

---

## Introduction

### 1 The Theme

The theme of this essay is rather simple, though its demonstration is not. It is that humans think reflexively or metamentally *because*—and often *in the forms* in which—they interpret each other. In this essay ‘metamental’ means ‘about mental’, and ‘reflexive mind’ means ‘a mind thinking about its own thoughts.’ To think reflexively or metamentally is to think about one’s thoughts deliberately and explicitly, as in my thinking that my current thoughts about metamentation are right. Thinking about thoughts requires understanding thoughts as thoughts, as mental structures that represent; it also requires an ability to relate thoughts to other thoughts and to recognize such interthought relations. Since metamentation is essential to and uniquely distinctive of human minds, the idea that it originates in interpreting other minds can be encapsulated in the slogan that minds are minded because minds mind minds. This word play translates thus: minds evolve into reflexive minds because they mind other minds—where ‘minding other minds’ means interacting and bonding with other minds, being concerned or curious about them, representing their relations to the world, manipulating and using these relations for some purpose, and the like. All of this amounts (in my terminology) to *interpreting* other minds in social contexts of cooperation, communication, education, politics, and so on. It follows that *intermental* relations among individuals, handled by a distinct competence for interpretation, are essential to the evolution of abilities to represent the *intramental* relations among thoughts typical of a reflexive mind.

I take ‘interpretation’ to be a convenient, short, and grammatically flexible label for what is known in philosophy as commonsense or folk psychology and in psychology as theory of mind, mindreading, or naive psychology. Interpreting is a cognitive rapport between an interpreter (she, in this book) and a subject (he), whereby she represents his mind-world relations from the

simplest, such as seeing or wanting, to complex propositional attitudes, such as desiring, believing, or intending, and factors these representations into her goal policies and strategies for action. Interpretation first evolved by natural selection to enable such factoring, thereby promoting the biological interests of the interpreter. So construed, interpretation was naturally selected among primates as a battery of practical skills that precede language and advanced thinking by a long evolutionary shot. In human ontogenesis the grip of natural selection gradually weakens and is replaced by forces of culture, whose grip on the mind may nevertheless be as universal and coercive as that of nature. This means that the emergence of mental reflexivity out of interpretation (and other enabling factors) should be understood as both evolution by natural selection (in early phylogenetic and ontogenetic stages) *and* development under cultural constraints (in later ontogenetic stages).

The idea that metamentation evolved out of interpretation is an exaggeration but not an extravagant one. It is an exaggeration because interpretation is not the sole reason for, and not the sole designer of, metamentation. Language is also a key player, although its mastery owes much to interpretation. It is an exaggeration also because the metamentation indebted to interpretation need not be all the metamentation there is. There are forms of thinking about pictures and sentences that may approximate reflexivity without interpretation, or without much of it, as the case of intelligent autism suggests. More important, the idea that reflexivity evolved out of interpretation is an exaggeration because also involved, crucially, were the abilities to pursue goals by imagining, planning, and solving problems—in short, mental advance work or mental rehearsal. Indeed, I will argue that metamentation begins as *interpretation mentally rehearsed*. Interpretation and mental rehearsal are thus the two pillars on which rests the construction of the primate mind and of its upper metamental floors in particular. A third pillar, equally vital, originates in a mutual physiological regulation between human infants and mothers and soon takes the form of comment-topic protoconversation or topical predication, as I will call it. This is a human development that moves interpretation from its earlier and narrower subject-world focus to a new triangular mind-world-mind pattern of mental sharing in which two individuals (interpreter and subject) share attitudes and information about items of common interest. It is in this pattern of mental sharing that interpretation conspires with mental rehearsal to develop metamentation. Soon after its emergence out of mutual regulation, topical predication is absorbed into communication with and interpretation of others. This is why, for all practical purposes, this third pillar will be counted here as part of interpretation.

Yet the notion that metamentation evolved out of interpretation is not extravagant, because the basic skills needed to think explicitly about or in

terms of other thoughts could not have emerged from any source or cognitive ability other than interpretation and can actually be reliably traced back to it both in primate evolution and child development. The patterns required for metamentation can be found solely in the domain of interpretation and are intelligible only as tasks of interpretation. Along with language, mental rehearsal brings these patterns inside the mind and makes them explicit objects of representation and manipulation. This is what this essay endeavors to demonstrate. The demonstration has a motivation worth making explicit, since it may run counter to prevailing views on the relation between mind and interpretation.

## 2 Motivation

In an earlier work I developed an evolutionary account of interpretation as a practically motivated adaptation (Bogdan 1997). That work left me with (at least) one puzzle that this essay tries to solve. The puzzle grew out of a familiar but troubling observation, which is that figuring out and explaining what people think and do is what interpretation does well and cognitive science doesn't, at least not yet. Hence the oft-heard proposal that cognitive science should tap the folk wisdom of interpretation for a better understanding of the mind. If the assumption behind this proposal is that interpreters have a tacit, naive but largely true knowledge of mental architectures (programs and functional mechanisms) that cognitive science lacks, then the strategy of tapping folk wisdom is a nonstarter. I argued elsewhere that the naive knowledge interpreters have of minds seems indifferent to and silent about mental architectures (Bogdan 1985, 1991b, 1993, 1994), for good evolutionary reasons (Bogdan 1997). Yet there is something else to tap in interpretation to get a scientific grip on human mentation. It is the decisive role that interpretation played in the evolution of primate minds. It is a historical fact (documented below) that interpretation coevolved with primate mentation, and there is a growing body of theory and evidence indicating that interpretation may have been heavily implicated in the design of many faculties of the primate mind. A study of that coevolution could then guide and enrich the understanding of primate minds. This essay pursues that promissory note in the narrow but crucial area of reflexive thinking. And it does so as follows.

## 3 Lines of Argument

The demonstration will run along three converging lines. The convergence is crucial because no single line could carry the whole weight of the thesis. One line is *conceptual*. It explores the structural similarity, and at times

isomorphism, between the tasks of interpretation and those of metamentation. To illuminate this conceptual parallel, I distinguish several key metamental tasks and analyze them in terms of categories and schemes required to handle the tasks. These categories and schemes turn out to represent objects of interpretation at different evolutionary stages. Hence the second, *evolutionary* line of argument. Its thrust is that interpretation is the chief model or blueprint for the categories and schemes of reflexive thinking. To make the idea plausible and biologically ground it, I begin with the background hypothesis, enjoying growing though not unchallenged influence, that primate social life was more apt—and more likely than foraging, tool use or other mechanical activities, directed at the physical world—to have fueled and molded the evolution of primate minds. Since primate social life selected for interpretation, more than for anything else, interpretation emerged as the mental activity most effective in the evolution of primate mentation. The phylogenetic and ontogenetic record favors this diagnosis, since it identifies the pressures for interpretive know-how as the strongest during primate childhood and suggests that metamental skills correlate consistently with skills for interpretation.

This brings in the third, *psychological* line of argument. Although drawing on some interspecies comparisons, the psychological story told below is mostly human, mostly developmental, and focused on checking when, how, and why advances in child interpretation precede, link up with, and facilitate, if not cause advances in, metamentation. Besides joining independent sources of evidence, the link between evolution and psychology has a further significance in this essay. The evolutionary debate over the primacy of physical work versus social life in driving primate mentation has a psychological echo in the developmental debate over the primacy of mechanical action (Piaget) versus social interaction (Vygotsky) in the formation of the child's mind. I think the echo is not fortuitous, not because phylogeny would recapitulate ontogeny, but because of a significant correlation (explored in chapter 3) between the unusually long and adult-dependent human childhood and the unique mind that results. This mental uniqueness seems to owe a good deal to the equally unique texture of the social and cultural surround in which human kids grow up and mature their mental faculties.

#### **4 Level of Analysis**

These converging lines of analysis will be pitched mostly at the level of the *tasks* executed (what is done) in interpretation and metamentation, and will remain silent about programs and brain mechanisms (how it is done). I often talk of programs or skills but think of them in terms of their tasks, not in terms of the nuts and bolts of their operation. The focal thesis—that metamen-

tation evolved out of the interpretation at work in mental rehearsal—should therefore be understood and judged in terms of tasks: it is the tasks of metamentation, however executed, that emulated those of interpretation. Pitching an evolutionary analysis at the level of tasks may look controversial and risky in the light of the widely shared belief that evolution selects for programs or mechanisms. This is true of the targets of selection but not of the *reasons* for selection. Selection is for *what* programs or mechanisms *do* that results in reproductive fitness, and what they do can be aptly and fruitfully analyzed in terms of tasks. I find a task analysis apt because evolutionary biology is a science of functions, in particular of functions that become adaptations, and adaptations can be fruitfully described in terms of tasks. I also find a task analysis apt because cognitive scientists often discern tasks before figuring out the underlying programs and mechanisms (as happened, for example, in the cases of grammar and vision). I think that the current understanding of interpretation and metamentation is at such a stage.

Another methodological choice needs to be noted. The demonstration attempted by bringing together data and arguments from evolution, development, and conceptual analysis is going to be inductive or rather detective, as it looks for a variety of clues that reveal patterns of interpretive tasks that metamorphosed into forms of metamentation. Although interdisciplinary in scope and indebted to empirical data, the demonstration is largely theoretical and often speculative. Scientists also speculate, often boldly, particularly in fields such as those covered in this book, but their speculations tend to be narrow, constrained leaps from domain-specific data, accepted theories, and other authoritative sources. (So they cite a lot.) My sort of speculation is less domain-specific, more global and integrative, more philosophical, as it looks for patterns and connections often lacking firm and narrow empirical moorings. (So I cite less. Often whole pages may go by without a citation. Sorry about that.) Yet this sort of speculation is worth pursuing and may yield benefits because the current understanding of the reflexive mind, still limited and fragmented, is unlikely to emerge from any single, compartmentalized precinct of cognitive science. The reflexive mind is a hard puzzle, one of the hardest, precisely because it may be the outcome of independent developments somehow strung together by interpretation. The story of this outcome will unfold as follows.

## 5 Plan

The essay is divided in two parts. The first sketches the evolutionary background of interpretation as stimulus and shaper of metamentation, the second charts key moments of this coevolutionary saga in terms of a comparative task

analysis. Chapter 1 argues that the phylogenetic and ontogenetic routes to metamentation begin in primate social life and the minds adapted to it. Among sundry kinds of socialized minds, only the human mind has the potential to turn reflexive. Why? Because of *how* it socializes. It is mind socialization through internalization of interpersonal relations during childhood. This hypothesis, first proposed by Lev Vygotsky, goes in the right direction but not far enough. This diagnosis sets the stage for chapter 2, where interpretation is found to be the missing link in the Vygotsky's story. Interpretation has solid and far-reaching evolutionary credentials among primates and is systematically implicated in the evolution of primate mentation in general and of thinking in particular. This later implication serves as a launching pad for metamentation, roughly as follows.

Social primates interact by generating and exploiting causal relations among themselves. So they must represent social causation under appropriate categories and schemes. Since interactions among primates are handled by interpretation, primate causal knowledge is represented under interpretive categories and schemes of subject-world relations. Mental rehearsal of social action involves manipulation of causal representations of subject-world relations. These representations are projected imaginatively and often calculated off-line. At some point late in human childhood, when the process is applied to one-self, conditions become ripe for developing categories and schemes for coding and mixing other-world and self-world relations as mental representations. Metamentation is just around the ontogenetic corner.

These developments occur only in the minds of human children and are responsible for the uniqueness of the resulting adult minds. Why? Chapter 3 argues that the answer should be sought in development itself. The human mind is unique because so is its development. Primate development is special in being very slow and adult-dependent, but its human version also involves a unique biophysiological regulation between infant and mother that grounds a give-and-take form of sentimental bonding and communication of emotions and experiences. Such sentimental bonding forms the basis for a truly novel ability, topical predication, which interpretation uses to design communication by shared meaning, language acquisition, and eventually metamentation. Thus concludes the first part of the essay.

The next four chapters chart the developmental progression from sentimental bonding to metamentation in conceptual, evolutionary, and psychological terms. Chapter 4 sets the stage by providing a conceptual profile of metamentation in terms that reveal its evolutionary complicity with interpretation. Metamentation operates through a battery of routines or sequences of tasks. The routines are decomposed into categories and schemes of representation that are objects of interpretation at different stages in the evolution of primate

minds. To simplify, but not by too much, this is to say that the abilities to represent metathoughts, as units of metamentation, evolved out of the abilities to represent triangular mind-world-mind relations, as units of interpretation. The stages of this evolution are surveyed in the subsequent chapters.

Chapter 5 is about situated interpretation and its earliest contributions to the edifice of metamentation. Situated interpretation is perceptually immersed in the here and now and has an interactive version in apes and an intersubjective one in human children. At this stage, the interpretational contributors to metamentation are the grasp of intentionality (or a good portion of it), apparently a primate-wide ability, and sentimental minding by sharing and communicating about emotions and experiences, a unique human specialty. Chapter 6 turns to unsituated interpretation and its contributions to metamentation: the category of propositional attitudes emulated by that of explicit metathought (the atom of metamentation) and the turn to self-interpretation, which discloses one's own attitudes as mind-world relations, on a par with those of others. Chapter 7 examines the ability to hold many minds in mind, by iterating attitude attributions and embedding some in others, and also the abilities to format attitudes in common terms and integrate across domains the information represented in the contents of attitudes. The result is a unified mind that can traffic in explicit representations about whatever interests it—an accomplishment that is new and surprising from an evolutionary standpoint.

Chapter 8 wraps things up. It construes autism as providing overall empirical confirmation for the main thesis: autistic people fail at metamentation because, and possibly to the extent that, they fail at interpretation; even those who master most of the skills of language, formal reasoning, and public representations fail to extend this mastery to the mental representations of others and themselves, and thus fail to become reflexive thinkers. After a look-back review of the argument for that conjecture, the essay concludes with a forward look at a few outstanding questions.

Since these chapters tell a constructive and gradually built story and give relatively little space to critical exegesis of and comparisons with other views, I thought it would help to indicate from the outset how the reader could relate this story to other major positions on the relation of interpretation to metamentation.

## **6 Polemical Side**

Besides its constructive role, the tripartite basis of my story—evolutionary, psychological, and conceptual—also has polemical import. I take my readership to fall into three groups: opponents, fellow travelers, and undecided. I do not expect any group to accept the thesis of this essay as is or be persuaded

by a single line of argument. Since the undecided are likely to decide relative to how opponents are argued out of their positions and how fellow travelers are persuaded to see it my way, it's going to be between the latter two groups.

The opponents must be shown that in primate evolution and particularly child development, alternative routes do not add up to reflexivity either empirically or conceptually. Several such routes can be envisioned. Those steeped in the rationalist tradition may assume that mental reflexivity is an innate gift, perhaps built into the brain architecture and maturing on its own. Even though some basic skills of interpretation seem innate in primates, the late development of metamentation in human childhood, mostly under the impact of language and culture, speaks against this innatist assumption. Followers of Jean Piaget may argue that metamentation develops out of formal abilities for logical and mathematical reasoning, these in turn developing out of sensorimotor schemes for physical action. Here the conceptual line is as effective as the empirical: there is nothing in those formal abilities or the more basic action schemes to serve as models for metamentation.

Language may also look sufficient to afford metamentation: thoughts are encoded linguistically and thus are frozen and stable enough to be subject to mental scrutiny, often by means of further thoughts linguistically encoded. This gambit, necessary to making thoughts explicit (in the ways required by metamentation) and linked to other thoughts, is far from sufficient. As this essay will endeavor to show, thoughts link up reflexively with other thoughts in ways and for reasons that are independent of language and are not exhausted by its rules and constraints, whether semantic or syntactic; topical predication is one such prominent example. Autistic people may handle well large fragments of language yet fail to predicate topically and to metamentate. Also telling is the fact that young children master language years before they metamentate. Finally, it may be thought that metamentation draws solely on abilities to plan or solve problems, but again higher primates and young children may be capable of such exploits without metamentating. Even mental rehearsal with linguistic expressions is not going to be enough; autistic people with reasonable language abilities might be able to mentally rehearse, but again they fail to metamentate normally. The missing link in all these theoretical schemes is intersubjective interpretation.

Ironically, it is the fellow travelers (perhaps the largest group) that may pose a greater challenge. For many of them may think (in a 'What's the big deal?' manner) that the evolution of metamentation out of interpretation is no surprise and no mystery, since, after all, interpreters naively theorize about perceptions, desires or, indeed, thoughts. That is what makes them interpreters. On some accounts, interpreters think about their mental states even before

thinking of those of others. Interpreters, then, would be reflexive thinkers by definition. Yet, there are good reasons to think that interpreters as such are not reflexive thinkers and certainly not by definition. Metamentation is the *joint* product of several developments (interpretation, topical predication, mental rehearsal), so metamentation can't be just interpretation or be derived solely from interpretation. Although interpretation provides the key tasks emulated by metamentation, the emulation is possible only because of these other contributions. It takes a probing look at evolution and development, not a definition or even a theory of interpretation, to prove this point and to show that the journey from interpretation to metamentation is no foregone conclusion. Apes may be credited with some interpretation, but they do not metamentate. Metamentation emerges late in human childhood, even though children have been interpreters, topical predicators, and mental rehearsers for several years.

Even the conceptual story is not as simple and straightforward as it may seem. The fact that interpretation is in general about mind-world relations and, in its intersubjective version, about mind-world-mind relations does not entail that thinkers think about their thoughts in the same ways or that they inherit metamentation from interpretation. The conceptual entailment is surely not visible to those philosophers and psychologists who envision a reflexive access to one's thoughts that is based on internal experiences (introspectionists) or practical-reasoning abilities (simulationists) and does not emulate the interpretation of others. Also important is the historical fact, again revealed only by evolution and development, that initially interpretation was not about *mind*-world, let alone mind-world-mind, relations. Apes and young human children represent only observable *subject*-world relations—such as gazing, seeing or being angry at something—whose mental component is meager and implicit. The mental component grows and becomes more explicit in later childhood when propositional attitudes are mastered, but even the categories of propositional attitudes are far from representing mental states in general, far from representing them reflexively, and far from originating in self-ascription. The turn to self-interpretation is a late development in childhood, and its explanation does not follow from just having the ability to interpret propositional attitudes, as many fellow travellers (and most philosophers) believe. When interpretation turns to self, it opens the way to, and provides a model for, metamentation. Yet even that process is not as obvious, simple, and predetermined as it may seem. There are still many variables needed to bring it to fruition. All in all, then, the fellow traveler may have at least as many reasons as the opponent or the undecided to read on. All are welcome.