The Logical Categories of Learning and Communication

Gregory Bateson

"Steps to an Ecology of Mind"

originally published 1972

The University of Chicago Press edition 2000 - pag. 279-308

Gregory Bateson was the son of pioneer geneticist William Bateson and the husband of renowned anthropologist Margaret Mead.

His work spanned many fields, from anthropology and communication theory to his studies of alcoholism and schizophrenia at the Veterans Administration Hospital in Palo Alto, California. His classic works include Naven, Balinese Character, coauthored with Margaret Mead, and Mind and Nature.

The Logical Categories of Learning and Communication*

All species of behavioral scientists are concerned with "learning" in one sense or another of that word. Moreover, since "learning" is a communicational phenomenon, all are affected by that cybernetic revolution in thought which has occurred in the last twenty-five years. This revolution was triggered by the engineers and communication theorists but has older roots in the physiological work of Cannon and Claude Bernard, in the physics of Clarke Maxwell, and in the mathematical philosophy of Russell and Whitehead. Insofar as behavioral scientists still ignore the problems of *Principia Maihematica,*[1] they can claim approximately sixty years of obsolescence.

It appears, however, that the barriers of misunderstanding which divide the various species of behavioral scientists can be illuminated (but not eliminated) by an application of Russell's Theory of Logical Types to the concept of "learning" with which all are concerned. To attempt this illumination will be a purpose of the present essay.

The Theory of Logical Types

First, it is appropriate to indicate the subject matter of the Theory of Logical Types: the theory asserts that no class can, in formal logical or mathematical discourse, be a member of itself; that a class of classes cannot be one of the classes which are its members; that a name is not the thing named; that "John Bateson" is the class of which that boy is the unique member; and so forth. These assertions may seem trivial and even obvious, but we shall see later that it is not at all unusual for the theorists of behavioral science to commit errors which are precisely analogous to the error of classifying the name with the thing named—or eating the menu card instead of the dinner—an error of *logical typing.*

Somewhat less obvious is the further assertion of the theory: that a class cannot be one of those items which are correctly classified as its nonmembers. If we classify chairs together to constitute the class of chairs, we can go on to note that tables and lamp shades are members of a large class of "nonchairs," but we shall commit an error in formal discourse if we count the *class of chairs* among the items within the class of nonchairs.

Inasmuch as no class can be a member of itself, the class of nonchairs clearly cannot be a nonchair. Simple considerations of symmetry may suffice to convince the nonmathematical reader: *(a)* that the class of chairs is of the same order of abstraction *(i.e.,* the same logical type) as the class of nonchairs; and further, *(b)* that if the class of chairs is not a chair, then, correspondingly, the class of nonchairs is not a nonchair.

Lastly, the theory asserts that if these simple rules of formal discourse are contravened, paradox will be generated and the discourse vitiated.

The theory, then, deals with highly abstract matters and was first derived within the abstract world of logic. In that world, when a train of propositions can be shown to generate a paradox, the entire structure of axioms, theorems, etc, involved in generating that paradox is thereby negated and reduced to nothing. It is as if it had never been. But in the real world (or at least in our descriptions of it), there is always *time,* and nothing which has been can ever be totally negated in this way. The computer which encounters a paradox (due to faulty programming) does not vanish away.

The "if . . . then . . ." of logic contains no time. But in the computer, cause and effect are used to *simulate* the "if then . . ." of logic; and all sequences of cause and effect necessarily involve time. (Conversely, we may say that in scientific explanations the "if . . . then . . ." of logic is used to simulate the "if . . . then . . ." of cause and effect.)

The computer never truly encounters logical paradox, but only the simulation of paradox in trains of cause and effect. The computer therefore does not fade away. It merely oscillates.

In fact, there are important differences between the world of logic and the world of phenomena, and these differences must be allowed for whenever we base our arguments upon the partial but important analogy which exists between them.

It is the thesis of the present essay that this partial analogy can provide an important guide for behavioral scientists in their classification of phenomena related to learning. Precisely in the field of animal and mechanical communication something like the theory of types must apply.

Questions of this sort, however, are not often discussed in zoological laboratories, anthropological field camps, or psychiatric conventions, and it is necessary therefore to demonstrate that these abstract considerations are important to behavioral scientists.

Consider the following syllogism:

(a) Changes in frequency of items of mammalian behavior can be described and

predicted in terms of various "laws" of reinforcement.

(b) "Exploration" as observed in rats is a category, or class, of mammalian behavior.

(c) Therefore, changes in frequency of "exploration" should be describable in terms of the same "laws" of reinforcement.

Be it said at once: first, that empirical data show that the conclusion (c) is untrue; and second, that if the conclusion (c) were demonstrably true, then either *(a)* or *(b)* would be untrue.[2]

Logic and natural history would be better served by an expanded and corrected version of the conclusion (c) some-what as follows:

(c) If, as asserted in *{b),* "exploration" is not an *item* of mammalian behavior but is a *category* of such items, then no descriptive statement which is true of *items* of behavior can be true of "exploration." If, however, descriptive statements which are true of items of behavior are also true of "exploration," then "exploration" is an item and not a category of items.

The whole matter turns on whether the distinction between a *class* and its *members* is an ordering principle in the behavioral phenomena which we study.

In less formal language: you can reinforce a rat (positively or negatively) when he investigates a particular strange object, and he will appropriately learn to approach or avoid it. But the very purpose of exploration is to get information about which objects should be approached and which avoided. The discovery that a given object is dangerous is therefore a *success* in the business of getting information. The success will not discourage the rat from future exploration of other strange objects.

A priori it can be argued that all perception and all response, all behavior and all classes of behavior, all learning and all genetics, all neurophysiology and endocrinology, all organization and all evolution—one entire subject matter—must be regarded as communicational in nature, and therefore subject to the great generalizations or "laws" which apply to communicative phenomena. We therefore are warned to expect to find in our data those principles of order which fundamental communication theory would propose. The Theory of Logical Types, Information Theory, and so forth, are expectably to be our guides.

The "Learning" of Computers, Rats, and Men

The word "learning" undoubtedly denotes *change* of some kind. To say *what kind* of change is a delicate matter.

However, from the gross common denominator, "change," we can deduce that our descriptions of "learning" will have to make the same sort of allowance for the varieties of logical type which has been routine in physical science since the days of Newton. The simplest and most familiar form of change is *motion,* and even if we work at that very simple physical level we must structure our descriptions in terms of "position or zero motion," "constant velocity," "acceleration," "rate of change of acceleration," and so on.[3]

Change denotes process. But processes are themselves subject to "change." The process may accelerate, it may slow down, or it may undergo other types of change such that we shall say that it is now a "different" process.

These considerations suggest that we should begin the ordering of our ideas about "learning" at the very simplest level.

Let us consider the case of specificity of response, or *zero learning*. This is the case in which an entity shows minimal change in its response to a repeated item of sensory input. Phenomena which approach this degree of simplicity occur in various contexts:

(a) In experimental settings, when "learning" is complete and the animal gives approximately 100 percent correct responses to the repeated stimulus.

(b) In cases of habituation, where the animal has ceased to give overt response to what was formerly a disturbing stimulus.

(c) In cases where the pattern of the response is minimally determined by experience and maximally determined by genetic factors.

(d) In cases where the response is now highly stereotyped.

(e) In simple electronic circuits, where *the circuit structure is not itself subject to change resulting from the passage of impulses within the circuit—i.e.,* where the causal links between "stimulus" and "response" are as the engineers say "soldered in."

In ordinary, nontechnical parlance, the word "learn" is often applied to what is here called "zero learning," *i.e.,* to the simple receipt of information from an external event, in such a way that a similar event at a later (and appropriate) time will convey the same information: I "learn" from the factory whistle that it is twelve o'clock.

It is also interesting to note that within the frame of our definition many very simple mechanical devices show at least the phenomenon of zero learning. The question is not, "Can machines learn?" but what level or order of learning does a given machine achieve? It is worth looking at an extreme, if hypothetical, case:

The "player" of a Von Neumannian game is a mathematical fiction, comparable to the Euclidean straight line in geometry or the Newtonian particle in physics. By definition, the "player" is capable of all computations necessary to solve whatever problems the events of the game may present; he is incapable of not performing these computations whenever they are appropriate; he always obeys the findings of his computations. Such a "player" receives information from the events of the game and acts appropriately upon that information. But his learning is limited to what is here called zero learning.

An examination of this formal fiction will contribute to our definition of zero learning.

(1) The "player" may receive, from the events of the game, information of higher or lower logical type, and he may use this information to make decisions of higher or lower type. That is, his decisions may be either strategic or tactical, and he can identify and respond to indications of both the tactics and the strategy of his opponent. It is, however, true that in Von Neumann's formal definition of a "game," all problems which the game may present are conceived as computable, *i.e.*, while the game may contain problems and information of many different logical types, the hierarchy of these types is strictly finite.

It appears then that a definition of zero learning will not depend upon the logical typing of the information received by the organism nor upon the logical typing of the adaptive decisions which the organism may make. A very high (but finite) order of complexity may characterize adaptive behavior based on nothing higher than zero learning.

(2) The "player" may compute the value of information which would benefit him and may compute that it will pay him to acquire this information by engaging in "exploratory" moves. Alternatively, he may make delaying or tentative moves while he waits for needed information.

It follows that a rat engaging in exploratory behavior might do so upon a basis of zero learning.

(3) The "player" may compute that it will pay him to make random moves. In the

game of matching pennies, he will compute that if he selects "heads" or "tails" at random, he will have an even chance of winning. If he uses any plan or pattern, this will appear as a pattern or redundancy in the sequence of his moves and his opponent will thereby receive information. The "player" will therefore elect to play in a random manner.

(4) The "player" is incapable of "error." He may, for good reason, elect to make random moves or exploratory moves, but he is by definition incapable of "learning by trial and error."

If we assume that, in the name of this learning process, the word "error" means what we meant it to mean when we said that the "player" is incapable of error, then "trial and error" is excluded from the repertoire of the Von Neumannian player. In fact, the Von Neumannian "player" forces us to a very careful examination of what we mean by "trial and error" learning, and indeed what is meant by "learning" of any kind. The assumption regarding the meaning of the word "error" is not trivial and must now be examined.

There is a sense in which the "player" can be wrong. For example, he may base a decision upon probabilistic considerations and then make that move which, in the light of the limited available information, was most probably right. When more information becomes available, he may discover that that move was wrong. But *this discovery can contribute nothing to his future skill.* By definition, the player used correctly all the *available* information. He estimated the probabilities correctly and made the move which was most probably correct. The discovery that he was wrong in the particular instance can have no bearing upon future in-stances. When the same problem returns at a later time, he will *correctly* go through the same computations and reach the same decision. Moreover, the set of alternatives among which he makes his choice will be the same set—and correctly so.

In contrast, an organism is capable of being wrong in a number of ways of which the "player" is incapable. These wrong choices are appropriately called "error" when they are of such a kind that they would provide information to the organism which might contribute to his future skill. These will all be cases in which some of the available information was either ignored or incorrectly used. Various species of such profitable error can be classified.

Suppose that the external event system contains details which might tell the organism: (a) from what set of alternatives he should choose his next move; and *(b)* which member of that set he should choose. Such a situation permits two *orders* of error:

(1) The organism may use correctly the information which tells him from

what set of alternatives he should choose, but choose the wrong alternative within this set; or

(2) He may choose from the wrong set of alternatives. (There is also an interesting class of cases in which the sets of alternatives contain common members. It is then possible for the organism to be "right" but for the wrong reasons. This form of error is inevitably self-reinforcing.)

If now we accept the overall notion that all learning (other than zero learning) is in some degree stochastic *(i.e.*, contains components of "trial and error"), it follows that an ordering of the processes of learning can be built upon an hierarchic classification of the types of error which are to be corrected in the various learning processes. Zero learning will then be the label for the immediate base of all those acts (simple and complex) which are not subject to correction by trial and error. Learning I will be an appropriate label for the revision of choice within an unchanged set of alternatives; Learning II will be the label for the revision of the *set* from which the choice is to be made; and so on.

Learning I

Following the formal analogy provided by the "laws" of motion *(i.e.*, the "rules" for describing motion), we now look for the class of phenomena which are appropriately described as *changes* in zero learning (as "motion" describes change of position). These are the cases in which an entity gives at Time 2 a different response from what it gave at Time 1, and again we encounter a variety of cases variously related to experience, physiology, genetics, and mechanical process:

(a) There is the phenomenon of habituation—the change from responding to each occurrence of a repeated event to not overtly responding. There is also the extinction or loss of habituation, which may occur as a result of a more or less long gap or other interruption in the sequence of repetitions of the stimulus event. (Habituation is of especial interest. Specificity of response, which we are calling zero learning, is characteristic of all protoplasm, but it is interesting to note that "habituation" is perhaps the only form of Learning I which living things can achieve without a neural circuit.)

(b) The most familiar and perhaps most studied case is that of the classical Pavlovian conditioning. At Time 2 the dog salivates in response to the buzzer; he did not do this at Time 1.

(c) There is the "learning" which occurs in contexts of instrumental reward and instrumental avoidance.

(d) There is the phenomenon of rote learning, in which an item in the behavior of the organism becomes a stimulus for another item of behavior.

(e) There is the disruption, extinction, or inhibition of "completed" learning which may follow change or absence of reinforcement.

In a word, the list of Learning I contains those items which are most commonly called "learning" in the psycho-logical laboratory.

Note that in all cases of Learning I, there is in our description an assumption about the "context." This assumption must be made explicit. The definition of Learning I assumes that the buzzer (the stimulus) is somehow the "same" at Time 1 and at Time 2. And this assumption of "sameness" must also delimit the "context," which must (theoretically) be the same at both times. It follows that the events which occurred at Time 1 are not, in our description, included in our definition of the context at Time 2, because to include them would at once create a gross difference between "context at Time 1" and "context at Time 2." (To paraphrase Heraclitus: "No man can go to bed with the same girl for the first time twice.")

The conventional assumption that context can be repeated, at least in some cases, is one which the writer adopts in this essay as a cornerstone of the thesis that the study of behavior must be ordered according to the Theory of Logical Types. *Without the* assumption of repeatable context (and the hypothesis that *for the organisms* which we study the sequence of experience is really somehow punctuated in this manner), it would follow that all "learning" would be of one type: namely, all would be zero learning. Of the Pavlovian experiment, we would simply say that the dog's neural circuits contain "soldered in" from the beginning such characteristics that in Context A at Time 1 he will not salivate, and that in the totally different Context B at Time 2 he will salivate. What previously we called "learning" we would now describe as "discrimination" between the events of Time 1 and the events of Time 1 *plus* Time 2. It would then follow logically that all questions of the type, "Is this behavior 'learned' or 'innate?" should be answered in favor of genetics.

We would argue that without the assumption of repeatable context, our thesis falls to the ground, together with the whole general concept of "learning." If, on the other hand, the assumption of repeatable context is accepted as somehow true of the organisms which we study, then the case for logical typing of the phenomena of learning necessarily stands, because the notion "context" is itself subject to logical typing.

Either we must discard the notion of "context," or we retain this notion and, with it, accept the hierarchic series— stimulus, context of stimulus, context of context of stimulus, etc. This series can be spelled out in the form of a hierarchy of logical types as follows:

Stimulus is an elementary signal, internal or external. Context of stimulus is a metamessage which *classifies* the elementary signal.

Context of context of stimulus is a meta-metamessage which classifies the metamessage. And so on.

The same hierarchy could have been built up from the notion of "response" or the notion of "reinforcement."

Alternatively, following up the hierarchic classification of errors to be corrected by stochastic process or "trial and error," we may regard "context" as a collective term for all those events which tell the organism among what *set* of alternatives he must make his next choice.

At this point it is convenient to introduce the term "context marker." An organism responds to the "same" stimulus differently in differing contexts, and we must therefore ask about the source of the organisms's information. From what percept does he know that Context A is different from Context B?

In many instances, there may be no specific *signal* or label which will classify and differentiate the two contexts, and the organism will be forced to get his information from the actual congeries of events that make up the context in each case. But, certainly in human life and probably in that of many other organisms, there occur signals whose major function is to *classify* contexts. It is not unreasonable to sup-pose that when the harness is placed upon the dog, who has had prolonged training in the psychological laboratory, he knows from this that he is now embarking upon a series of contexts of a certain sort. Such a source of information we shall call a "context marker," and note immediately that, at least at the human level, there are also "markers of contexts of contexts." For example: an audience is watching *Hamlet* on the stage, and hears the hero discuss suicide in the con-text of his relationship with his dead father, Ophelia, and the rest. The audience members do not immediately telephone for the police because they have received information about the context of Hamlets context. They know that it is a "play" and have received this information from many "markers of context of context"—the playbills, the seating arrangements, the curtain, etc, etc. The "King," on the other hand, when he lets his conscience be pricked by the play within the play, is ignoring many "markers of context of context."

At the human level, a very diverse set of events falls within the category of "context markers." A few examples are here listed:

(a) The Pope's throne from which he makes announcements *ex cathedra,* which announcements are thereby endowed with a special order of validity.

(b) The placebo, by which the doctor sets the stage for a change in the patients subjective experience.

(c) The shining object used by some hypnotists in "inducing trance."

(d) The air raid siren and the "all clear."

(e) The handshake of boxers before the fight.

(f) The observances of etiquette.

These, however, are examples from the social life of a highly complex organism, and it is more profitable at this stage to ask about the analogous phenomena at the pre-verbal level.

A dog may see the leash in his master*s hand and act as if he knows that this indicates a walk; or he may get information from the sound of the word "walk" that this type of context or sequence is coming.

When a rat starts a sequence of exploratory activities, does he do so in response to a "stimulus?" Or in response to a context? Or in response to a context marker?

These questions bring to the surface formal problems about the Theory of Logical Types which must be discussed. The theory in íts original form deals only with rigorously digital communication, and it is doubtful how far it may be applied to analogue or iconic systems. What we are here calling "context markers" may be digital *(e.g.,* the word "walk" mentioned above); or they may be analogue signals — a briskness in the master's movements may indicate that a walk is pending; or some *part* of the coming context may serve as a marker (the leash as a part of the walk); or in the extreme case, the walk itself in all its complexity may stand for itself, with no label or marker between the dog and the experience. The perceived event itself may communicate its own occurrence. In this case, of course, there can be no error of the "menu card" type. Moreover, no paradox can be generated because in purely analogue or iconic communication there is no signal for "not."

There is, in fact, almost no formal theory dealing with analogue communication and, in particular, no equivalent of Information Theory or Logical Type Theory, This gap in formal knowledge is inconvenient when we leave the rarified world of logic and mathematics and come face to face with the phenomena of natural history. In the natural world, communication is rarely either purely digital or purely analogic Often discrete digital pips are combined together to make analogic pictures as in the printer's halftone block; and sometimes, as in the matter of context markers, there is a continuous gradation from the ostensive through the iconic to the purely digital. At the digital end of this scale all the theorems of information theory have their full force, but at the ostensive and analogic end they are meaningless.

It seems also that while much of the behavioral communication of even higher mammals remains ostensive or analogic, the internal mechanism of these creatures has become digitalized at least at the neuronal level. It would seem that analogic communication is in some sense more primitive than digital and that there is a broad evolutionary trend toward the substitution of digital for analogic mechanisms. This trend seems to operate faster in the evolution of internal mechanisms than in the evolution of external behavior. Recapitulating and extending what was said above:

(a) The notion of repeatable context is a necessary premise for any theory which defines "learning" as *change.*

(b) This notion is not a mere tool of our description but contains the implicit hypothesis that for the organisms which we study, the sequence of life experience, action, etc, is somehow segmented or punctuated into subsequences or "contexts" which may be equated or differentiated by the organism.

(c) The distinction which is commonly drawn between perception and action, afferent and efferent, input and out put, is for higher organisms in complex situations not valid. On the one hand, almost every item of action may be reported either by external sense or endoceptive mechanism o the C.N.S., and in this case the report of this item becomes an input. And, on the other hand, in higher organisms, perception is not by any means a process of mere passive receptivity but is at least partly determined by efferent control from higher centers. Perception, notoriously, can be changed by experience. In principle, we must allow both for the possibility that every item of action or output may create an item of input; and that percepts may in some cases partake of the nature of output. It is no accident that almost all sense organs are used for the emission of signals between organisms. Ants communicate by their antennae; dogs by the pricking of their ears; and so on.

(d) In principle, even in zero learning, any item of experience or behavior may be regarded as either "stimulus" or "response" or as both, according to how the

total sequence is punctuated. When the scientist says that the buzzer is the "stimulus" in a given sequence, his utterance implies an hypothesis about how the organism punctuates that sequence. In Learning I, every item of perception or behaviour may be stimulus or response or *reinforcement* according to how the total sequence of interaction is punctuated.

Learning II

What has been said above has cleared the ground for the consideration of the next level or logical type of "learning" which we shall here call Learning II. Various terms have been proposed in the literature for various phenomena of this order. "Deutero-learning,"[4] "set learning,"[5] "learning to learn," and "transfer of learning" may be mentioned.

We recapitulate and extend the definitions so far given:

Zero learning is characterized by *specificity of response,* which—right or wrong—is not subject to correction.

Learning I is *change in specificity of response* by correction of errors of choice within a set of alternatives.

Learning II is *change in the process of Learning /, e.g.,* a corrective change in the set of alternatives from which choice is made, or it is a change in how the sequence of experience is punctuated.

Learning III is *change in the process of Learning II, e.g.,* a corrective change in the system of *sets* of alternatives from which choice is made. (We shall see later that to demand this level of performance of some men and some mammals is sometimes pathogenic.)

Learning IV would be *change in Learning III,* but probably does not occur in any adult living organism on this earth. Evolutionary process has, however, created organisms whose ontogeny brings them to Level III. The combination of phylogenesis with ontogenesis, in fact, achieves Level IV.

Our immediate task is to give substance to the definition of Learning II as "change in Learning I," and it is for this that the ground has been prepared. Briefly, I believe that the phenomena of Learning II can all be included under the rubric of changes in the manner in which the stream of action and experience is segmented or punctuated into contexts together with changes in the use of context markers.

The list of phenomena classified under Learning I includes a considerable (but not exhaustive) set of differently structured contexts. In classical Pavlovian contexts, the contingency pattern which describes the relation between "stimulus" (CS), animals action (CR), and reinforcement (UCS) is profoundly different from the contingency pattern characteristic of instrumental contexts of learning.

In the Pavlovian case: *If* stimulus and a certain lapse of time: *then* reinforcement.

In the Instrumental Reward case: // stimulus and a particular item of behavior; *then* reinforcement.

In the Pavlovian case, the reinforcement is not contingent upon the animals behavior, whereas in the instrumental case, it is. Using this contrast as an example, we say that Learning II has occurred if it can be shown that experience of one or more contexts of the Pavlovian type results in the animals acting in some later context as though this, too, had the Pavlovian contingency pattern. Similarly, if past experience of instrumental sequences leads an animal to act in some later context as though expecting this also to be an instrumental context, we shall again say that Learning II has

occurred.

When so defined, Learning II is adaptive only if the animal happens to be right in its expectation of a given contingency pattern, and in such a case we shall expect to see a measurable *learning to learn*. It should require fewer trials in the new context to establish "correct" behavior. If, on the other hand, the animal is wrong in his identification of the later contingency pattern, then we shall expect a delay of Learning I in the new context. The animal who has had prolonged experience of Pavlovian contexts might never get around to the particular sort of trial-and-error behavior necessary to discover a correct instrumental response.

There are at least four fields of experimentation where Learning II has been carefully recorded:

(a) In human rote learning. Hull[6] carried out very careful quantitative studies which revealed this phenomenon, and constructed a mathematical model which would simulate or explain the curves of Learning I which he recorded. He also observed a second-order phenomenon which we may call "learning to rote learn" and published the curves for this phenomenon in the Appendix to his book. These curves were separated from the main body of the book because, as he states, his mathematical model (of Rote Learning I) did not cover this aspect of the data.

It is a corollary of the theoretical position which we here take that no amount of rigorous discourse of a given logical type can "explain" phenomena of a higher type. Hull's model acts as a touchstone of logical typing, automatically excluding from explanation phenomena beyond its logical scope. That this was so—and that Hull perceived it—is testimonial both to his rigor and to his perspicacity.

What the data show is that for any given subject, there is an improvement in rote learning with successive sessions, asymptotically approaching a degree of skill which varied from subject to subject.

The context for this rote learning was quite complex and no doubt appeared subjectively different to each learner. Some may have been more motivated by fear of being wrong, while others looked rather for the satisfactions of being right. Some would be more influenced to put up a good record as compared with the other subjects; others would be fascinated to compete in each session with their own previous showing, and so on. All must have had ideas (correct or incorrect) about the nature of the experimental setting, all must have had "levels of aspiration," and all must have had previous experience of memorizing various sorts of material. Not one of Hull's subjects could have come into the learning context uninfluenced by previous Learning II.

In spite of all this previous Learning II, and in spite of genetic differences which might operate at this level, all showed improvement over several sessions. This improvement cannot have been due to Learning I because any recall of the specific sequence of syllables learned in the previous session would not be of use in dealing with the new sequence. Such recall would more probably be a hindrance. I submit, therefore, that the improvement from session to session can only be accounted for by some sort of adaptation to the *context* which Hull provided for rote learning.

It is also worth noting that educators have strong opinions about the value (positive or negative) of training in rote learning. "Progressive" educators insist on training in "insight," while the more conservative insist on rote and drilled recall.

(b) The second type of Learning II which has been experimentally studied is called "set learning." The concept and term are derived from Harlow and apply to a rather special case of Learning II. Broadly, what Harlow did was to present rhesus monkeys with more or less complex *gestalten* or "problems." These the monkey had to solve to get a food reward. Harlow showed that if these problems were of similar "set,” *i.e.,* contained similar types of logical complexity, there was a carry-over of learning from one problem to the next. There were, in fact, two orders of contingency patterns involved in Harlow*s experiments: first the overall pattern of instrumentalism *(if* the monkey solves the problem, *then* reinforcement); and second, the contingency patterns of logic within the specific problems.

(c) Bitterman and others have recently set a fashion in experimentation with "reversal learning." Typically in these experiments the subject is first taught a

binary discrimination. When this has been learned to criterion, the meaning of the stimuli is reversed. If X initially "meant" *R1* and Y initially meant R2, then after reversal X comes to mean R2, and Y comes to mean R1. Again the trials are run to criterion when again the meanings are reversed. In these experiments, the crucial question is: Does the subject learn about the reversal? I.*e.*, after a series of reversals, does the subject reach criterion in fewer trials than he did at the beginning of the series?

In these experiments, it is conspicuously clear that the question asked is of logical type higher than that of questions about simple learning. If simple learning is based upon a *set* of trials, then reversal learning is based upon a set of such sets. The parallelism between this relation and Russells relation between "class" and "class of classes" is direct.

(d) Learning II is also exemplified in the well-known phenomena of "experimental neurosis." Typically an animal is trained, either in a Pavlovian or instrumental learning con-text, to discriminate between some X and some Y; *e.g.*, between an ellipse and a circle. When this discrimination has been learned, the task is made more difficult: the ellipse is made progressively fatter and the circle is flattened. Finally a stage is reached at which discrimination is impossible. At this stage the animal starts to show symptoms of severe disturbance.

Notably, (a) a naive animal, presented with a situation in which some X may (on some random basis) mean either *A* or B, does not show disturbance; and *(b)* the disturbance does not occur in absence of the many context markers characteristic of the laboratory situation.[7]

It appears, then, that Learning II is a necessary preparation for the behavioral disturbance. The information, "This is a context for discrimination," is communicated at the beginning of the sequence and *underlined* in the series of stages in which discrimination is made progressively more difficult. But when discrimination becomes impossible, the structure of the context is totally changed. The context markers *(e.g.*, the smell of the laboratory and the experimental harness) now become misleading because the animal is in a situation which demands guesswork or gambling, *not* discrimination. The en-tire experimental sequence is, in fact, a procedure for putting the animal in the wrong at the level of Learning II.

In my phrase, the animal is placed in a typical "double bind," which is expectably schizophrenogenic.[8]

In the strange world outside the psychological laboratory, phenomena which belong to the category Learning II are a major preoccupation of anthropologists, educators, psychiatrists, animal trainers, human parents, and children. All who think about the processes which determine the character of the individual or the processes of change in human (or animal) relationship must use in their thinking a variety of assumptions about Learning II. From time to time, these people call in the laboratory psychologist as a consultant, and then are confronted with a linguistic barrier. Such barriers must always result when, for example, the psychiatrist is talk-ing about Learning II, the psychologist is talking about Learning I, and neither recognizes the logical structure of the difference.

Of the multitudinous ways in which Learning II emerges in human affairs, only three will be discussed in this essay:

(a) In describing individual human beings, both the scientist and the layman commonly resort to adjectives descriptive of "character." It is said that Mr. Jones is dependent, hostile, fey, finicky, anxious, exhibitionistic, narcissistic, passive, competitive, energetic, bold, cowardly, fatalistic, humorous, playful, canny, optimistic, perfectionist, careless, careful, casual, etc. In the light of what has already been said, the reader will be able to assign all these adjectives to their appropriate logical type. All are descriptive of (possible) results of Learning II, and if we would define these words more carefully, our definition will consist in laying down the contingency pattern of that context of Learning I which would expectably bring about that Learning II which would make the adjective applicable.

We might say of the "fatalistic" man that the pattern of his transactions with the environment is such as he might have acquired by prolonged or repeated experience as subject of Pavlovian experiment; and note that this definition of "fatalism" is specific and precise. There are many other forms of "fatalism" besides that which is defined in terms of this particular context of learning. There is, for example, the more complex type characteristic of classical Greek tragedy where a man's own action is felt to aid the inevitable working of fate.

(b) In the punctuation of human interaction. The critical reader will have observed that the adjectives above which purport to describe individual character are really not strictly applicable to the individual but rather describe *transactions* between the individual and his material and human environment. No man is "resourceful" or "dependent" or "fatalistic" in a vacuum. His characteristic, whatever it be, is not his but is rather a characteristic of what goes on between him and something (or somebody) else.

This being so, it is natural to look into what goes on between people, there to find contexts of Learning I which are likely to lend their shape to processes of Learning II. In such systems, involving two or more persons, where most of the important events are postures, actions, or utterances of the living creatures, we note immediately that the stream of events is commonly punctuated into contexts of learning by a tacit agreement between the persons regarding the nature of their relationship—or by context markers and tacit agreement that these context markers shall "mean" the same for both parties. It is instructive to attempt analysis of an ongoing interchange between A and B. We ask about any particular item of A's behavior: Is this item a stimulus for B? Or is it a response of A to something B said earlier? Or is it a reinforcement of some item provided by B? Or is A, in this item, consummating a reinforcement for himself? Etc.

Such questions will reveal at once that for many items of A's behavior the answer is often quite unclear. Or if there be a clear answer, the clarity is due only to a tacit (rarely fully explicit) agreement between A and B as to the nature of their mutual roles, *i.e.,* as to the nature of the contextual structure which they will expect of each other.

If we look at such an exchange in the abstract:

$a_1b_1a_2b_2a_3b_3a_4b_4a_5b_5$....., where the *a's* refer to items of A's behavior, and the *b's* to items of B's behavior, we can take any a_1 and construct around it three simple contexts of learning. These will be:

i. (a_i b_i a_{i+1}), in which a_i is the stimulus for b_i

ii. (b_{i-1} a_i b_i), in which a_i is the response to b_{i-1}, which response B reinforces with b_i

iii. (a_{i-1} b_i-$_1$ a_i), in which a_i is now A’s reinforcement of B's b_{i-1}, which was response to a_{i-1}.

It follows that a_i may be a stimulus for B or it may be A's response to B, or it may be A's reinforcement of B.

Beyond this, however, if we consider the ambiguity of the notions "stimulus" and "response," "afferent" and "efferent"—as discussed above—we note that any a_i may also be a stimulus for A; it may be A’s reinforcement of self; or it may be A’s response to some previous behavior of his own, as is the case in sequences of rote behavior.

This general ambiguity means in fact that the ongoing sequence of interchange between two persons is structured only by the person's own perception of the sequence as a series of contexts, each context leading into the next. The particular manner in which the sequence is structured by any particular person will be determined by that person’s previous Learning II (or possibly by his genetics).

In such a system, words like "dominant" and "submissive", "succoring" and "dependent" will take on definable meaning as descriptions of segments of interchange. We shall say that "A dominates B" if A and B show by their behavior that they see their relationship as characterized by sequences of the type a_1 b_1 a_2 where a_1 is seen (by A and B) as a signal defining conditions of instrumental reward or punishment; b_1 as a signal or act obeying these conditions; and a_2 as a signal reinforcing b_1.

Similarly we shall say that "A is dependent on B" if their relationship is characterized by sequences a_1 b_1 a_2, where a_1 is seen as a signal of weakness; b_1 as a helping act; and a_2 as an acknowledgement of b_1

But it is up to A and B to distinguish (consciously or unconsciously or not at all) between "dominance" and "dependence." A “command” can closely resemble a cry for "help."

(c) In psychotherapy, Learning II is exemplified most conspicuously by the phenomena of "transference." Orthodox Freudian theory asserts that the patient will inevitably bring to the therapy room inappropriate notions about his relationship to the therapist. These notions (conscious or unconscious) will be such that he will act and talk in a way which would press the therapist to respond in ways which would resemble the patient's picture of how some important other person (usually a parent) treated the patient in the near or distant past. In the language of the present paper, the patient will try to shape his interchange with the therapist according to the premises of his (the patient’s) former Learning II.

It is commonly observed that much of the Learning II which determines a patient's transference patterns and, in-deed, determines much of the relational life of all human beings, (a) *dates from early infancy,* and *(b) is unconscious.* Both of these generalizations seem to be correct and both need some explanation.

It seems probable that these two generalizations are true because of the very nature of the phenomena which we are discussing. We suggest that *that* is learned in Learning II is a way of *punctuating events.* But a *way of punctuating* is not true or false. There is nothing contained in the propositions of this learning that can be tested against reality. It is like a picture seen in an inkblot; it has neither correctness nor incorrectness. It is only a *way* of seeing the inkblot.

Consider the instrumental view of life. An organism with this view of life in a new situation will engage in trial-and-error behavior in order to make the situation provide a positive reinforcement. If he fails to get this reinforcement, his purposive philosophy is not thereby negated. His trial-and- error behavior will simply continue. The premises of "purpose" are simply not of the same logical type as the material facts of life, and therefore cannot easily be contradicted by them.

The practitioner of magic does not unlearn his magical view of events when the magic does not work. In fact, the propositions which govern punctuation have the general characteristic of being self-validating.[9] What we term "con-text" includes the subject's behavior as well as the external events. But this behavior is controlled by former Learning II and therefore it will be of such a kind as to mold the total context to fit the expected punctuation. In sum, this self-validating characteristic of the content of Learning II has the effect that such learning is almost ineradicable. It follows that Learning II acquired in infancy is likely to persist through life. Conversely, we must expect many of the important characteristics of an adult’s punctuation to have their roots in early infancy.

In regard to the unconsciousness of these habits of punctuation, we observe that the "unconscious" includes not only repressed material but also most of the processes and *habits* of gestalt perception. Subjectively we are aware of our "dependency" but unable to say clearly how this pattern was constructed nor what cues were used in our creation of it.

Learning III

What has been said above about the self-validating character of premises acquired by Learning II indicates that Learning III is likely to be difficult and rare even in human beings. Expectably, it will also be difficult for scientists, who are only human, to imagine or describe this process. But it is claimed that something of the sort does from time to time occur in psychotherapy, religious conversion, and in other sequences in which there is profound reorganization of character.

Zen Buddhists, Occidental mystics, and some psychiatrists assert that these matters are totally beyond the reach of language. But, in spite of this warning, let me begin to speculate about what must (logically) be the case.

First a distinction must be drawn: it was noted above that the experiments in reversal learning demonstrate Learning II whenever there is measurable learning *about* the fact of reversal. It is possible to learn (Learning I) a given premise at a given time and to learn the converse premise at a later time without acquiring the knack of reversal learning. In such a case, there will be no improvement from one reversal to the next. One item of Learning I has simply replaced another item of Learning I without any achievement of Learning II. If, on the other hand, improvement occurs with successive reversals, this is evidence for Learning II.

If we apply the same sort of logic to the relation between Learning II and Learning III, we are led to expect that there might be replacement of premises at the level of Learning II *without* the achievement of any Learning III.

Preliminary to any discussion of Learning III, it is there-fore necessary to discriminate between mere replacement without Learning III and that facilitation of replacement which would be truly Learning Hl.

That psychotherapists should be able to aid their patients even in a mere replacement of premises acquired by Learning II is already no mean feat when we consider the self-validating character of such premises and their more or less unconscious nature. But that this much can be done there is no doubt.

Within the controlled and protected setting of the therapeutic relationship, the therapist may attempt one or more of the following maneuvers:

(a) to achieve a confrontation between the premises of the patient and those of the therapist—who is carefully trained not to fall into the trap of validating the old premises;

(b) to get the patient to act, either in the therapy room or outside, in ways which will confront his own premises;

(c) to demonstrate contradiction among the premises which currently control the patient's behavior;

(d) to induce in the patient some *exaggeration or caricature* (e.g., in dream or hypnosis) of experience based on his old premises.

As William Blake noted, long ago, "Without Contraries is no progression." (Elsewhere I have called these contradictions at level II "double binds.")

But there are always loopholes by which the impact of contradiction can be reduced. It is a commonplace of learning psychology that while the subject will learn (Learning I) more rapidly if he is reinforced every time he responds correctly, such learning will disappear rather rapidly if reinforcement ceases. If, on the other hand, reinforcement is only occasional, the subject will learn more slowly but the resulting learning will not easily be extinguished when reinforcement ceases altogether. In other words, the subject may learn (Learning II) that the context is such that absence of reinforcement does not indicate that his response was wrong or inappropriate. His view of the context was, in fact, correct until the experimenter changed his tactics.

The therapist must certainly so support or hedge the contraries by which the patient is driven that loopholes of this and other kinds are blocked. The Zen candidate who has been assigned a paradox *(koan)* must labor at his task "like a mosquito biting on an iron bar."

I have argued elsewhere ("Style, Grace, and Information in Primitive Art," see p. 128) that an essential and necessary function of all habit formation and Learning II is an *economy* of the thought processes (or neural pathways) which are used for problem-solving or Learning I. The premises of what is commonly called "character"—the definitions of the "self" —save the individual from having to examine the abstract, philosophical, aesthetic, and ethical aspects of many sequences of life. "I don*t know whether it's good music; I only know whether I like it."

But Learning III will throw these unexamined premises open to question and change.

Let us, as was done above for Learning I and II, list some of the changes which we shall be willing to call Learning III.

(a) The individual might learn to form more readily those habits the forming of which we call Learning II.

(b) He might learn to close for himself the "loopholes" which would allow him to avoid Learning III.

(c) He might learn to change the habits acquired by Learning II.

(d) He might learn that he is a creature which can and does unconsciously achieve Learning II.

(e) He might learn to limit or direct his Learning II.

(f) If Learning II is a learning of the contexts of Learning I, then Learning III should be a learning of the contexts of

those contexts.

But the above list proposes a paradox. Learning III *(i.e.,* learning *about* Learning II) may lead either to an increase in Learning II or to a limitation and perhaps a reduction of that phenomenon. Certainly it must lead to a greater flexibility in the premises acquired by the process of Learning II —a *freedom* from their bondage.

I once heard a Zen master state categorically: "To become accustomed to anything is a terrible thing."

But any freedom from the bondage of habit must also denote a profound redefinition of the self. If I stop at the level of Learning II, "I" am the aggregate of those characteristics which I call my "character." "I" am my habits of acting in context and shaping and perceiving the contexts in which I act. Selfhood is a product or aggregate of Learning II. To the degree that a man achieves Learning III, and learns to perceive and act in terms of the contexts of contexts, his "self" will take on a sort of irrelevance. The concept of "self" will no longer function as a nodal argument in the punctuation of experience.

This matter needs to be examined. In the discussion of Learning II, it was asserted that all words like "dependency," "pride," "fatalism," refer to characteristics of the self which are learned (Learning II) in sequences of relationship. These words are, in fact, terms for "roles" in relationships and refer to something artificially chopped out of interactive sequences. It was also suggested that the correct way to assign rigorous meaning to any such words is to spell out the formal structure of the sequence in which the named characteristic might have been learned. Thus the interactive sequence of Pavlovian learning was proposed as a paradigm for a certain sort of "fatalism," etc.

But now we are asking about the contexts of these contexts of learning, *i.e.*, about the larger sequences within which such paradigms are embedded.

Consider the small item of Learning II which was mentioned above as providing a "loophole" for escape from Learning III. A certain characteristic of the self—call it "persistence"—is generated by experience in multiple sequences among which reinforcement is sporadic. We must now ask about the larger context of such sequences. How are such sequences generated?

The question is explosive. The simple stylized experimental sequence of interaction in the laboratory is generated by and partly determines a network of contingencies which goes out in a hundred directions leading out of the laboratory into the processes by which psychological research is designed, the interactions between psychologists, the economics of research money, etc, etc.

Or consider the same formal sequence in a more "natural" setting. An organism is searching for a needed or missing object. A pig is rooting for acorns, a gambler is feeding a slot machine hoping for a jackpot, or a man must find the key to his car. There are thousands of situations where living things must persist in certain sorts of behavior precisely *because* reinforcement is sporadic or improbable. Learning II will simplify the universe by handling these instances as a single category. But if Learning III be concerned with the contexts of these instances, then the categories of Learning II will be burst open.

Or consider what the word "reinforcement" means at the various levels. A porpoise gets a fish from the trainer when he does what the trainer wants. At level I, the fact of the fish is linked with the "rightness" of the particular action. At level

II, the fact of the fish confirms the porpoise's under-standing of his (possibly instrumental or dependent) relationship with the trainer. And note that at this level, if the porpoise hates or fears the trainer, pain received from the latter may be a positive reinforcement confirming that hate. ("If it*s not the way I want it, Til prove it.")

But what of "reinforcement" at level III (for porpoise or for man) ?

If, as I have suggested above, the creature is driven to level III by "contraries" generated at level II, then we may expect that it is the resolving of these contraries that will constitute positive reinforcement at level III. Such resolution can take many forms.

Even the attempt at level III can be dangerous, and some fall by the wayside. These are often labeled by psychiatry as psychotic, and many of them find themselves inhibited from using the first person pronoun.

For others, more successful, the resolution of the contraries may be a collapsing of much that was learned at level II, revealing a simplicity in which hunger leads directly to eat-ing, and the identified self is no longer in charge of organizing the behavior. These are the incorruptible innocents of the world.

For others, more creative, the resolution of contraries reveals a world in which personal identity merges into all the processes of relationship in some vast ecology or aesthetics of cosmic interaction. That any of these can survive seems almost miraculous, but some are perhaps saved from being swept away on oceanic feeling by their ability to focus in on the minutiae of life. Every detail of the universe is seen as proposing a view of the whole. These are the people for whom Blake wrote the famous advice in the "Auguries of Innocence:"

To see the World in a Grain of Sand,

And a Heaven in a Wild Flower,

Hold Infinity in the palm of your hand,

And Eternity in an hour.

The Role of Genetics in Psychology

Whatever can be said about an animals learning or in-ability to learn has bearing upon the genetic make-up of the animal. And what has been said here about the levels of learning has bearing upon the whole interplay between genetic make-up and the changes which that individual can and must achieve.

For any given organism, there is an upper limit beyond which all is determined by genetics. Planarians can probably not go beyond Learning I. Mammals other than man are probably capable of Learning II but incapable of Learning III. Man may sometimes achieve Learning III.

This upper limit for any organism is (logically and presumably) set by genetic phenomena, not perhaps by individual genes or combinations of genes, but by whatever factors control the development of basic phylar characteristics.

For every change of which an organism is capable, there is the *fact* of that capability. This fact may be genetically determined; or the capability may have been learned. If the latter, then genetics may have determined the capability of learning the capability. And so on.

This is in general true of all somatic changes as well as of those behavioral changes which we call learning. A man's skin tans in the sun. But where does genetics enter this picture? Does genetics completely determine his *ability* to tan? Or can some men increase their ability to tan? In the latter case, the genetic factors evidently have effect at a higher logical level.

The problem in regard to any behavior is clearly not "Is it learned or is it innate?" but "Up to what logical level is learning effective and down to what level does genetics play a determinative or partly effective role?"

The broad history of the evolution of learning seems to have been a slow pushing back of genetic determinism to levels of higher logical type.

A Note on Hierarchies

The model discussed in this paper assumes, tacitly, that the logical types can be ordered in the form of a simple, unbranching ladder. I believe that it was wise to deal first with the problems raised by such a simple model.

But the world of action, experience, organization, and learning cannot be completely mapped onto a model which excludes propositions about the relation *between* classes of different logical type.

If C_x is a class of propositions, and C_2 is a class of propositions about the members of Cj; C_3 then being a class of propositions about the members of C_2; how then shall we classify propositions about the relation *between* these classes? For example, the proposition "As members of C_x are to members of $C_{2)}$ so members of

C_2 are to members of C_3" cannot be classified within the unbranching ladder of types.

The whole of this essay is built upon the premise that the relation between C_2 and C_3 can be compared with the relation between C_1 and C_2. I have again and again taken a stance to the side of my ladder of logical types to discuss the structure of this ladder. The essay is therefore itself an example of the fact that the ladder is not unbranching.

It follows that a next task will be to look for examples of learning which cannot be classified in terms of my hierarchy of learning but which fall to the side of this hierarchy as learning about the relation between steps of the hierarchy. I have suggested elsewhere ("Style, Grace, and Information in Primitive Art") that art is commonly concerned with learning of this sort, *i.e.*, with bridging the gap between the more or less unconscious premises acquired by Learning II and the more episodic content of consciousness and immediate action.

It should also be noted that the structure of this essay is *inductive* in the sense that the hierarchy of orders of learning is presented to the reader from the bottom upward, from level zero to level III. But it is not intended that the explanations of the phenomenal world which the model affords shall be unidirectional. In explaining the model to the reader, a unidirectional approach was necessary, but within the model it is assumed that higher levels are explanatory of lower levels and vice versa. It is also assumed that a similar reflexive relation—both inductive and deductive—obtains among ideas and items of learning as these exist in the lives of the creatures which we study.

Finally, the model remains ambiguous in the sense that while it is asserted that there are explanatory or determinative relations between ideas of adjacent levels both upward and downward, it is not clear whether direct explanatory relations exist between separated levels, e.g., between level III and level I or between level zero and level II.

This question and that of the status of propositions and ideas collateral to the hierarchy of types remains unexamined.

Notes:

* This essay was written in 1964 while the author was employed by the Communications Research Institute, under a Career Development Award (K3-NH-21, 931) from the National Institute of Mental Health. It was submitted as a position paper to the "Conference on World Views" sponsored by the Wenner-Gren Foundation, August 2-11, 1968. The section on "Learning III" was added in 1971.

[1] A. N. Whitehead and B. Russell, *Principia Mathematica,* 3 vols., 2nd ed., Cambridge, Cambridge University Press, 1910-13.

[2] It is conceivable that the same *words* might be used in describing both a class and its members and be true in both cases. The word "wave" is the name of a class of movements of particles. We can also say that the wave itself "moves,” but we shall be referring to a movement of a class of movements. Under friction, this metamovement will not lose velocity as would the movement of a particle.

[3] The Newtonian equations which describe the motions of a "particle" stop at the level of "acceleration." *Change of acceleration* can only happen with progressive deformation of the moving body, but the Newtonian "particle" was not made up of "parts" and was therefore (logically) incapable of deformation or any other internal change. It was therefore not subject to rate of change of acceleration.

[4] G. Bateson, "Social Planning and the Concept of Deutero-Learning," *Conference on Science, Philosophy and Religion, Second Symposium,* New York, Harper, 1942.

[5] *H.* E. Harlow, "The Formation of Learning Sets," *Psychology. Review,* 1949, 56: 51-65.

[6] E. L. Hull, *et al., Mathematico-deductive Theory of Rote Learning,* New Haven, Yale University, Institute of Human Relations, 1940.

[7] H. S. Liddell, "Reflex Method and Experimental Neurosis," *Personality and Behavior Disorders*, New York, Ronald Press, 1944.

[8] G. Bateson, *et al*, "Toward a Theory of Schizophrenia," *Behavioral Science,* 1956, 1: 251-64.

[9] J. Ruesch and G. Bateson, *Communication: The Social Matrix of Psychiatry*, New York, Norton, 1951.