human folly for the claim that people are behaving functionally? Appreciating the argument from value computation depends on understanding that many species, including humans, are known to systematically perform substantially better than random in a growing number of well-studied domains, reaching narrow targets of evolutionarily defined behavioral success. This is what it means to say humans (and other species) are known to solve certain adaptive problems well. The very existence of individuals and populations depends on the ongoing successful computation of the answers to a range of value-dependent, action-regulatory problems to within very narrow tolerances. Although entropy is a formidable opponent, and our systems all break down sooner or later, animals on their passage through cycles of replication exhibit consistent, impressive, temporary triumphs over it. For example, the world is full of substances, but random selection of these, or random motor operations on these, will not prevent the organism from starving to death or poisoning itself. Courtship, mating, and parenting are far more complex. Explaining how this is regulated computationally is the task.

The study of motivational incommensurability gives us a method for setting an irreducible lower bound on the number of different evolved content-specific procedures or computational elements involved in valuation, as well as insight into their heterarchical organization into domains. (Of course, the actual number of evolved conceptual elements is likely to be larger because there are other kinds of computational advantages to content sensitivity than to serve as inputs to motivational operations). Cases of motivational incommensurability are numerous, and easily identified. Distinct and incommensurable evolved motivational principles exist for food, sexual attraction, parenting, kinship, incest avoidance, coalitions, disease avoidance, friendship, predators, provocations, snakes, spiders, habitats, safety, competitors, being observed, behavior when sick, certain categories of moral transgression, and scores of other entities, conditions, acts, and relationships. Consequently, evolved content specializations must also exist for these separate domains. (For the original versions of this argument, on why organisms

cannot evolve a general-purpose inclusive fitness-maximizing device, and so necessarily depend on at least some content-specific machinery, see Cosmides & Tooby, 1987; Tooby & Cosmides, 1992).

A motivational domain is a set of represented inputs, contents, objects, outcomes, or actions that a functionally specialized set of evaluative procedures was designed by evolution to act over (e.g., representations of foods, contaminants, animate dangers, people to emulate, potential retaliations to provocations). Not only is there an irreducible number of domains, but there is an irreducible set of domain-specific criteria or value-assigning procedures operating within each domain (e.g., for food: salt, sweet, bitter, sour, savory, fat affordances, putrefying smell avoidance, previous history with the aversion acquisition system, temporal tracking of health consequences by immune system, stage of pregnancy, boundaries on entities and properties considered by the system, perhaps maggot-ridden food avoidance, and scores of other factors). When the required assignments of value within a domain (such as food) cannot all be derived from a common neurocomputational procedure, then the number of motivational elements must necessarily be multiplied to account for the data.

The computational challenge with respect to motivation is to produce a set of programs that can duplicate human value-regulated behavior. As an important scientific goal, we need to begin the construction of an inventory of evolved value and choice criteria and procedures that are (in some way) built into our species-typical architectures, and of the evolved neurocomputational programs that derive, expand, and enrich them. To do this, we need to examine evolved valuation problems that humans can be shown to solve (or indeed any valuation-requiring behavior that humans are known to exhibit) and look at the set of valuation criteria that are needed to accomplish the task. We need to see how small the set of initial evolved value elements can be made that can still fully account for the data, being open to the parsimony considerations posed by the possible involvement of domain-general and domain-specific procedures for ontogenetically elaborating value criteria (e.g., the derivation of secondary reinforcers from primary reinforces by pattern associator systems). If it can be shown at any point that the so-far-identified derivation procedures (operating realistically in a naturally structured environment) cannot derive the required valuation-regulated behavior from the so-far-identified list of evolved value elements, then either new value elements should be added to the list to account for the new sets of behaviors to be explained or a new procedure must be added (whichever the data supports). So, for example, at present we are not compelled to posit a separate motivation for locomotion, because locomotion is instrumental to achieving other valued outcomes (although we do need to posit a value-based effort computation system that transduces locomotion, among other things, to explain why the same individual will walk 10 feet for a given reward but not 10 miles). Nevertheless, we do need to posit separate evolved motivational elements to account for sexual behavior and feeding behavior, because well-engineered choice in both these areas cannot be achieved by the same value criteria. Altogether, there has not been very much progress over the last century toward constructing such an inventory, because we have been shrugging off the issue of motivational innateness through the shell game of implying that any

given motivation is secondarily acquired, without obliging ourselves to computationally specify how and from what. The field needs to settle on a well-validated, irreducible set of motivational first movers. In our experience, a serious analysis of any domain often leads to the discovery that the irreducible minimum motivational feature set is surprisingly large (see, for an analysis of incest avoidance, Lieberman et al., 2003, in press, under review).

The outputs of these rich, indispensable systems of valuation computation are loosely referred to as *feeling*, saturating our experience with their voluminous, dense, intricate textures, and guiding our mental operations and bodies into fitness-enhancing realizations, choices, behaviors, and preparatory activities. They also deliver inputs to (but should not be confused with) a parallel, minimalist system of value distillation that produces a stripped down set of proprietary content that is used in certain aspects of decision-making. This subsidiary system provides the basis for intuitive and formal concepts such as utility, reward, payoff, and reinforcement. Why is this subsystem needed, in addition to the richer system it derives from? The realities of the physical world, the fact that we cannot be in two places at the same time, and the finite processing limitations on our neural circuits mean that many choices are necessarily mutually exclusive. In order to make choices in a way that usually promotes fitness, our architectures need to be able to discriminate alternative courses of action on the basis of computed indices of their probable fitness consequences. To serve this purpose, the minimum valuation-proprietary form of content is therefore a form of representational tagging with computed scalar utilities (or their equivalent) assigned to whatever representational parsing there is of goals, plans, situations, outcomes, or experiences. That is, the system must reliably develop so as to translate complex high-dimensional valuation representations involving rich content—such as *frightening* or *disgusting* or *irritating*—into unidimensional magnitudes. This is required so that situation-representations or sensory inputs can be ordered by payoff. Although the motivational system is far richer than just a utility computing system, we know this unidimensional neural currency must exist as one aspect of the motivational system, or the system could not be designed to make mutually exclusive choices nonreflexively in a way that tracked higher fitness payoffs. This form of payoff representation must be scalar so that magnitudes can be ordered, and should in addition have properties of a ratio scale so the computational system can arbitrate competing goals under different probability distributions. That is, this subsystem must be able to do more than ordinally rank outcomes, or it could not shift from one course of action to another upon discovering a shift in the probabilities of success among the alternatives (which common experience and conditioning studies show is routinely done).

Although only a small piece of the motivational system, this minimalist subsystem attracts disproportionate attention, and is often mistaken by certain research communities to constitute essentially the whole of motivation. This belief is seductive for researchers in fields like economics and learning theory because utility-style conceptualizations are easy to mathematically formalize and test. By focusing only on the question of what procedures would be needed to use pre-existing utilities to make choices, many researchers overlook the existence of the rest of the

motivational architecture that encompasses it. There is all too little research, for example, into the irreducibly complex input and processing systems needed to transform the entire universe of human experience and situation representation into payoff magnitudes. When their attention is drawn to the contrast, most researchers will admit that the rich universe of feeling cannot be captured by a set of flat, unidimensional utilities, and so utilities by themselves cannot be an adequate model of or explanation for this universe of valuation. It is time to move cognitive science into an exploration of this larger realm.

5 Evolved Systems for Motivational Computation Use Conceptual Structure in Targeted Ways, so Motivational Computation and Knowledge Computation Cannot Be Isolated from Each Other into Separate Systems

Valuation processes typically involve many of the same elements of conceptual structure that are the traditional objects of cognitive science (representations of persons, foods, objects, animals, actions, events). This means that the evolution of innate motivational elements will mandate the evolution of an irreducible set of conceptual elements as well. Why? A valuation is not meaningful or causally efficacious in the regulation of behavior unless it includes some form of specification of what is valued. That is, the specification of what the value applies to generally involves conceptual structure.

For example, for natural selection to cause safe distances from snakes to be preferred to closeness to snakes, it must build the recognition of snake-like entities into our neurocomputational architecture. This system of recognition and tagging operations is computationally a snake concept, albeit a skeletally specified one. Evidence supports the view that humans and related species do indeed have a valuation system specialized to respond to snakes (e.g., Marks, 1987; Mineka & Cook, 1993; Mineka et al., 1984; Yerkes & Yerkes, 1936). This one consideration alone forces us to add to a fourth innate idea to Kant's space, time, and causality. Yerkes and Yerkes's finding counts as empirically based philosophical progress, and as straightforward progress in the cognitive science of knowledge as well—derived (*pace* Fodor) from evolutionarily motivated theories of function.

In other words, the evolved motivation argument not only establishes the necessity of evolved motivational elements: it also resurrects the argument for the necessity of innate knowledge-like conceptual structure. Moreover, it does this in a way that is not vulnerable to the counterargument that objective knowledge (putatively) can be discovered by some general learner alone. This is because evolved conceptual structure is not present in the architecture (only) as "objective" knowledge. For the purposes of this argument, the elements of conceptual structure under discussion evolved to be in the architecture in order to be the object of intrinsically unlearnable motivational valuations. It is the specificity of the coupling to the particular valuation procedure that individuates the concept with respect to this set of motivational functions (e.g., [your children: beloved], [snakes: suspect]). Of course, although we think the neurodevelopmental basis of a lot of conceptual

structure was built in to the developmental programs by natural selection because it helped in computing accurate representations of evolutionarily important external relationships (see, e.g., Spelke, 1990), that is not the kind of selection pressure being discussed here. The requirements of motivation and actionselected for certain aspects of conceptual structure, and these aspects of conceptual structure may or may not be the same features of conceptual structure that were favored because they promoted the efficient acquisition of accurate representations of the world. (It seems extremely likely that conceptual structure was shaped by both sets of selection pressures.) In any case, conceptual elements (sexual rival) that evolved to serve motivational functions must be innately individuated by the way the motivational system distinguishes them for its operations (like jealousy).

That is, the evolution of content-discriminating motivational systems necessarily involves the evolution of crosscoupled, motivation-discriminated conceptual structure. Our evolved architecture is riddled with valuation processes, including (but not limited to) systems for generating, specifying, distinguishing, and ranking goal-states. To compute actions that differentially increase the probability of reaching a given goal, that goal-state (and action-relevant aspects of the situation the goal-state is embedded in) must be computationally definable, recognizable, and distinguishable from non-goal-states and alternative goal-states. More generally, if the successful functioning of an evolved adaptation requires a valuation process underivable from anything else, and if that valuation process requires the participation of a specific concept or category whose relationship to the rest of the valuation process cannot be derived, then the conceptual element must be, in some sense, innately (that is, evolutionarily) specified. You cannot systematically hit narrow targets unless there is a specification of the target. And in the realm of motivation, findings from evolutionary biology, behavioral ecology, and evolutionary psychology provide domain after domain where animals, including humans, efficiently hit the evolved targets that natural selection predicts they should.

For example, normally developing humans were naturally selected to have sex with healthy, reproductively mature members of the opposite sex (Symons, 1979). For a computational system to cause this, there must be evolved, reliably developing conceptual machinery that distinguishes human from nonhuman, male from female, mature from immature or senescent, healthy from unhealthy, live from dead (and so on) in order to assign one attribute higher valuation than the other. As one surveys the conceptual requirements of each motivational system about which there is evidence, the list of reliably developing, evolutionarily discriminated concepts becomes inescapably long. In traditional cognitive and philosophical terms, evolved motivational computation requires massive nativism.[7] Of course, this is not the claim that every adult value discrimination is innate. For example, if the representation of *healthy* gives *living* for free by derivation, then the

7. We use the terms *innate*, *nativism*, and so on because, given the discourse practices of philosophers and cognitive scientists, they are the closest counterpart to a more biologically elaborate concept. That is, while genetic determinism is an incoherent position, so is environmental determinism.

live versus *dead* distinction need not be a separately selected component of the motivational system (although this distinction might be important, for different reasons, in systems motivating behavior around potentially dangerous animals; Barrett & Behne, in press).

These representations need not be rich representations—neural and genetic economizing will mean that they will often be encoded using what can be called *minimal sufficient specification*. The minimal sufficient specification is the most economical cognitive machinery necessary for recognizing a representation by some evolutionarily constant feature it manifests neurodevelopmentally. The specification must tag representations so that the specific motivational operation will be able to find its proper objects. For example, adult concepts of male and female are undoubtedly very rich. Yet all the developing sexual valence circuit might need (in principle) is a single innately privileged psychophysical cue that causes males to be reliably distinguished from females, binarily indicating which is which for motivational purposes, with another binary parameter for setting the sex targeted for attraction. The sorting of tokens into types by the conceptual projections of the motivational system then allows a richer psychophysical template to be formed than is initially used, and conceptual enrichment to occur. (Evidence suggests, for example, that the historically contingent concept of *race* is a projection of a coalitional categorization system that evolved for sorting individuals into alliance sets; Kurzban et al., 2001.) The specific psychophysical (or other) cues that motivational systems use as inputs to accomplish the initial sorting of represented entities are expected to be minimal, subtle, strange, and abstractly contentful,

---

Everything develops from a jointly codetermined interaction among the genes, the environment, and the state of the organism at a given time. More precisely, in addition to zygotic organization, the organism inherits two sets of determinants rather than just one—a genetic inheritance and a less well conceptualized environmental inheritance (Tooby & Cosmides, 1990, 1992; Tooby et al., 2003). The environmental system of inheritance consists of the properties of the world that participate in the organism's development and life-processes and that persist from generation to generation. These two sets have been inherited together repeatedly across a number of generations. This repetition has allowed natural selection to coordinate the interaction of stably replicated genes with stably persisting environmental regularities, so that this web of interactions produces the reliable development of a highly organized, highly functional, and largely species-typical design. When we call something *innate*, we do not mean that it is "encoded entirely in the genes," that it is genetically determined, that it does not develop, that the environment played no role or a lesser role in its development, and so on—nothing real has those properties: not eyes, nor eye color, nor aortas, nor otoliths. What we mean is that it reliably develops across the species' normal range of environments. Reliable development (innateness) is caused by the interaction of the ancestrally coordinated set of environmental regularities and genetic regularities. We do not mean *present at birth* if by that one means *expressed at birth*. An innate feature could be the product of selection, a by-product of selection, or a property fixed by stochastic processes. In each of these cases, it is a regular part of the architecture of the organism. Regardless of whether something was itself selected, if it was a regular part of the architecture, it could have been a cause of selection. We are most interested in exploring innate functional organization, which is recognizable because it consists of reliably developing properties that are nonramdomly organized according to biologically functional engineering criteria: eyes see, and sexual jealousy interferes with one's mate's potential extrapair copulations, but the color of blood does not help it carry oxygen or nutrients.

compared to the richly elaborated adult representations we are familiar with. Of course, there is a balance between neural and genetic economy on the one side and worthwhile improvements in performance made through adding evolved criteria on the other. In the case of human sexual attraction, there is substantial evidence that the irreducible set of evolved criteria used and traded off against each other are complexly multidimensional (Buss, 1991) and not simply binary (or all members of each sex would be equivalently attractive).

Returning to our snake avoidance system, we can see it has a series of components. It has a psychophysical front end: one of its subcomponents assigns the evolved, internal tag *snake* through visual and biomechanical motion cues to a perceptual representation of some entity in the world. It has a second subcomponent that maps in a parameter *distance* between the *snake* and the valued entity (like *self* or *child*). Obviously, the distance-representing component is used by many systems. However, it also must have a component that assigns and updates different specific valuation intensities for different distances, so that further away is better than closer. The metric of valuation against distance (and its update rules) is proprietary to snakes, but the output value parameter it produces must be accessible to other systems (so that distance from snakes can be ranked against other goods, like getting closer in order to extract your child from the python's coils). Snake, distance, and the identity-distance valuation metric all necessarily operate together for this simple system to work. Snakes, the entity to be protected, and distance cannot be assigned to one computational process and valuation to another. Even in this simple example, conceptual and valuation functions indivisibly interpenetrate each other, with the representations necessarily coexisting within the same structure. As this form of analysis is applied to the other tasks humans perform, we think it will be impossible to escape the general conclusion that cognitive science intrinsically involves motivation, and the science of motivation intrinsically involves cognitive science. (Opposing views are not only implicit in the comparative neglect of motivation, except as a factor in learning, but are sometimes explicit. Fodor (2000), for example, considers the study of "cognitive"[8] processes and "conative" processes to be functionally separate, rather than coevolved aspects of the same unified systems of representation and action.)

The snake system also must interface with other shared systems for planning, situation representation, emotion, and action (e.g., systems that produce inferences that some potential actions represent improvements; that some potential outcomes are negatively valued; that motivate the choice of better outcomes over worse ones). The emotion system is particularly interrelated (Cosmides & Tooby, 2000b; Tooby & Cosmides, 1990). The function of the rich representation *frightening*

8. It is important to clarify that when we use the word *cognitive*, we intend it to be understood solely as a synonym for *information-processing* or *computational*, and not as an adjective that distinguishes say, thinking or knowing from feeling or acting. We are looking for cognitive—that is, computational—models of motivation and knowledge. We also use the word *representation* more loosely than most (e.g., as any computational product), because limiting it to knowledge-like structures with counterparts in the environment invites the acceptance of folk concepts and intuitions that we resist.

(as opposed to mere negative utility) is that in its associated emotion mode, fear orchestrates perception, hormones, the cardiopulmonary system, memory, and so on, so that they perform better, given the kinds of imminent action the architecture is likely to decide on and the long-term recalibration it derives from the event. (Emotions are conceptualized as evolved modes of operation of the entire psychological architecture, rather than a separate kind of mental activity.) The snake avoidance system also has another component. Although the details are not clear, it presumably recalibrates on the basis of individual experience, possibly slowly habituating in the absence of negative experiences or observations, and increasing sharply if snake contact leads to injury. It also narrowly accepts inputs from the social world—a conspecific expressing fear toward a snake (but not toward rabbits or other stimuli) in order to recalibrate the individual's snake valuation (Mineka & Cook, 1993; Mineka et al., 1984). Presumably this evolved because the system operates more functionally by upregulating or downregulating fear as a function of the local distribution of fear intensities in others, which index to some degree the local rate at which venomous snakes are encountered.

The key point here is that even this apparently simple one-function motivational system involves a series of evolved content-specific conceptual elements, including snakes, distance, conspecifics, that fear-faces have specific referents in the world, that snakes are a privileged referent of a fear-face (for snake fear to be recalibrated), and the output of fear itself. Of course, not all of these elements are unique to the snake system (although several are), but their pattern of distribution among motivational systems is heterarchical and itself not something that could be derived by content-independent operations acting on experience.

It is important to recognize that many kinds of motivational architectures are possible, not just ones that specify a single privileged goal-state and initiate means-ends inference. That structure seems an unlikely candidate, for example, for snake avoidance or sexual attraction. A particular bad event (like an imagined snake bite) need not be specifically represented as a negative goal-state in the snake avoidance system, with distance acquiring its significance through backward induction and means-ends analysis. More probably, the distance-fear relationship fills the representation of space with a motivational manifold that itself motivates avoidance (closeness is increasingly unpleasant). In the case of sex, it seems likely that the motivational system has a great deal of structure, with an evolved multidimensional path of motivational elicitation that intrinsically motivates many steps that guide the organism (foresightfully or not) to what is functionally (but not necessarily representationally) the goal-state. Computationally speaking, action-inviting affordances are not the same thing as represented goal-states.

The relevant question that will need to be addressed as the cognitive science project proceeds is how complex and how specifically detailed the architecturally derived motivational and conceptual machinery has to be to account for known, well-defined cases of human behavioral success. Computational explicitness, if insisted on, can play an important role in pushing cognitive science to deal more productively with the issues raised by the fact that the human neurocomputational architecture solves a large family of complex, distinct, evolutionarily recurrent adaptive problems. It is illuminating to try to map out a subsystem that can handle

even very simple, direct motivational phenomena. Such an attempt rapidly makes clear how much our intuitions hide the computational intricacy that underwrites the approximation humans achieve of evolutionarily adaptive valuation in their daily affairs. The requirement to build something program-like as opposed to labeling black boxes will awaken the field to the true magnitude of the scientific problems posed by motivation. It will correspondingly inhibit the tendency to imbue black boxes with magical powers.

The case of socially recalibrated intensities of snake avoidance show that natural selection can and does evolve procedures that accept social inputs when it is evolutionarily advantageous to do so. While the discussion of the machinery that underlies cultural phenomena lie beyond the scope of this chapter, we wish to warn against the casual acceptance of the widespread idea that social inputs processed by content-independent learning procedures are the primary explanation for the origin of human valuation. Here are a few reasons. Functionally well-calibrated valuation is indexical. What is good to value for some individuals is not good to value for others. Individuals are in daily social conflict over whose values prevail. (Because of inherent conflicts of interest in social species, a system that simply adopts others' values would be rapidly selected out. Others' values are processed [1] prudentially, in terms of the incentives they provide for the organism's own already-existing value system; and [2] as evaluated clues to what might lead to the best behavioral payoffs, given the individual's evolved meta-value criteria.) Although we cannot explore them here, there are insurmountable learnability barriers preventing the social acquisition of necessary values solely through content-independent procedures. For example, the courses of action the monitored individual did not choose and traded off against are invisible because they are counterfactual. Therefore, any observed course of action gives insufficient information from which to deduce the valuation systems of others. We only succeed at deducing some of the values of others because we share the same underlying sets of content-sensitive value systems, which allow us to know, a priori, what values others are likely to hold.

Even granting that some values could be acquired through content-independent processes operating on social inputs (which we dispute), the motivational unlearnability argument would continue to apply to the aspects of motivational systems whose parameters are not wholly accounted for by social information. It is easy to identify large numbers of these. For example, one major class involves valuations that develop independently of those held by others in the social group. The argument applies even more strongly for those values that develop in opposition to widely shared values, often eliciting strong negative sanctions from others. The idea that the child is a tape recorder passively absorbing values from others is easily contravened by ordinary experience: children resist foods urged by their parents; they resist treating objects valued by adults with the same care and reverence; they resist acquiring many skills valued by adults; most adolescents in religious and traditional schools notoriously do not adopt the urged or modeled values toward premarital sexual behavior. These and many similar observations lead us to the *social learnability test*: if it can be shown that the social world resists or fails to support certain motivations, then those motivations cannot have

been acquired from the social world. Many value-related phenomena meet the conditions for this argument. Indeed, humans ubiquitously pursue goals for which they are punished—and the development of valuation for these goals develops in spite of, and not because of, the existence of the social world.

6 The Evolved Function of the Cognitive Architecture Is the Generation of Biologically Successful Action Rather Than The Fixation of True Belief

Value and action have been relatively neglected by cognitive scientists because a commonly held view is that "the proper function of cognition is" (as Fodor puts it) "the fixation of true beliefs" (2000, p. 68). A consideration of the evolutionary dynamics acting on cognitive architectures shows that this view is at best incomplete, and more usually misleading. Before going further, however, it is important to point out that such a starting point, as self-evident as it may seem to be, commits us to a set of philosophical concepts that have no clear definitions in engineering terms. Whenever we are dealing with the designs of organisms, we are dealing with engineering questions. Philosophically, of course, it has proven extremely difficult to specify exactly what it means to call something a belief, to call a belief true, or to explicate reference, at least in an uncontroversial way. Although we seem to have clear intuitions about the meaning of such concepts as truth, knowledge, belief, representation, and reference, this may not be because they are what they seem to be. Indeed, a synthesis of evolution and computationalism suggests that these intuitions have led us away from a correct scientific understanding of the organic engineering phenomena they are used to represent. The situation may not be so different from what happened to many other equally irresistible intuitive concepts under the onslaught of modern physics (e.g., intuitions about space, time, causality, solid objects, and empty space bear little resemblance to the scientific concepts). We need to be prepared to have these venerable epistemological concepts transformed by our understanding of the nature, origin, and function of the computational systems that they inhabit as control elements. A quite different possibility is that they seem self-evident because they are conceptual primitives built into our cognitive architecture—as naturally selected Kantian *a priori*s, so to speak. These primitives are needed, for example, in theory-of-mind computations (Leslie, 1987) and in other scope-setting operations (Cosmides & Tooby, 2000a).

An alternative approach to their elucidation is to start out with engineering concepts drawn from biology, physics, and computer science. From there, the task is to see if it is possible (in principle) to build systems that have the same competences that animals (including humans) do. Once that is done, then it is possible to reexamine the architecture and its operation and see (1) what causally clear, well-described properties might serve as the evolutionarily tailored computational counterparts to our intuitive concepts of truth, belief, representation, reference, and so on; (2) how our engineering counterparts to these concepts might differ in certain key respects from their use in other accounts; and (3) the evolutionary-functional reasons why natural selection engineered reduced and transformed

versions of these concepts into our cognitive architectures as metarepresentational conceptual primitives (Cosmides & Tooby, 2000a).[9] Through this process, we might be able to get a fresh perspective on certain questions.

For animals, the accomplishment of sets of ancestral adaptive problems was enhanced by the evolution of behavior regulatory systems, which over evolutionary time coalesced into what, on histological grounds, is usually viewed as a single entity, the nervous system (as well as a few other architectural features, such as the endocrine system). The nervous system's functional identity is as a control system (or a set of control systems), analogous in many ways to control systems in manufacturing, robotics, engine design, architecture, and aviation. A control system is, by its very nature, a very different kind of thing from a scientist or a philosopher. Scientists and philosophers often stress the importance of arriving at true beliefs, while control systems exist solely to generate successful behavior. Correct action (action leading to successful propagation) is the functional product that the brain evolved to furnish, as disease resistance is the functional product of the immune system.[10]

For animals, knowledge only exists because ancestrally its production served as a means to correct action. Therefore, the designs of systems for the acquisition of knowledge in our architecture owe their functional organization to the evolved, systematic role they played ancestrally in regulating correct action. While this is sometimes acknowledged, less often explored are the downstream revisions this requires us to make in our thinking and scientific practice. The usual move is to argue that successful action self-evidently seems to depend on the attainment of true belief, so that the primary functional identity of the brain must be as a knower, a reasoner, and an acquirer of truths. Alternatively, some define cognition exclusively as knowledge-related mental operations, banishing by definition other operations from cognitive science. Either move justifies viewing the mission of cognitive science as primarily to explain the acquisition of knowledge (e.g., Fodor, 2000).

There have been a series of negative consequences for cognitive science that stem from its primary emphasis on knowledge acquisition. First, it assumes that at computational and neural levels, procedures for knowing are functionally separable from procedures for action regulation, and so can be successfully conceptualized and studied independently. We think that motivation, for the reasons discussed, shows that this is not the case. Second, it reduces the scope of cognitive science to a far smaller jurisdiction than what humans (and so human brains)

9. This project would require a book-length treatment, and in this chapter we can only offer a few remarks on the way to discussing motivational unlearnability. We do wish to warn the reader of our occasional departures from common accounts of truth, belief, reference, and representation; for further discussion, see, e.g., Cosmides and Tooby (1987, 2000); Tooby and Cosmides (1992).

10. Fodor (2000) dismisses this view because of its affinities with pragmatism. Pragmatism founders on the vagueness of its foundational standard: what *works*. In contrast, the engineering perspective of evolutionary functionalism is based on a very precise, formalizable concept: ancestrally, a systematic enhancement of successful design propagation.

actually do. From an evolutionary control theory perspective, there is not just a cognitive science of such things as language, intuitive physics, and number, but a cognitive science of parenting, eating, kinship, friendship, alliance, groups, mating, status, fighting, tools, minds, foraging, natural history, and scores of other ancient realms of human action. Third, it diverts cognitive scientists away from studying conceptual structure, motivation, and action as a single integrated system (which it seems likely to be), with motivation, in particular, in cognitive eclipse. Fourth, it ignores the many causal pathways whereby our evolved architecture should have been designed to manufacture, store, communicate, and act on the basis of representations that would not qualify as a rational architecture's efficient attempt at constructing true beliefs.[11] But the most intriguing reason to consider the implications of the brain as a control system is that it might give us better insight into what the phenomenon of knowledge is (i.e., insight into its ontology and engineering), as well as into the ontology of truth, belief, and representation.

7 Knowledge Is the Product of Evolutionarily Valid Inference and Came into Existence in Order to Serve as Potential Parameters for Biologically Successful Behavioral Regulation

From an evolutionary-functional perspective, knowledge is the total set of regulatory discriminations in the organism that allow its actions to be generated and adjusted so that they mesh successfully with the potentially variable features of its world. Of course, there are regulatory units in the genetic systems of bacteria that bind environmental factors (such as the lac operon) that qualify as embodying knowledge in this sense. However counterintuitive this engineering definition might initially seem to some, it becomes less so as regulatory problems get more complex and evolved regulatory systems get more sophisticated. As this happens, at least some sets of regulatory discriminations resemble more and more strongly our modern, intuitive conception of what knowledge ought to look like.

11. There are many evolutionary-functional reasons why "the fixation of true belief" is an inaccurate description of the goals or design criteria of the cognitive system, of which the following is a partial list. The first is discussed in the text: that values play an inextricable role in effectively setting truth criteria in systems engineered to take action that is designed to be successful (Neyman & Pearson, 1928, 1933). Leaving aside the necessary coparticipation of value in the definition of truth (discussed in the text), the existence of conflicts of interest in social life constitutes the source of many other deviations from truth-seeking as an engineered goal of all cognitive mechanisms. The system may be required to reason about value, and there is no truth of the matter for valuation. Individuals may adopt beliefs (e.g., God is three in one; Darwinism is irrelevant) because they socially coordinate them with others. The recomputation required to adopt the true belief may be too costly, at least for a period of time, so that temporary denial (as in grief) may be functional. The introduction of true information may be too disruptive to successful functioning, as when you choose not to look down when climbing a cliff face. To the extent a data store is computed for communication to others rather than to be acted on by oneself, then the optimal impact on others will be the criterion and not truth-value. The attribution of fault or blame to social rivals illustrates one of the many situations in which individuals may develop, disseminate, and "believe"—act as if—something is true that they have grounds for knowing is false.

In particular, many circuits for making discriminations in the service of action control will be indices that change in coordination with states of the external world. For example, one could imagine a binary neural register that is set to zero at night and one during the day, a register that evolved to regulate a single activity, such as sleep. Taken together, the parameter value of the index, and its location in the circuit structure, can for engineering purposes be called a representation, and its value constitutes a belief. From an engineering perspective, it is a true belief when it is successfully tracking the discriminated conditions that it evolved to parameterize. Representations are settings in a computational architecture designed to regulate behavior; they derive their existence and meaning from the causal properties of the architectures they inhabit. On this view, belief, truth, representation, and reference have a mechanism-relative, mechanism-anchored, and evolutionary function-specific character that delivers us from many of the puzzles that emerge when we attempt to make their character transcend mechanism (Cosmides & Tooby, 2000a; Tooby & Cosmides, 1992b; for kindred views see German & Leslie, 2000).[12] An indefinitely rich aspect of the external world such as night and day can be indexed to operation-defined parameters whose design is shaped to regulate a particular set of activities such as sleep or fear of leaving the concealment of one's home base. Operations on a belief do not have to be truth-preserving with respect to a superset of logical operations that might conjoin it with the total set of other beliefs in the system (assuming they were represented in such a way as to even make that possible). They only need to be success promoting within the scope of operations that regulates biologically significant behavior. What sets the definition this register uses for day and night are the engineering criteria built into its input and decision-making circuitry—that is, the circuitry that flips the register from one state to the other. These criteria will be set over evolutionary time by the relative fitness consequences of the various design variants made available by mutation. (I.e., it will hill-climb toward the variant that is "best" in the sense of producing the highest long-term fitness.) The register that results can be thought of as a "concept" of day versus night. The "meaning" of this concept can be explicated functionally and computationally in terms of the states of the world it evolved to track and, especially, the computational systems it evolved to interact with and regulate. Using this approach, one can isolate different

12. For some (but not for us), some kind of indexing of what a given representation is "about" (i.e., refers to or tracks) in the external world is diagnostic of representation. For a discussion of some of the functions of tags or representations about representations (metarepresentation), see Cosmides and Tooby (2000) and Leslie (1987). The evolution of a capacity to tag some representations with respect to a system of common reference serves at least one obvious function: it allows different kinds of information about cognitively defined environmental entities to be brought together as likely to be inferentially relevant to each other. In our view, the idea of reference is coherent not because it involves a relationship between a representation and the world but because it involves the coordination between at least two systems of representation (such as a perceptual parsing and predicted consequences of action made on the basis of that parsing), embedded in a system (or communicating community of systems) that can take action on the basis of these representations in a way that can be evaluated using some criterion of success (such as biological success).

components of meaning in an engineering sense. Loosely speaking, reference constitutes the states of the world that the register evolved to track. Another component of meaning is the set of input criteria that sets the value of the register. A third component of meaning is the set of action-regulating procedures that take as input the representation in the register. And a fourth component of meaning—what might be called *sense*—has to do with the set of inferences that can be made using the content of the register as an input.

Among more sophisticated organisms, it will usually be the case that action must be regulated by a space of discriminations that cannot be parameterized by mapping sensory inputs directly. Better kinds of actions could be orchestrated if unobservable states of the world could be determined through computation. What is this predator intending (Barrett, 2005; Barrett et al., under review)? What is the degree of genetic relatedness between this person and me (Lieberman et al., 2003, in press, under review)? Which coalition is this person likely to ally with (Kurzban et al., 2001)? Because the world repeatedly faced by members of a species over evolutionary time has a rich, stable, recurrent causal and statistical structure (the environment of evolutionary adaptedness), this problem can be evolutionarily solved by an additional process: *evolutionarily valid inference*. By inference, we mean the application of any neurocomputational procedure that uses some registers to set the value of other registers. We in no way mean to limit the structure of these procedures to the set normatively recognized in logic, statistical inference, and decision theory. Indeed, we think traditional inferential methods, to the extent they may be neurally realized within some representational systems, constitute only a small subset of the procedures embodied in the mind. Most inferential procedures will be what we have called *ecologically rational*—that is, they improve the performance of the animal because the structure of the inferential procedure reflects some enduring relationships in the structure of the world (Cosmides & Tooby, 1996; Gigerenzer et al., 1999; Shepard, 1984, 1987; Tooby & Cosmides, 1992b). Logically valid inferences are (within some representational systems) a small subset of evolutionarily valid inferences. An evolutionarily valid inference rule is any rule whose application produces (1) on average for a given species over its recent evolution, (2) within its proper cognitive domain, (3) a change in its set of computational parameters, so (4) the range of potential actions of the organism is adjusted, so that (5) they mesh with the potentially variable features of its world, with (6) greater biological success.

For example, among our mammalian ancestors, the female who nursed an infant was almost always its mother. This evolutionarily reliable statistical relationship meant that infant caretaking predicted genetic relatedness between mother and offspring, as well as relatedness among offspring cared for by the same mother. Another relatedness-predicting relationship ancestrally existed between the length of subadult coassociation and relatedness. Evidence supports the prediction that these enduring relationships in the world selected for a set of ecologically rational procedures specialized for inferring genetic relatedness. These evolved procedures take observations about the duration of coresidence and the existence of common caretaking as input, and transform them to set the values of a system of regulatory variables that evolved to track genetic relatedness between

individuals (Lieberman et al., 2003, in press, under review). This neurocomputational system of regulatory variables was selected for because these variables are used to (1) upregulate or downregulate tradeoffs between one's own welfare and that of kin, and (2) generate appropriate intensities of aversion to sex with genetic relatives (incest avoidance). We believe that these representations also influence (to some extent) the formation of explicit, linguistically accessible representations of kinship, but are not isomorphic with them. They are simultaneously and inseparably motivational and cognitive. They drive inferentially constructed plans. At least with respect to these two action systems (and perhaps to others), these regulatory variables represent "true" genetic relatedness. However, because the scope and fitness consequences for helping and for incest avoidance are different, the brain may represent two different (but related) values for genetic relatedness between a given pair of individuals. Each is "true" (functionally well calibrated), with respect to the action-regulatory system it inhabits, but they may be different. Females may, for example, represent individuals as more highly related for purposes of incest avoidance than as objects of altruism, because the asymmetric consequences of a miss versus a false alarm are different for incest avoidance and kin assistance.

Evolutionarily valid inferential procedures can exploit the fact that some relationships among elements of the ancestral world remained statistically true during the species' evolution. This means that the determination of the state of some variables allowed the probabilistic inference of the state of other variables, using procedures whose principles of transformation reflect these enduring relationships (i.e., if *i* nursed *j*, then set the register tracking the genetic degree of relatedness between individuals *i* and *j* to .499). These relationships need not be sensorily detectable or logically warranted, because architectures that build in the best Kantian a priori assumptions about unobservable relationships (embodied in procedures, data formats, etc.) outcompete others that lack such assumptions (Tooby & Cosmides, 1992a, 1992b). Moreover, we expect that there are many internal systems of representation (involving what Fodor would call central processes) that are not set simply or primarily by the immediate mapping of perceptual systems. They consist of libraries of operations and networks of representations linked by tags. These tags identify the inferential procedures that can operate on them. There are also tags to identify which evaluation procedures, decision-making procedures, differential memory operations, and so on can operate on them. These include a very rich set of evolved systems of conceptual structure, including many specialized systems for the construction of representations of persons, predators (Barrett et al., under review), minds (Baron-Cohen, 1995; Leslie, 1987), coalitions (Kurzban et al., 2001; Price et al., 2002), social interactions (Cosmides & Tooby, 1989, 1992, 2005), kinship (Lieberman et al., 2003, in press, under review), artifacts (Boyer, 2001; German & Barrett, in press; German & Johnson, 2002), and many other classes of entities. If an evolved action-regulation system regularly requires distinctions of a certain kind (cheater, predator, coalition member, gender, manipulable object, own child, mother), then specialized systems of representation tagging may evolve to provide the distinctions or create an evolved cognitive ontology (Boyer, 2001; Cosmides & Tooby, 1989, 1992; Kurzban et al., 2001). Indeed, valuation processes may play a significant role in defining certain ontological

domains (such as food, dangers, and exchanges) and the ontological affordances that invite domain-specific processes.

Selection should favor the evolution of ecologically rational procedures, concepts, and concept-generating systems on the basis of (1) how inferentially productive they are; (2) the degree to which they support informative distinctions in evolutionarily important valuation processes; (3) how easy it is to obtain relevant perceptual inputs (if these are required or useful); (4) how relevant they are to regulating important, evolutionarily recurrent activities for the organism; and (5) how naturally they can be derived from other reliably developing computational elements of the architecture. The aggregate effect of these functional criteria on shaping our cognitive architectures will make them look very different from what one would expect if knowledge acquisition alone were the criterion of functional performance. The developing picture is one of an evolutionary micro-Kantianism that shapes experience in far more detailed ways than giving form to space, time, and causality. These ecologically rational procedures pour experience into evolved, and often motivationally significant, categories such as *mother*, *predator*, *male*, *my child*, *coalition*, *domestic sharing unit*, *meat*, and so on. All together, these evolved procedures (and evolved metasystems for deriving procedures) constitute a very productive system for massively unpacking the fragmentary samples of perceptual and other inputs into a strongly structured set of representations of the world, and of the values of the actions that can be taken in it.[13]

8 The Computation of Truth Is Inextricably Bound to the Evolved and Computed Standards of Valuation Expressed in Our Evolved Architecture

The population of modern humans embodies neurocomputational architectures that acquired their engineering compromises from an immense series of encounters

13. There are many sources of input—initial parameter setting—in addition to sense data. For one thing, any somatic developmental interaction with the world could be used by natural selection to build a parameter setting system, and not just the senses as traditionally conceived. The organism may, in its developmental rules, be designed to assemble different computational settings on the basis of different nutrient flows, chemical exposures, endocrine levels, uterine environments, and so on—factors that provide another kind of grounding for inference aside from sense data. For example, a large number of regulatory parameter settings are unpacked from being on one of the two developmental pathways orchestrated by sex determination (i.e., organisms are often designed to think and choose differently depending on whether they are males or females). Moreover, the genetic material can itself receive signals when in the parent that are transmitted to the offspring and unpacked in the form of different developmental trajectories. This can happen, for example, through the setting and transmission of methylation patterns, piggy-backed on the outside of unchanged, inherited DNA sequences (Haig, 2002; Tooby et al., 2003). Third, there is nothing that would rule out knowledge from being inferentially developed from built-in premises and rules for their elaboration, whether or not at some processing stage it is admixed with inputs from the senses. Fourth, our species-typical endowment of evolutionarily valid inference procedures (which include the motivational assignment systems discussed earlier) can itself be viewed as an important kind of "input"—the introduction of content into our minds from the reliable development of the inherited design rather than from the senses.

that differentially preserved some design features and discarded others. This differential preservation was based on the degree to which they successfully solved recurrent ancestral adaptive problems in real, consequential environments. Our ancestors not only held beliefs (to use the folk concept) but acted on them, and the relative propagative success of those actions built some procedures for belief acquisition at the expense of others. Since the pioneering work of Neyman and Pearson (1928, 1933) it has been clear that for systems designed to realize values through making decisions that lead to actions, the optimal criteria for truth determination sensitively depend on the values the system is designed to realize (Tooby & Cosmides, 1990, 1992a).

Signal detection theory with its hits, misses, false alarms, and correct rejections, for example, is a well-known and straightforward application of Neyman-Pearsonian decision theory, in which the values of the four outcomes must be computed to set the threshold criterion for when to decide the signal has been detected. Since representational systems evolved as input parameters into action systems, the need to integrate value weighting into "truth" criteria would necessarily have ramified through every aspect of our cognitive architecture. Consider a simple dichotomous case (Tooby & Cosmides, 1990): the shortest path to walk to a destination would take a hominid under the overhanging branches of tree. There is either a leopard in the tree or there is no leopard in the tree. There are different payoffs to the four possible outcomes defined by act and state of the world: the hominid avoids walking under the tree, and there was a leopard in the tree (hit); the hominid avoids walking under the tree, and there was no leopard in the tree (a false alarm); the hominid walks under the tree, and there is no leopard in the tree (a correct rejection); and the hominid walks under the tree, and there is a leopard in the tree (a fatal miss). The cost of a leopard attack is large (death); the benefit of walking a straight line is a few calories saved. The best strategy for the choice system (its truth setting for the purpose of action regulation) is to act as if the leopard is in the tree, even if in 999 times out of 1,000 it is not. On the other hand, if a group of hominids were hunting a leopard, they might not even bother to look in an unpromising tree that, under identical information conditions but with different purposes, each individually would have avoided for possibly harboring a leopard. Similar shifts in truth criteria can be expected in making judgments about whether a predator is dead or merely asleep, for example (Barrett & Behne, in press).

The coevolutionary dependence of truth standards on value applies to every component of our evolved neurocomputational architecture. The design of every system should have been impacted by this relationship. Because knowledge acquisition systems evolved to form the basis of action, the kinds of actions the system has evolved to engage in will build in different procedures for establishing truth criteria for different kinds of functions. This kind of Neyman-Pearsonian value shift is why genetic relatedness representation may effectively fractionate in its downstream passage to the incest avoidance system and to the kin-assistance system. Wherever there has been an evolutionarily recurrent relationship between a kind of knowledge to be acquired and the kinds of uses to which it is put, there is the possibility that natural selection has introduced procedures for calibrating

differentiated sets of truth criteria. What the criteria for truth ought to be for an engineered cognitive system cannot be determined in the absence of value criteria. Even logical operations, which are supposed to be perfectly truth-preserving, cannot be trusted to give true conclusions in engineered systems, because the correspondence between the representations in the architecture and the conditions in the world they supposedly index cannot be made operationally perfect. There is always some possibility that a valid transformation will produce a conclusion outside of the scope within which the representational system evolved to work. Our architectures may be designed to disregard such logically valid conclusions, when they can be detected.

In the area of knowledge acquisition, value may play a more significant role than simply triggering occasions and activities within which knowledge is acquired. The motivational architecture may be constitutive of the organization and acquisition of children's knowledge, shaping or creating principles of knowledge acquisition. To take one out of many possible examples, valuation procedures may play an important role in setting the boundaries of concepts, shifting to some extent our understanding of prototypicality effects. To begin with the familiar, the perceived world "is not an unstructured total set of equiprobable co-occurring attributes" (Rosch, 1978, p. 29); it has a correlational structure. Attributes come in clusters: objects that share many properties—prototypical items—are information rich clusters of attributes. For prototypic items, knowing one property allows one to predict the presence of many other properties. Rosch argued that our cognitive architecture is designed to detect the correlational structure in the perceived world and produce categories that mirror it: categories with a family resemblance structure. Prototypes are "just those members of a category that most reflect the redundancy structure of the category as a whole" (p. 37). This is one clear area where domain-general learning procedures can produce a large and valuable set of data structures (although domain-specific skeletal organizing principles play at least as big a role in conceptual structure [Gelman, 1990]). Roschian prototype effects have been one experimentally validated theory for explaining perplexities that arise from considering instances where classical definitions of concepts conflict with people's intuitions: for example: Is the pope a bachelor? Was Jesus? Is a eunuch? An infant boy? A homosexual male? However, whereas correlated attributes may explain some aspects of the rapidly fading concept of bachelor, it is possible that conceptual projections of valuation procedures are another. That is, concepts may be generated, and their properties partially determined, by a calculus of the value their constituent criteria play in predicting the value of the instance for regulating behavior. If a major, socially shared function of the concept of *bachelor* is to make inferences about potential marriage partners, then other criteria contributing to this function may be imported into the concept in addition to the most probabilistically informative threshold tests organizing the concept (being male and unmarried). These may also lead to patterns of exclusion or peripheralization of instances with low value for the contemplated activity (the pope, a child, etc.). This is a different explanation for prototypicality judgments from those that emphasize instances that "most reflect the redundancy structure of the category as a whole." At least in Austen's world of *Pride and Prejudice*, more

attractive men and more prosperous men would be judged more prototypic, even though their attributes are rarer. Their use in choice and goal-state setting would explain the tendency of prototypic representations of instances to incorporate aspects of the ideal (based on valuation) rather than simply correlated attributes (based on frequencies). In addition, value criteria should play a role in defining the boundaries of the category over which correlations of attributes are computed. For example, there is no logical reason why early fruiting bodies should not count as fruit, but they are so distant from being edible that they are not considered instances that help to define the category. An experimental program to test this approach would see whether the internal structure of concepts reflected not only correlated attributes but also value criteria rendering them more or less valuable for the actions the category supports. Both ought to be present in stabilizing the meaning and boundaries of categories. Frequency-defined attributes are inferentially powerful; value-diagnostic attributes are motivationally informative. A typical prediction would be that (for example, in the case of fruit) prototype effects would only be partially accounted for by statistical frequencies of attributes, with prototypes shifted in the direction of increasing value. Rotten, unripe, or otherwise inedible fruit would not be considered central to the category even when their ecological frequency is greater (as it usually is). Indeed, the concept of *fruit* may be something like: any fruiting body whose appearance warrants further investigation as potentially edible enough in the near future to be worth harvesting.

9 Conclusion

We are not making any claims about information encapsulation. We are not claiming that all elements of each computational adaptation evolved from "the beginning" for the functions they presently serve. We are not claiming that, for example, all of the functional elements used for the operation of the snake avoidance motivational adaptation are unique to the snake avoidance system. We are not claiming that there are no general mechanisms for motivation. We are not claiming that the environment plays no role in the development of these systems, or that evolved systems operate the same way regardless of developmental environment. We do think that each adaptation is a collection of elements many of which are shared in different configurations among adaptations, some of them quite broadly. The specialization of an adaptation for a function does not lie in the specialization of all parts to its function. The specialization lies in the way the particular interrelationship of the parts is coordinated to solve the specialized adaptive problem with particular efficiency. This may require the evolved introduction of only a single new element into the evolved developmental programs—a minimal sufficient specification, for example, that can individuate an additional proper object of a certain class of motivations or inferences.

We are claiming that (1) an initial, irreducible set of category-recognizing, value-assigning, and value-responsive procedures must be built into our species-typical set of developmental programs; that (2) every evolved motivational system must have evolved conceptual machinery to express its necessary set of evaluative distinctions (e.g., in the case of sexual attraction, tags that distinguish the

representational identities of adult from child, male from female, human from nonhuman, healthy from unhealthy); that (3) such evolved conceptual elements are numerous; because (4) the rules required for regulating action and assigning value will necessarily be different for each adaptive problem domain in which the criteria of biological success are functionally incompatible (e.g., you necessarily pick the best available mate by different criteria from those for picking the best food, the safest refuge, or the neediest child); that (5) many of these evolved elements will be by their nature functionally specialized, content sensitive, domain specific, and content generative; and that (6) the architecture operates jointly on values and representations of states of affairs within the same computational systems, so that knowledge-representing cognitive processes often cannot be intelligibly separated from motivational processes. More generally, the claim is that successful performance on value-related adaptive problems poses an insurmountable *ought from is* learnability barrier that cannot be crossed, even in principle, by content-independent learning architectures, whatever their implementation. Given data about which valuation problems humans solve, this is a method not only for demonstrating the general case for innate ideas but also for identifying specific sets of such computational elements.