MASSACHUSETTS INSTITUTE OF TECHNOLOGY

ARTIFICIAL INTELLIGENCE LABORATORY

A.I. memo No. 603 November, 1980

Jokes and the Logic of the Cognitive Unconscious

Marvin Minsky

Abstract Freud's theory of jokes explains how they overcome the mental "censors" that make it hard for us to think "forbidden" thoughts. But his theory did not work so well for humorous nonsense as for other comical subjects. In this essay I argue that the different forms of humor can be seen as much more similar, once we recognize the importance of *knowledge about knowledge* and, particularly, aspects of thinking concerned with recognizing and suppressing *bugs* -- ineffective or destructive thought processes. When seen in this light, much humor that at first seems pointless, or mysterious, becomes more understandable.

Acknowledgements: This report describes research done at the Artificial Intelligence Laboratory of the Massachusetts Institute of Technology. Support for the laboratory's artificial intelligence research is provided in part by the Advanced Research Projects Agency of the Department of Defense under Office of Naval Research contract N00014-80-C-0505 and in part by the Office of Naval Research under Office of Naval Research contract N00014-79-C-0260.

INTRODUCTION

> *A gentleman entered a pastry-cook's shop and ordered a cake; but he soon brought it back and asked for a glass of liqueur instead. He drank it and began to leave without having paid. The proprietor detained him. "You've not paid for the liqueur." "But I gave you the cake in exchange for it." "You didn't pay for that either." "But I hadn't eaten it".*
>
> --- from Freud (1905). [0]

In trying to classify humorous phenomena, Sigmund Freud asks whether this should be called a joke, *"for the fact is we do not yet know in what the characteristic of being a joke resides."* Let us agree that some of the cake-joke's humor is related to a logical absurdity -- leaving aside whether it is in the logic itself, or in keeping track of it. Later Freud goes on to ask what is the status of a "knife without a blade which has no handle?" This absurdity has a different quality; some representation is being misused -- like a frame without a picture.

Freud, who never returned to the subject after writing his 1905 book on the theory of jokes {0}, suggested that "censors" in the mind form powerful, unconscious barriers that make it difficult to think "forbidden" thoughts. But jokes can elude these censors -- to create the pleasure of unearned release of psychic energy, which is discharged in the form of laughter. He explains why jokes tend to be compact and condensed, with double meanings: this is to fool the childishly simple-minded censors, who see only innocent surface meanings and fail to penetrate the disguise of the forbidden wishes.

But Freud's theories do not work as well for humorous nonsense as for humorous aggression and sexuality. {0} In this essay I try to show how these different forms of humor can be seen as much more similar, once we make certain observations about the nature of commonsense reasoning. Here is our thesis:

1. *Common sense logic is too unreliable for practical use.* It cannot be repaired, so we must learn to avoid its most common malfunctions. Humor plays a special role in learning and communicating about such matters.

2. *It is not enough to detect errors in reasoning; one must anticipate and prevent them.* We embody much of our knowledge about how to do this in the form of "censors" that suppress unproductive mental states. This is why humor is so concerned with the prohibited.

3. *Productive thinking depends on knowing how to use Analogy and Metaphor.* But analogies are often false, and metaphors misleading. So the "cognitive unconscious" must suppress inappropriate comparisons. This is why humor is so concerned with the nonsensical.

4. *The consequences of intellectual failure are manifest in one's own head, while social failures involve other people.* Intellect and Affect seem less different once we theorize that the *"cognitive unconscious"* considers faulty reasoning to be just as "naughty" as the usual "Freudian" wishes.

5. *Humor evolved in a social context.* Its forms include graciously disarming ways to instruct others about inappropriate behavior and faulty reasoning. This deviousness makes the subject more confusing.

Our theory emphasizes the importance of *knowledge about knowledge* and, particularly, aspects of thinking concerned with recognizing and suppressing *bugs* -- ineffective or destructive thought processes. When seen in this light, much humor that at first seems pointless, or mysterious, becomes more understandable. {1}

I. PROBLEMS OF COMMON SENSE REASONING

When you tell a young child *"I am telling a lie"* then, if he is old enough to reason so, he will think: *"If that is false, then he's not telling a lie. But, then it must be true. But then, it must be a lie, for it says so. But then ---"*. And so on, back and forth.

A child might find this situation disagreeable for several reasons. It challenges the belief that propositions are always either true or false. It threatens to propagate through his knowledge-structure, creating other inconsistencies. And he can make no progress when his mind returns again and again to the same state. {2} Common sense can go awry in endless ways. Beliefs can be wrong from the start, one can make mistakes from each step to the next, and one can wander aimlessly, getting nowhere. But before we discuss these, we should observe what actually happens when you say things like *"this statement is false."* Often the listener first seems puzzled, then troubled, and finally laughs. *"That's funny,"* he says. *"Tell me another liar joke".*

THE PROBLEM OF TRUTH: *How do we know where to begin?* The conclusions of even the best reasoning can be no better than its premises. In mathematics, this is of little concern (because one cares more where premises lead than where they come from). But in real life, few propositions are perfectly trustworthy. What does one do when an accepted "fact" turns out false? One of my children was once entranced by an ornamental fish-shaped bowl with four short legs. After a while she announced, somewhat uncertainly: *"Some fish don't have legs."*

We never know anything for certain. What should one do upon reaching a conclusion that appears false -- erase all the assumptions? When a long held belief turns false, should one erase all that has been deduced from it? When an acquaintance once tells a lie, should one reject everything else he ever said? There is no simple, single rule: each person must find his own ways to maintain his knowledge about his knowledge. [1]

WHENCE THE RULES OF INFERENCE? *How do we know how to infer?* Most people believe that *if most A's are B's, and if most B's are C's, then most A's are C's.* Though false, this has great heuristic value, especially for children who encounter few exceptions. *Psychologically,* I see no great difference between heuristic and logical reasoning; deduction is "just another" kind of evidence. We firmly believe a deduced conclusion only when it seems plausible on other grounds as well. At one time, many philosophers held that faultless "laws of thought" were somehow inherent, a priori, in the very nature of mind. This belief was twice shaken in the past century; first when Russell and his successors showed how the logic men employ can be defective, and later when Freud and Piaget started to reveal the tortuous ways in which our minds actually develop.

After Russell observed that the seemingly frivolous *"Who shaves the Barber who shaves everyone who does not shave himself?"* was a truly serious obstacle to formalizing common sense logic, he and others tried to develop new formalisms that avoided such problems -- by preventing the fatal self-references. But the proposed substitutes were much too complicated to serve for everyday use.

I am inclined to doubt that anything very resembling formal logic could be a good model for human reasoning. (The paper by Hewitt and Kornfeld might suggest a possible avenue of compromise [2].) In particular, I doubt that any logic that prohibits self-reference can be adequate for psychology: no mind can have enough power -- without the power to think about Thinking itself. Without Self-Reference it would seem immeasurably harder to achieve Self-Consciousness -- which, so far as I can see, requires at least some capacity to reflect on what it does. {3}

If Russell shattered our hopes for making a completely reliable version of commonsense reasoning, still we can try to find the islands of "local consistency," in which naive reasoning remains correct. It seems that only certain kinds of expressions lead to paradoxes and inconsistencies, and it seems worth taking some risks, gambling for greater power -- provided we can learn, over time, to avoid the most common disasters. We all know the legend of the great mathematician who, warned that his proof would lead to a paradox if he took one more step. He replied *"Ah, but I shall not take that step."* {4} One would miss the point to treat this as a "mere" joke. What it means, really, is that we build into our minds two complementary functions:

We work to discover "islands of power" within which commonsense reasoning seems safe.

We work also to find and mark the unsafe boundaries of those islands.

In civilized communities, guardians display warnings to tell drivers about sharp turns, skaters about thin ice. Similarly, our philosophers and mathematicians display paradigms -- like the Barber, the Tortoise, and the Liar -- to tell us where to stop -- and laugh. I suggest that when such paradigms are incorporated into the mind, they form intellectual counterparts to Freud's emotional censors. This would help explain why purely logical nonsense so often has the same humorous quality as do jokes about injury and discomfort -- the problem that bothered Freud. The cake-joke reminds us, somewhat obscurely, to avoid a certain kind of logical absurdity -- lest we do ourselves some vaguely understood cognitive harm. Hence our thesis: since we have no systematic way to avoid all the inconsistencies of commonsense logic, each person must find his own way by building a private

collection of "cognitive censors" to suppress the kinds of mistakes he has discovered in the past.

HEURISTIC CONTROL OF LOGIC: *How do we know what next to do?* I once tutored a student having trouble with middle-school geometry. I began to explain the axioms, and how proofs were structured. *"I understand all that,"* he said, *"only I was sick the day the teacher explained how to find the proofs."*

It is not enough just to know the principles of reasoning; one must know also when to apply them. We each know millions of facts, and perhaps millions of rules of inference. Which to apply to what, and when? There is a basic problem of direction, of not meandering, lest one aimlessly derive billions of inferences, all perfectly logical but none relevant to any goal. First, some "plan" is required. Next, one must avoid circling -- returning to the same place again and again. Finally, to avoid confusion, one needs an administrative structure to keep track of what one is doing, and why.

The new science called Artificial Intelligence is concerned with just such issues of the efficiency and effectiveness of Reason -- matters rarely discussed in Logic or Philosophy, which focus on verifying that proofs are valid, or that arguments are sound, rather than on how proofs are discovered. Much has been learned, in "AI", about how to avoid excessive meandering and confusion, by using goal-structures and plans -- techniques for insuring progress. Using such methods, some modern computer programs can thread their ways through some quite complicated situations.

Nevertheless, the problem of meandering is certain to re-emerge once we learn how to make machines that examine themselves to formulate their own new problems. Questioning one's own "top-level" goals always reveals the paradox-oscillation of ultimate purpose. How could one decide that a goal is worthwhile -- unless one already knew what it is that is worthwhile? How could one decide when a question is properly answered -- unless one knows how to answer that question itself? Parents dread such problems and enjoin kids to *not take them seriously.* We learn to suppress those lines of thoughts, to "not even think about them" and to dismiss the most important of all as nonsensical, *viz.* the joke *"Life is like a bridge." "In what way?" "How should I know?"* Such questions lie beyond the shores of sense and in the end it is Evolution, not Reason, that decides who remains to ask them.

II. CENSORSHIP

Just like taboos in human societies, certain things must not be thought inside the Mind. The best way for a child to learn not to do a certain bad thing would be to learn *not to even to think of it.* But isn't that like trying "not to think of a monkey"? Contrast two ways: (i) suppress an idea already in the mind or (ii) prevent it from being thought in the first place:

(i) Stop thinking that!
(ii) Don't even (or ever) think that!

It is easy to begin to make a type (i) censor: wait for the specified "bad" event to happen, and then suppress it. But how does one prevent it from recurring? It seems harder to begin to make a type (ii) censor, because it must be able to recognize the repressed thought's Precursors --- but it is easier to see what to do next, since each Precursor usually leads to several options, and suppressing one still leaves the others. So we shall discuss only the second censor-type.

So, our censor-builder has to learn to recognize Precursors -- mind-brain states that precede a recognizable (and, here, to be prohibited) activity. To do this, it will need a short term memory (to remember what *just* happened) and a long term memory (to store the result of learning). The latter may eventually become quite large, because a prohibited event may have many different precursors. In any case, a experienced type (ii) censor can recognize its joke by the situation and need not wait for the punch line. A type (i) censor makes you wait till the comedian finishes: only then can you complain *"Oh, I've heard that one before!"*

To place these in a larger framework, Let's consider a simple "two-brain" theory. An "A-brain" has sensory inputs from the world, and motor outputs to the world. The B-brain's inputs come from the interior of A -- so B can perceive "A-states" --- and its outputs go into A, so B can affect activities in A. Thus, B can "see" what is happening inside A, and act to influence it, just as A can "see" and affect what happens in the world. B need not -- and probably can not -- know what A-events "mean," vis-a-vis the world, but B is in a position to recognize such metapsychological conditions such as A being "meandering, circling, or confused." {5}

When a B-censor acts, it must disturb the A-brain so as to suppress the undesired activity. (It would be even better for B to remember, from past events, which is a *good* way to go, but that is outside this essay's concern.) In any case, the point is that precursor-sensitive censors can do their work *before* the problems that they evade actually arise. Probably, also, they can do this so quickly and gently as to produce no noticeable mental phenomenology. This would explain why censors are (in the) unconscious.

The censorship theory explains why a joke is not so funny if you've heard it before; this is because a new censor has been constructed, or an old one extended. Freud touches on "novelty" as a component of humor, but never dwells on why old jokes get few laughs. I presume that he simply considered it too obvious to mention, that censors are learners.

How big must the censor memory be, to protect us from naive reasoning mistakes? Probably not so large for formal logic, considering how rarely we discover a new paradox. But for avoiding nonsense in general, we might accumulate millions of censors. For all we know, this "negative meta-knowledge" -- about patterns of thought and inference that have been found defective or harmful -- may be a large portion of all we know.

Consider the activities called *play* and *practice.* The insensitive learning theories of behavioristic psychology regard *play* not at all, and see practice as reinforcing something repetitive. But *practice* (I conjecture) is often far from mere repetition and refinement of the same thing; often it is exploratory, testing out a skill's minute variations and perturbations -- and learning which of them

to enhance or suppress. Similarly *play*, (commonly seen as "mere") is often also an exploration of variations on a larger scale. Many other everyday activities can so be seen as ways to learn to avoid bugs and mistakes.

I know a young child whose sense of humor tends toward jokes like *"what if the spoon were rubber,"* apparently observing that a flexible spoon would be absurd because the food would fall out of it. Is he enforcing a "must-be-rigid" property of some spoon-frame, or is he censoring the use of some spoon-frame that lacks the property? The humor-behavior of children also needs more study.

III. MEANING AND METAPHOR

> *Two villagers decided to go bird-hunting. They packed their guns and set out, with their dog, into the fields. Near evening, with no success at all, one said to the other, "We must be doing something wrong". "Yes", agreed his friend. "Perhaps we're not throwing the dog high enough."*

When you want to fasten a screw and therefore reach for a certain screwdriver, your mind has chosen to see that screwdriver as a screw-driver; you could have seen it as a kind of dull knife, or as a hammer without a head. When we see something only in its "intended" aspect, we are "confusing the thing with itself." As Korzybski [3] intoned, *"whatever a thing is, it is not".* **{6}, {7}**

FRAMES: I suggested in [4] that perceptions are ordinarily interpreted by the mind in terms of previously acquired description-structures called *Frames.* A frame is a way to represent a stereotyped situation, like being in a certain kind of room, or going to a certain kind of party. Attached to each frame are several kinds of information; some about how to use the frame, some about what one might expect to happen next, some about what to do if those expectations are not confirmed, and so forth. This theory was proposed to explain the speed and virtual absence of noticeable phenomenology in perceiving and thinking, and here I propose to sketch just enough of it to explain some features, and some "bugs," of reasoning. Then we can return to unconscious censorship and error-correction.

Each frame includes, among other things, a variety of *terminals* to which other frames are attached. Thus a chair-frame specifies that a (certain kind of) chair has a *seat*, a *back*, and four *legs*. The details of these would be described, not in the chair-frame itself, but in other frames attached to its terminals. Each frame includes also a set of features which, if enough of them are present, may activate the Frame. So, when you see enough parts of a chair, these will activate one of your chair-frames which, in turn, will activate the sub-frames attached to its terminals. These, then, will "look for" other chair-parts that were not recognized at first -- because of being unusual, or partially hidden from view, or whatever. Finally, if some elements required by the frame are not seen at all -- one rarely sees all the legs of a chair, and *never* all the sides of a box -- the missing elements are supplied by *default.* This is easy because most terminals of most frames have certain sub-frames already attached as *default assignments.* When one reads in a story about some shoe or some chair,

these cause one, "by default" to assume a certain kind of shoe or chair.

The concept of default goes much further. When one sees a person in a sitting posture then, even if every part of his chair is hidden from view, one will pseudo-see a chair under him. Unless your attention is drawn to the fact, you never notice that no chair was actually seen. This is because the "sitting" frame includes a "must-be-supported-by" sub-frame terminal, and this is attached to some "typical" chair-frame, for its *default assignment.* {8}

The chair-frame that is selected can depend, however, on the context for default assignments are weak and easy to "displace". If there are other chairs around, the invisible chair will be assumed to be like one of them. If the scene is set in a park, then a park-bench frame might be activated to serve as default. If one then noticed an arm-chair arm, the system would replace the weakly-attached bench-frame by one that better suits what was seen -- and one now "sees" an armchair.

According to the theory in [4] this is done very swiftly because the "corresponding" terminals of related frames are already pre-connected to one another. This makes it easy to change a faltering interpretation or a frustrated expectation. Shifting from one related frame to another should be so fast and efficient as to be imperceptible to introspection. This is why one so easily recognizes any chair as "a chair", even though particular chairs are so different from one another. I do not suggest that all this happens by magic; *the interconnecting network is constructed over a lifetime of experience.* In [5] I discuss how new frames arise usually as revised versions of older ones, bringing those "common terminals" along with them. In [6] are more details, but one must read between the lines of that paper because it does not use the terminology of frames.

Frames and frame-systems are used at conceptual as well as perceptual levels, and there we find other kinds of *frame-systems* -- families of interconnected frames -- that are not transformed so easily, hence more effort is noticed. Consider, for example, Wittgenstein's paradigmatic question about defining "game." [7] The problem is that *there is no property common to all games*, so that the most usual kinds of definition fail. Not every game has a ball, nor two competing teams; even, sometimes, there is no notion of "winning." In my view, the explanation is that a word like "game" points to a somewhat diffuse "system" of prototype frames, among which some frame-shifts are easy, but others involve more strain. The analogy between football and chess is strained only a little, more with solitaire, and so on. Shifting from one familiar kind of kitchen-chair to another is imperceptible, but changing a park-bench to an arm-chair would be strain enough to "surprise".

Now I propose that much of commonsense logic itself is based on learning to make shifts between frames that have terminals in common. For example, if a situation fits a frame like *A implies B, and B implies C*, a simple frame-shift re-represents it as *A implies C.* Seen in this light, Freud's cake story appears to display some sort of incorrect-logic script in which each consecutive pair of sentences matches some such tried-and-true kind of reasoning step.

I presume that when we "understand" this sort of story, we represent it in our minds as a series of pairs of overlapping assignments of things to terminals of such frames. And somewhere along the

way, in the cake story, there is an improper assignment-change. Is it the payment moving from the cake to the drink? Is it a pivot between "owns" and "possesses?" Each listener must make his own theory of what is wrong -- and devise his own way to avoid this confusion in the future. Some people will do better at this than others.

METAPHOR: All intelligent persons also possess some larger-scale frame-systems whose members seemed at first impossibly different -- like water with electricity, or poetry with music. Yet many such analogies -- along with the knowledge of how to apply them -- are among our most powerful tools of thought. They explain our ability sometimes to to see one thing -- or idea -- as though it were another, and thus to apply knowledge and experience gathered in one domain to solve problems in another. It is thus that we transfer knowledge via the paradigms of Science. We learn to see gases and fluids as particles, particles as waves, and waves as envelopes of growing spheres.

How are these powerful connections discovered? For the simple ones, there is no great problem: some frames are easily recognized as similar because their terminals accept the same sorts of entities; these could be located and classified by simple algorithms, e.g., searches for best match. As for those most subtle, once-in-a-lifetime insights, whose analogical powers are hidden deep in the procedural structures that operate on them, we hardly need a general theory to account for them since -- like favorable evolutionary mutations -- few are ever discovered by any single individual, and those can be thereafter transmitted through the culture. In any case, putting aside the origins of those rare, greatest insights, each individual must have his own ways to build new connections amongst his frames. I will contrast particular methods against general methods, arguing that the two have problems that seem somewhat opposite in character -- that the errors of "particular" methods can be managed additively, while bugs in the "general" methods must be repaired subtractively. This last point will eventually bring us back to censors.

PARTICULAR ANALOGIES: In the course of thinking, we use different frames from one moment to the next. But frequently one of the active frames will fail -- "that's not a door, only a big window." Winston, in [8] suggests that whenever such an error is (somehow) detected, described, and corrected, we can attach to the failing frame a "pointer" to some other frame that has been found to work in this circumstance. The pointer must contain, of course, a description of the failure circumstance. A family of frames connected in such a way is called a *difference network.* {9} We can explain [4] the definition difficulty (e.g., that of defining "game") by supposing that such words point not to any single frame, but into such a network.

Any such link between two frames implies a generalization of the form *"A's are like B's, except for D".* Thus, such a link is a fragment of an analogy. Of course, like any generalization, it will likely soon fail again and need refinement. Winston's thesis [8] suggests ways to do this. The important point here is that, particular analogies discovered in the course of experience, can be remembered by adding *positive,* active links between frames.

GENERAL ANALOGY METHODS: What if one is confronted with a novel situation that does *not* arouse any particular frame? Then, it makes sense to try some "general" method, e.g., to compare the situation to some large class of frames, and, select the one that "best fits." Such a

method can do much better than chance, but what if it yields a result that does more harm than good? I will argue that one now must build a censor, and that there is a general principle here that learning theorists have not appreciated: *positive general principles need always to be supplemented by negative, anecdotal censors.*

For, it hardly ever pays to alter a general mechanism to correct a particular bug. Almost every instance-specific modification of a best-match mechanism would reduce its general usefulness. So when the general mechanism yields a bad result, one can only remember to suppress its subsequent appearances -- that is, to build a censor. If a child wants his sibling's toy, he will first try seizing it -- the most general way to get things one wants. Once the parents see to it that this is censored, then he can find other ways. But he must *not* learn *in general* not to take things he wants, lest he become helpless.

Some forms of humor, notably puns, turn on changing the meaning-sense of a word. (Besides the easily distinguished dictionary senses, most words have also many others that would never be separated in a dictionary. "Lift," for example, has different implications when an object weighs one gram, or a thousand, {7} and really to understand lifting requires a network for making appropriate shifts among such different micro-senses.) While verbal sense-shifting can be funny, and even useful, it is dangerous, and especially hazardous to be subject to the fortuitous, meaningless sense-shifts that depend on superficial word-sound similarities. In a fragment of a schizophrenic's transcript, the patient sees a penny in the street, says "copper, that's a conductor," then must run to a street car to speak to the conductor. Perhaps this disorder is one of frame-shift control, either disabling the bad-analogy suppressors or irresponsibly enhancing the general analogy-finder.

The element that seems to me most common to all the different kinds of humor is that of unexpected frame-substitution, in which a scene is first described from one viewpoint and then suddenly -- typically by a single word -- one is made to view all the scene-elements in another, quite different way. Some such shifts are insightful, of course, while others are mere meaningless accidents. Next we turn to a kind that could be turning points in each individual's personal evolution.

IV. TRAUMATIC COGNITIVE EXPERIENCES

"Yields truth when appended to its own quotation"
yields truth when appended to its own quotation.
--W. V. Quine

In the popular psychology of our day, Intellect is seen as straightforward, deliberate, conscious, and emotionally neutral; but in regard to emotional matters, the public has come generally to accept the psychoanalytic view that Affect is dominated by unknown terrors and traumas, lurking in the unconscious since childhood. Now I want to challenge this affect-intellect "dumbbell". I certainly do not mean to suggest that there are no differences of kinds in mental affairs -- only that this

particularly popular distinction, however useful in everyday life, does more harm than good in psychology. [6]

In any case, the Affect-Intellect distinction would lose much of its force if Reason, too, employed a powerful "cognitive unconscious" -- and that is exactly what I shall argue: that Intellect, too, has its own buried secrets. In Freud's scenario, the ego develops largely in contexts of fear of deprivation, punishment or mutilation; of anxiety about uncertainty and insecurity; terror of losing the esteem or person of parent or attachment-figure. New families of censors -- new domains of repression -- are created when wishes conflict enough with reality to justify attempting to keep them from becoming conscious.

Does anything like that happen in the intellectual sphere? One might suppose not, because a necessary element is missing -- that of a beloved -- or punitive -- authority figure. However, while Freudian taboos originate from outside, the child needs no external authority-figure to point out his gross cognitive failures; he needs no parent to scold him, when an encounter with paradox throws his mind into a frightening cyclone. The momentary loss of mental control should provoke anxiety in its own right.

But, one might ask, if we bear the scars of frightening cognitive experiences, why do they not reveal themselves (like the affect-laden ones) in nightmares, compulsions, phobias, and the like? Perhaps they do; it would not show in the interpretations of present-day psychiatry. But every teacher knows (and fears) the rigid inhibition of children's cognitive phobias: *"I don't want to learn this; I couldn't possibly do that"*. Let us speculate on how such fears might originate. Consider the paradox of *Nearness*: Every child must once have said to himself:

> *Hmmm. Ten is almost Eleven. And Eleven is nearly Twelve. And so on; Ninety-Nine is nearly a Hundred. But then Ten must be almost a Hundred!*

To an adult, this is not even a stupid joke. But we each must once have thought something like: "there is obviously something wrong here. Is it in my premises or in my logic? Well, what's my premise? Obviously, that *"if A is near B, and if B is near C, then A must be near C."* Nothing wrong with that. So there must be something wrong with my logic. But I'm only using something like: *"If A implies B, and if B implies C, then A implies C."* "How could *that* be wrong? No way!"

To be sure, not everyone remembers such experiences as frightening. In fact, *ad hominem,* readers of essays like this one would more likely complain that they *like* such problems and cannot see them as "traumatic." No matter, such readers are just the ones who have found ways to transform -- what did Freud call it? -- to *sublimate* such problems into constructive thinking about thinking. In any case, in one private manner or another, everyone somehow comes to deal with such problems, and I see only one practical way: we must each grow for ourselves some structure, of large complexity and little elegance, to tell us when --and when not -- to trust each such pattern of inference. For example, we each learn never to repeat a *near* deduction more than a few times in any one argument. Furthermore, the accumulation of such experiences leads us eventually to realize that this is not peculiar to "near" alone: perhaps one shouldn't use *any* inference method too many

times. What is "too many?" There is, I fear, no elegant answer. We each have to learn and master large bodies of knowledge about the limitations of each species of reasoning.

It might be useful to try to catalog the kinds of cognitive incidents that must have so baffled each of us in early life. Each reader must recall the distress of being made to discern the arbitrary boundaries between one ocean and another, or at trying to answer *"which came first, the chicken or the egg?"* Every child must have wondered about his origin and from whence came the *first* person. Only the dullest child never found for himself some sort of Zeno paradox, or Achilles and Tortoise problem. I remember being especially disturbed to discover that there were questions that adults had *no* way to answer.

MYSTICAL EXPERIENCE: Another family of disturbances arise when we question our own purposes. *"Why should I do this,"* -- whatever it is -- one asks, and proposes some answer. But then one is impelled to continue, "why should I want *that,"* and so forth. There is a not uncommon phenomenon -- sometimes called mystical experience -- from which a person emerges with the conviction that some unsolvable problem (like the purpose of existence) has been completely explained; one can't remember quite how, only that it was answered so well as to leave no doubt at all. This, I venture, reflects some mental mechanism (perhaps one of last resort) that, in a state of particularly severe turmoil or stress, can short-circuit the entire intellectual process -- *by creating the illusion that the problem has been settled.* Powerful but dangerous, such a mechanism short-cuts the canons of normal confirmation. One kind of confusion-cycle is thereby broken, but this may damage other ways in which sane minds confront beliefs with evidence. Then, anything can happen. {10}

V. HUMOR AND EVOLUTION

If you wish to study a granfalloon
just remove the skin of a toy balloon.
-- Kurt Vonnegut, in *Cat's Cradle*

In the 1912 edition Freud, still perplexed about the purpose of nonsense, recounts a joke of this form: {11}

> *"A man at the dinner table dipped his hands in the mayonnaise and then ran them through his hair. When his neighbor looked astonished, the man apologized: "I'm so sorry. I thought it was spinach."*

We have argued that learning about bugs is central to the growth of reason. But reason itself grows in no vacuum; most of our ideas -- *and our ideas about ideas* -- come via our families and cultures, and this poses some special communication problems. For one, it is risky to point out the mistakes of a person one wants to please. So this must be done in some "conciliatory" manner -- and humor seems involved in this. For another, if learning about bugs involves a special kind of memory, then

this communication must somehow engage that memory. In this section we propose that humor -- and more specifically, *laughter* -- is innately enabled to do this, too.

But first we digress to discuss an important methodological problem: *Why is it so hard to explain why jokes are funny?* Why, for that matter, is it so hard to say precisely what is a joke? We have already mentioned Wittgenstein's problem of defining "game": one can find no single quality common to all the different kinds of examples -- and one finds a similar problem in attempting to define "humor." *But we did not stop to ask why this is so.* One might suppose it a mere surface difficulty, and hope that we may yet find a single underlying structure from which all funny things spring -- some basic "grammar of humor," or "comical deep structure." Not so, I fear; when we look deeper for that underlying structure of humor we shall still find a vexing lack of unity. I argue that this is a consequence of the way things usually evolve in biology.

In mechanisms designed by plan, it is reasonable to ask about purpose or cause. What is the purpose of that beam in this house? Simple: to hold up this roof -- and, perhaps, to hold those walls apart. But, when we ask questions about structures created by evolution, we find that only rarely does one evolutionary increment serve a single purpose -- and rarely is one alone in serving any particular purpose. Behavior emerges from a network of interdependent mechanisms, and one cannot expect any compactly circumscribed theory (or mechanism) completely to "explain" any single surface component of behavior. What a theory *can* do, though, is to describe some fragment of that larger network of interacting subsystems.

Humor, like games, serves and exploits many different needs and mechanisms. It lacks sharp, natural boundaries because those underlying things themselves overlap and exploit one another. When we employ a *word* like "humor," one has the illusion of designating something sharper than this kind of complex web of relations among laughter, faulty reasoning, taboos and prohibitions, and unconscious suppressor mechanisms. But, I think the very clarity of words is itself a related illusion; as noted in {7}, language itself works only because oversimplification is more useful than realistic confusion -- that is, in real life, if not for thinking about psychology.

ROLES OF LAUGHTER: Consider what happens when a thought-situation comes to be perceived as funny or absurd: further reasoning is drowned in a flood of activity -- furious motions of thorax, abdomen, head, limbs and face, accompanied by loud barking, wheezing, and choking noises. To a Martian, an epileptic seizure would be less alarming. Adults, of course, can train themselves to suppress this, but that is another matter.

> *Laughter disrupts reasoning:* The laughter reaction is so distracting as to keep the mind from proceeding further along the prohibited or ridiculous path it has started. Whatever that line of thought, the disruption prevents you from "taking it seriously," from acting upon it or considering its further logical consequences.

At the same time, laughter exercises another, complementary function.

Laughter focusses attention: While disrupting further reasoning, laughter takes a firm grip on the thought itself, holding up the absurdity in sharp focus. Perhaps the joke-thought is given full attention, holding the incongruity in some "short term memory" -- so that "censor-learning" agency can absorb it.

Thus "humor" might serve to mediate the process in which the censors learn, much as "pleasure" is often supposed to mediate ordinary learning. {12}

EVOLUTION OF HUMOR: How might all this have evolved? We conjecture that it happened while our evolving minds passed through stages of increasing ability to reflect -- to think not merely about the physical problems at hand, but about how we should apply our mental resources to them; in a word, when we were learning to *plan. In order to make realistic plans, we had to learn to take account of what we could make our minds do.*

This ability could not have emerged all at once. There must have been intermediate steps -- such as the appearance of multi-level schemes like the one suggested above, in which an A-mind is monitored by a B-mind. Eventually we became able to symbolize and manipulate *representations* of plans, and this allowed the first direct references to our own mental activities. Now, suddenly, we could do such tricks as to relate statements to their own quotations, and make propositions that (for better or for worse) could discuss their own truth -- as in Quine's *"yields truth"* tour de force. {13}

In any case, once able to accomplish intricate chains of reasoning, we became vulnerable to new kinds of bugs: faulty variable bindings, subtle changes of sense, and more obscurely circular logic. This same epoch probably saw also the emergence of Language for social communication, which also converged toward using more concise and manipulable symbolic representations. And, because we could not weaken the expressiveness of symbol-manipulation without losing its power, we had to evolve those censor memories. Of course, what actually happened was surely not this simple, but we had better return again to our speculations about laughter.

Laughter's facial component suggests that it evolved in connection with social communication. It appears to be derived (ethologically) in part from a "conciliatory" expression, but it includes also a baring of teeth that suggests a defensive-aggressive mixture. {14}

Laughter's bizarre vocal component also suggests social functions that combine ancestral "releasers" for both conciliation and aggression. Perhaps it came somehow to serve as a signal to induce another person to stop whatever he was doing: whether because dangerous, pointless, objectionable, ridiculous, or otherwise forbidden.

Later, this function became internalized. If a person could feel and hear himself laugh, grimace, and shake, why not exploit these side-effects also to make one's own self to stop doing something ridiculous or prohibited? Perhaps, literally, men first learned to laugh at their own mistakes, and later learned to censure themselves in silence.

I make no plea for this particular scenario, only that something of this general sort must have happened, in which several pre-existing complexes grew together. They each brought along a variety of older interactions with other systems and purposes -- that were exploited to produce the puzzling combinations of conciliation, aggression, sexuality, and nonsense that are now mixed together in humor. If other mental structures also share this kind of tangled ethological ancestry, then a mind grown this way must now resemble a great spider-web, in which many threads of different biological purpose intersect in many nodes of different multi-purpose mechanisms. If so, the goals of psychological theories must be to describe different fragments of that web -- each to make a map of some few of those threads and nodes.

By a curious coincidence, our *theories* of how minds work must probably have themselves this same peculiar, web-like character -- albeit for a different reason. For (I think) the only way a person can understand anything very complicated is to understand it at each moment only locally -- like the spider itself, seeing but a few threads and crossings from each viewpoint. Strand by strand, we build within our minds these webs of theory, from hard-earned locally intelligible fragments. The mind-spider's theory is correct to the extent that the model in his head corresponds to the mechanism in his head. {15}

Finally it is probably futile to ask *precisely* what humor is. Korzybski's injunction applies here especially: the word "humor" points to no real thing. Instead, in each different person's mind it points to a slightly different web-model. We each use -- and mean -- the word a little differently -- just as we each laugh at different jokes. Now I do not mean to hint that the problem is unreal, or even that it is especially incomprehensible. I only want to suggest that "humor" may be less a Thing-Part *of* the mind, and more a Thing-Theory *in* the mind. This makes it no less worthy of study, but one must be clear about what one is doing. If we get confused between making theories about theories and making theories about things, we may spin forever. {16}

TIME CONSTANTS: According to our thesis, familiar types of jokes should seem less funny, because the censors have learned more about them. Why, then, do some kinds of jokes remain so persistently funny? People tire of old nonsense jokes, but not of jokes about forbidden aspects of sex. Does this falsify our theory? Not necessarily; it may mean only that the censors for this particular subject are much slower to learn and to change. Is that plausible?

Most psychological theories of our day -- both popular and professional -- seem to suppose that all memories are made of the same stuff, stored in the same huge container. (I argue otherwise in [6].) But, on reflection, we see that some memories *ought* to be less changeable than others. Contemplate, for example, the plight of a mother with a new infant, that will demand her time and attention for several years. Why doesn't she wonder *"Why am I doing this,"* or *"what could this baby do for me to justify such a sacrifice."* One might argue *"To preserve the race,"* or *"because you will love it,"* or *"because it will repay you some day,"* but these would hardly convince any rational person; raising a child is not, let's face it, a notably sensible enterprise.

Conventional wisdom recognizes that love is far from rational, and holds that an "instinctive" attachment is somehow created and somehow protected from casual alteration. Clearly so, but

we must know more about those "somehow"s. This maternal self-questioning doesn't usually go too far, perhaps because we are protected by the web of personal pleasure and social compulsion surrounding child-rearing. But the problem is real, and there are occasional (and invariably concealed) tragedies in which a young mother's frustration overwhelms her attachment.

The simplest way might be to build attachment into a special kind of memory that, once established, tends to persist for several years. After all, long time-constants characterize other aspects of attachment. Some persons always choose different partners of similar appearance; as though unable to alter a fixed stereotype. Others find themselves in the grip of undesired infatuations, that reason declares inappropriate. And most familiar is the seemingly inexorable time-span of mourning -- the year or two it takes to adjust to separation or loss. All these could be by-products of adaptations of older mechanisms whose slowness was/is of value in our sociobiological evolution. {17}

Perhaps one can also interpret in this light the prolonged, mourning-like depression associated with sexual assault, presuming that the momentary association of violence with sexuality somehow impairs the entire attachment machinery. The large time-constants make recovery slow, from a profound disturbance of normal social attachments. No matter if the victim manages to view the incident "rationally;" this does not automatically restore those sluggish mechanisms to their normal state, and one must suffer the prolonged functional deprivation of an important mind-part.

All this suggests that the curious robustness of sexual humor may reflect only that the associated censors are among the "slow learners" of the mind, like retarded children. Perhaps they indeed *are* retarded children-- the nearly static remnants of our own early selves. They change only slowly, and our tireless enjoyment of certain censured subjects may be a side-effect of that circumstance.

So, we finally conclude: jokes are not really funny at all, but reflect the most serious of concerns; the pursuit of sobriety through the suppression of the absurd.

Cambridge, Mass.
May-July, 1980

NOTES

[Note 0] Freud seemed somewhat puzzled by "nonsense jokes" and suggested, to explain the worst of them, that they "give the teller the pleasure of misleading and annoying, by rousing the expectation of a joke and then frustrating the listener" -- who in turn -- "damps down his annoyance by determining to tell them himself later on." The enjoyment of nonsense, he goes on, might also reflect a wish to return to a childhood of relaxed, careless thought in which one *"puts words together without regard to the condition that they should make sense, in order to obtain from them the pleasurable effect of rhythm or rhyme. Little by little he is forbidden this enjoyment, till all that remains to him are significant combinations of words. But when he is older attempts still emerge at disregarding the restrictions that have been learned."* [0, p.125] In connection with alcoholic cheerfulness, Freud recounts a pun: "It is most nstructive to observe how standards of joking sink as spirits rise" -- and later -- *"the grown man becomes a child, who finds pleasure in having the course of his thoughts freely at his disposal without paying regard to the compulsion of logic."* [0, p.127]

Freud's later image of childhood was different, with more emphasis on fears, frustrations, and the oppression of a growing self-image that emerges from internal models of authority and attachment figures. In the present essay's conception of the growth of logic in the child, I suggest a comparable self-image of Rationality -- only here with less need for an external human model, because confusion automatically imposes its own sanctions.

It was not quite accurate to say that Freud never returned to the subject, for in 1927 [9] he published a brief essay in which, still regarding jokes as a source of pleasure, he now portrays humor (in contrast) as a way to ward off suffering. The super-ego, like a parent, comforts the frightened childlike ego, repudiating reality by suggesting that however dangerous the world may seem, it is nothing but a game for children. Freud is troubled, though, that this seems out of character for the superego; perhaps the thesis of this essay resolves that. In any case, all this impinges on the area of "adult theory" that I hesitate to discuss here, for the reasons noted in {16}.

[Note 1] *The Society of Mind.* Some of the ideas in this essay originated in my earlier work with Seymour Papert, especially the ideas about the construction of the mind through exploitation of different agencies by one another. The present paper, along with [4], [5], [6] and [12] are all related, but I have yet to attempt a single, comprehensive account. The key idea is to reject the conventional view of the mind as a Single Agent that either thinks of something or doesn't; rather, the mind is composed of many smaller minds, themselves composed of yet smaller ones. It would mean little to talk about what these separately "think", for each becomes specialized to perform functions meaningful only vis-a-vis those few others that it has connections with. The phenomenological observations that a "person" makes about himself emerge in a very indirect way from those interactions.

One (of many) reasons to consider decentralized psychological theories is that they seem potentially better able to provide for mechanisms that exploit *knowledge about knowledge* than

do the control structures that have become traditional in the literatures of AI and of Cognitive Psychology. It is perfectly understandable that the early years of the "Information Processing Approach" should have focussed on the astonishing power of the newly-invented single-process serial computer. But "meta-knowledge" is not easily accommodated by such machines, and I see the strain showing in attempts to realize more ambitious ideas about intelligent learning -- *viz*, in [10] and [11].

[Note 2] *Partial Mental state.* I don't mean to suggest that the *entire* brain ever repeats the same global state. One could say but little about "mental states" if one imagined the Mind to be a single, unitary thing. But if we envision a mind (or brain) as composed of many partially autonomous "agents", then we can talk of repeating a "partial mental state" -- that is, a *subset of the states of those agents.* This is discussed more precisely in [6]. The notion of partial mental state allows one to speak of entertaining *several partial states at once* -- to the extent they are compatible -- that is, they do not assign different states to the same individual agents. And even when they conflict, the concept still has meaning if such conflicts can be settled within the Society. In [6] I argue that such local mechanisms for conflict resolution could be the antecedents of what we know later as *reasoning* -- useful ways to combine different fragments of knowledge.

[Note 3] *Consciousness.* As I see it, "consciousness" phenomena can emerge from the operation of a "self-referent" mechanism when it tries to account for some of what it itself is doing. I doubt we possess any especially direct and powerful ways to do this, so we probably do it much as we understand anything else -- that is, by making and refining models that may never be particularly accurate. Technically, discussing such matters is messy because of the web of different meanings for "self-reference" itself. Should we call a system self-referent just because it operates on data derived from its own operation? It might be considered so if it can be seen as trying to describe itself. The self-reference of planning -- as when one says *"I shall do X"* -- is different from simply expecting to do X, because it entails (usually) subsidiary plans like *"I shall remain firm and not allow Y to cause me to 'change my mind'"*. Should we call a system self-referent when, only by accident, a structure inside it happens to resemble (in an outside observer's opinion) a larger-scale description of itself? In any case it seems quite clear that our psychological self-models are far from technically accurate -- yet serve a variety of social-heuristic-survival needs. In [12] I discussed how models that are technically quite wrong can still be useful; in particular it suggests an explanation of the illusion of free will.

In general, I see little reason to suppose that the "conscious" parts of our minds have any direct access to our own "top-level goals." -- or even to suppose that any such "top level" exists, i.e., that the mental society has a strict hierarchy. Even if there were, Evolution probably would have found a way to block the rest of the mind from being able to examine it too closely -- if only to keep the logical bull out of the teleological china shop. {10} In any case I suspect that if there *is* a top-level it consists of agencies access to which would reveal nothing intelligible (for reasons described in {12}). In [5] I discuss some possible mechanisms that might be "central" in thinking, because their use would be as limited, scarce resources. These might indeed play roles in our articulate, self-model formulations but, again, those models could be very unrealistic.

[Note 4] *Non-monotonic logic.* I discussed these problems with logic very briefly in [4]. Doyle's 1980 thesis [11] has a very imaginative discussion of such matters. One might ask whether the safe regions are like separate islands, or is logic generally safe if we avoid scattered pitfalls.

[Note 5] *Confusion.* When a person can say "I'm confused," he is a large step beyond merely being confused because, presumably, he has gone on enough to have recognized a somewhat specific metapsychological condition.

[Note 6] *Familiar.* Schank [13] points out that it is only when one doesn't quite recognize something that one thinks *"that looks familiar".* When something is decisively recognized there is no such phenomenology; one notices the matching process only when the match is imperfect and strained.

[Note 7] *Ambiguity.* All but the most childish jokes have two or more meanings "condensed" into one expression or situation. It is commonplace to wonder about non-humorous ambiguities of words and phrases, and how the mind decides which thought is intended. It is not so often realized that *thoughts themselves can be ambiguous* -- because a (partial) mental state can be a precursor of many others, depending on the computational context within the mind. [5]

[Note 8] *Excuses.* This is oversimplified in many ways. Defaults are not mere conveniences; they are perhaps our most powerful means for making generalizations, because they activate whatever is "typical" of some class by activating the memory of a typical individual of that class. And even when they are not useful, such assumptions are usually innocuous -- provided that they are weakly enough attached so that they are easily displaced by "reality". However, this hypothetical default mechanism has many potential bugs of its own. In particular, one must not make mistakes about "supported-by" for all sorts of personal safety reasons. If there is no visible support *and* no intervening object then this is a levitation absurdity.

I suspect that *ethnic humor* exemplifies a larger-scale sociobiological bug that emerges from the frame mechanism. Why are nonsense jokes so often used also to deride alien social groups? A popular theory is that it provides an opportunity to display aggression, and no doubt this is true. But there may also be a more technical reason. I argue in [4] that *it is hard to understand a story about a person unless one is provided with a specific person-frame -- it does not matter how stereotyped.* Communication is simplified when the listener does not have to choose a frame for himself. Thus bigotry may emerge spontaneously as a side-effect of this circumstance of representation; when, for example one makes a joke about stupidity, it is psycho-computationally convenient to project it onto some stereotype -- preferably alien to avoid conflict with known reality. Obviously, this may become a runaway process, as that stereotype accumulates undeserved absurdities. Sooner or later one loses track of the humorous origin of that structure. It will be hard to eradicate prejudice without understanding the importance (and value) of stereotypes in ordinary thinking.

[Note 9] Many of the ideas about the importance of discerning differences came as early as [14] and [15], but [8] had the first clear idea of a difference-network. More recently, in [16] Winston has considered using frames for more complex analogies. The serious study of "bugs" was perhaps first considered in the work of Papert on early education [17] and then in the doctoral theses of Sussman [18] and Goldstein [19].

[Note 10] *Defense.* Lewis Thomas remarks in [20] that philosophers and linguists *"are compelled to use as their sole research instrument the very apparatus they wish to study."* He recounts that, while listening to a proof of Godel's theorem: "just as I was taking it all in, I suddenly felt something like the flicking of a mercury wall switch and it all turned to nonsense in my head" and suggests -- presuming one can tell when this scientist-poet is serious -- that *there is a "scrambler in the brain, a protective device preserving the delicate center of the mechanism of language against tinkering and meddling, shielding the mind against information with which it has no intention of getting involved."* Perhaps he's right.

[Note 11] *Spinach.* A reader mentioned that she heard this joke about *brocolli*, not mayonnaise. This is funnier, because it transfers a plausible mistake into an implausible context. In Freud's version the mistake is already too silly: one *could* mistake spinach for brocolli, but not for mayonnaise. I suspect that Freud *transposed the wrong absurdity* when he determined to tell it himself later on. Indeed, he (p.139) seems particularly annoyed at this joke -- and well he might be if, indeed, he himself damaged it by spoiling the elegance of the frame-shift. I would not mention this were it not for the established tradition of advancing psychiatry by analyzing Freud's own writings.

[Note 12] *Enjoyment.* As Freud demanded, we have to explain why humor is pleasant, yet is so often about unpleasant -- painful, disgusting, or forbidden matters. There is nothing really funny about most good jokes - except perhaps in the skill with which the content is disguised. Now, there is no shortage of explanations of how this reversal of sign might come about: the censor-energy theory, the "I'm glad it didn't happen to me" theory, the minimizing the importance of reality theory, and so forth. Yet the question remains all the more important and difficult because, everyone supposing the matter to be obvious, no one seems to have proposed any sophisticated theories of *pleasure* itself -- so all those commonsense "mini-theories" seem built on sand.

What is pleasure, and why do we like it? It is not a tautology. Clearly pleasure involves a complex web concerned with: learning and goals; with activities one wants to continue and/or repeat; and with anticipations and rehearsals of such. What makes the issue elusive, I think, is that we "sense" pleasure only through an elaborately constructed illusion -- typical of our tendency to represent things to ourselves as though they were more coherent than they really are.

We indulge in such illusions even when we say the seemingly simplest things of the form *"I feel xxx"*. These exemplify a "single agent" concept of mind, an illusion of coherency that denies that within the mind different agencies play different roles at the same moment. For example,

one part may be rewarded for disciplining another. Yet if, as I suppose, the pleasure phenomenon does indeed involve a complex web of different activities, we still need to explain why they seem phenomenologically to have so much in common.

Here is one conjecture: what was initially common to all those activities was only a metapsychological feature -- *they were just the ones most subject to being disturbed by the "noise" of other members of the Society of Mind.* This would create an ecological niche, within that Society, for the emergence of a special brain center that, when activated by any of them, tends to depress all the others. (One does not so welcome the offer of one kind of pleasure, while involved with another.) Once such a center comes into existence, that very fact would facilitate the formation, in a growing mind, of a concept or belief in the commonality of the mechanisms associated with it.

A second conjecture: such a centralization would better enable each member of the mental society to be able to tell when it has done something to satisfy a need of another -- -- without having to understand which one, or any of the what or why of it. But note how my use of the word "satisfy" betrays my intention by intimating that same unintended uniformity. It is perhaps this betrayal that fools us all into thinking it a tautology, unworthy of study, to ask *"why do we like pleasure".*

Seymour Papert has suggested to me yet another, more sociobiological conjecture about how evolution might have gathered these ingredients together and provided them with a single, common, output. It would make it easier for a mother to tell when she has satisfied her child's need of the moment, without having to learn specific signs for each such need. Similarly, this would help one person tell when he is convincing another, and so forth. All that is accomplished by attaching a suitable collection of different internal processes to one single "consummatory act" -- to use Tinbergen's term.

[Note 13] My daughter Julie watched "yields truth" for a while and then said, *"well, it's not exactly a paradox, but it does keep saying that it's true, back and forth."*

[Note 14] *Displacement.* Lorenz [21] and Tinbergen [22] frequently observed peculiar, seemingly pointless behaviors when an animal is poised between fight and flight. What better time to consider negotiating? So, one might *a priori* expect to find ambiguities in the primordial germs of social sign-systems.

[Note 15] I don't mean all this to seem pessimistic. It is not necessary, in understanding something, to have all one knows about it active in the mind at one time. One *does* need to have access to fragments of maps, at various levels of detail, of what one knows. The thesis of Kuipers [23], which proposes a theory of how a person's knowledge of a city might be represented in computational terms, might be re-interpreted as a metaphor for how minds might deal with their own knowledge.

[Note 16] *Theories.* There is, I think, a special problem in making theories about the psychology of adults, in whom cultural evolution has had its full interaction with individual and organic evolution. Consider that complex of laughing, smiling, good-natured interactions called "good humor", in which we as often enjoy engaging and surprising frame-shifts of high quality as we enjoy nonsense worthy only of being suppressed. The joking ambiance is used as much to develop analogies as to restrict them -- and my distinction between positive and negative seems to fall apart. If anything remains uniform in the different varieties of humor, it is perhaps only that highlighting of manipulating unexpected frame-shifts -- but in a bewildering variety of ways. Does this mean the theory is refuted? I think not, but to explain why I must digress to discuss "adult development."

No, a baby theory is not necessarily refuted by an adult counterexample Most psychologists have not sufficiently appreciated the full power of an adult intellect to re-construct itself -- to exploit and rearrange its earlier components. Stage after stage of intricate developments are superimposed, in which both internal and cultural influences modify earlier ones. This process would be complicated enough if it were spontaneous -- that is, if only internal factors were involved. But much of it is also socially institutionalized; everyone in the culture "knows" which concerns, and which manners of behavior, are "appropriate" for four-year-olds, nine-year-olds, college students, or professors of philosophy.

The most powerful theoreticians of developmental psychology have struggled to untangle the different principles and forms of these influences; still only the surface has been touched. We yet know far too little to confidently declare that such-and-such a behavioral exhibition either illustrates or refutes a given psychogenetic hypothesis. In my view, the most profitable activity in the present era is to experiment, not on people to see if an hypothesis about the mind is true or false, but on computers to see if a proposed theory-fragment can be made part of a system that shows mind-like activity. To be sure, this can never show, by itself, that the theory in question then must resemble a human mechanism -- but it seems hardly worth trying to verify *that* until a theory shows signs of meeting a first condition -- that it be capable of contributing to life-like activity. In short, we are unlikely to discover much of that which is, until we discover more of that which could be.

[Note 17] *Attachment.* Of course, the sociobiology of reproduction is far more complicated than suggested here. Provisions for child-raising conflict in many ways with provisions for gene-dissemination. Many different sociobiological islands of stability exist, both potentially and actually.

An "acquisition envelope" hypothesis: Here is a different example of how both individual and social development might be influenced by a genetic control of a learning-rate. It is a commonplace observation that persons who learn second languages after adolescence rarely acquire the phonetic competence of native speakers; they rarely come to speak the new language "without an accent". When told to pronounce something "like *this*, not like *that*," they seem to sense too little difference to know what changes to make. I conjecture that this reflects a post -

pubertal change in a brain mechanism -- and may illustrate an important way for genetic control to affect cognitive development.

More precisely, the conjecture is that (i) phonetic learning occurs in some particular brain structure whose capacity to learn new discriminations is (ii) shut off by a genetic mechanism linked to pubertal changes. *It is linked to puberty because that is the biological moment when one's role shifts from learning to teaching*. Its "evolutionary purpose" is to prevent the parent from learning the child's language; this makes the child learn the adult language.

After all, a young parent's principal goal is not language instruction -- it is communication. If it were easy for the parent to adopt the child's idiosyncratic phonology, that's what would happen! But then the parent would learn the child's language -- and the child would have less drive, or opportunity, to learn the adult's. And, over the span of generations, it would be hard for any common social language to develop at all!

When we talk of innate vs. acquired aspects of development, we must face the problem of encoding for structures whose acquisition cannot be genetically anticipated -- e.g., details of cognitive or linguistic structure. Our idea is first to look instead for ways for genetics to affect directly the "acquisition envelopes" -- *the control structures that mediate how things are learned.* The foregoing hypothesis illustrates, I think, one way that brain genetics might circumvent the complexity of devising direct constraints on the representations of not-yet-acquired cognitive structures.

REFERENCES

[0] Freud, Sigmund. *Jokes and Their Relation to the Unconscious*, 1905, (transl. Strachey) Standard Edition, vol.8, Hogarth Press, 1957.

[1] Doyle, Jon. *Truth Maintenance Systems for Problem Solving.* M.I.T., Artificial Intelligence Laboratory, AI/TR-419. January 1978.

[2] Kornfeld, William A. *Using Parallel Processes for Problem Solving.* M.I.T., Artificial Intelligence Laboratory, AI Memo 561. Cambridge, Ma., Dec. 1979.

[3] Korzybski, Alfred. *Science and Sanity.* Lancaster, Pa.: Science Press, 1941.

[4a] Minsky, Marvin. *A Framework for Representing Knowledge.* M.I.T., Artificial Intelligence Laboratory, AI Memo 306. Cambridge, Ma., June 1974.

[4b] Minsky, Marvin. "A Framework for Representing Knowledge" (condensed version). *The Psychology of Computer Vision.* Edited by P. H. Winston. New York: McGraw-Hill, 1975.

[5] Minsky, Marvin. "Plain Talk About Neurodevelopmental Epistemology." *Proceedings of the Fifth International Joint Conference on Artificial Intelligence.* Cambridge, Ma., August 1977. Condensed version in *Artificial Intelligence*, edited by Winston and Brown, Vol. 1, MIT Press, 1979.

[6] Minsky, Marvin. "K-lines: A Theory of Memory." *Cognitive Science*, Vol. 4, No. 2 (April 1980), 117-133.

[7] Wittgenstein, L. *Philosophical Investigations.* Oxford, 1953.

[8] Winston, P.H. "Learning Structural Descriptions by Examples." *Psychology of Computer Vision.* Edited by P. H. Winston. New York: McGraw-Hill, 1975.

[9] Freud, Sigmund. *Humour*, 1927, (transl. Strachey) Standard Edition, Vol. 21, Hogarth Press, 1957, 161-166.

[10] Davis, Randall. "Meta-rules: Reasoning about Control." To appear in *Artificial Intelligence*, 1981.

[11] Doyle, Jon. *A Model for Deliberation, Action and Introspection.* Ph.D Thesis, M.I.T.,Artificial Intelligence Laboratory, Cambridge, Ma., August 1980.

[12] Minsky, Marvin. "Matter, Mind and Models." *Proceedings of IFIP Congress 1965.* May, 1965, 45-49. Reprinted in *Semantic Information Processing*, MIT Press, 1968.

[13] Schank, Roger. "Language and Memory." *Cognitive Science*, Vol. 4, 1980, 243-284.

[14] Newell, A; Shaw, J.C.; and Simon, H.A.. *Preliminary Description of General Problem Solving Program*, (GPS-1). CIP Working Paper No. 7, December 1957.

[15] Minsky, Marvin. *Heuristic Aspects of the Artificial Intelligence Problem.* Lincoln Laboratory, M.I.T., Lexington, Mass. Group Report No. 34-55. ASTIA Doc.No. AS236885, December 1956.

[16] Winston, Patrick H. *Learning by Understanding Analogies.* M.I.T.,Artificial Intelligence Laboratory, AI memo 520. Cambridge, Ma., June 1979.

[17] Papert, Seymour. *Mindstorms. Children, Computers and Powerful Ideas.* New York: Basic Books, 1980.

[18] Sussman, Gerald J. *A Computational Model of Skill Acquisition.* Ph.D.Thesis, M.I.T., Artificial Intelligence Laboratory, IA/TR-297. Cambridge, Ma., August 1973.

[19] Goldstein, Ira P. *Understanding Simple Picture Programs.* Ph.D. Thesis,M.I.T., Artificial Intelligence Laboratory, IA/TR-294. Cambridge, Ma., April 1974.

[20] Thomas, Lewis. "The Scrambler in the Mind." *The Medusa and the Snail.* New York: Bantam Books, 1980.

[21] Lorenz, Konrad. *King Solomon's Ring.* New York: Thomas J. Crowell, 1961.

[22] Tinbergen, Niko. *The Study of Instinct.* Oxford University Press, 1951.

[23] Kuipers, Benjamin. *Representing Knowledge of Large-Scale Space.* Ph.D. Thesis, M.I.T., Artificial Intelligence Laboratory, IA/TR-418. Cambridge, Ma., July 1978.

ACKNOWLEDGMENTS

I thank Howard Cannon, Danny Hillis, William Kornfeld, David Levitt, Gloria Rudisch, and Richard Stallman for suggestions. Gordon Oro provided the dog-joke.