THE LANGUAGE OF THOUGHT

JERRY A. FODOR
Massachusetts Institute of Technology

THOMAS Y. CROWELL COMPANY · NEW YORK · ESTABLISHED 1834

Published simultaneously in Canada by Fitzhenry & Whiteside, Ltd. Toronto.

Library of Congress Cataloging in Publication Data

Fodor, Jerry A.
The language of thought.

(The Language and thought series)
Bibliography: p.
Includes index.
1. Cognition. 2. Languages—Psychology. I. Title.
BF311.F56 153.4 75-4843
ISBN 0-690-00802-3

Thomas Y Crowell Company
666 Fifth Avenue
New York, New York 10019

Typography by Jules Perlmutter

Manufactured in the United States of America

CONTENTS

1
FIRST APPROXIMATIONS

I'm the only President you've got.
LYNDON B. JOHNSON

The main argument of this book runs as follows:

1. The only psychological models of cognitive processes that seem even remotely plausible represent such processes as computational.
2. Computation presupposes a medium of computation: a representational system.
3. Remotely plausible theories are better than no theories at all.
4. We are thus provisionally committed to attributing a representational system to organisms. 'Provisionally committed' means: committed insofar as we attribute cognitive processes to organisms and insofar as we take seriously such theories of these processes as are currently available.
5. It is a reasonable research *goal* to try to characterize the representational system to which we thus find ourselves provisionally committed.
6. It is a reasonable research *strategy* to try to infer this characterization from the details of such psychological theories as seem likely to prove true.
7. This strategy may actually work: It is possible to exhibit specimen inferences along the lines of item 6 which, if not precisely apodictic, have at least an air of prima facie plausibility.

The epistemic status of these points is pretty various. I take it, for example, that item 3 is a self-evident truth and therefore requires no justification beyond an appeal to right reason. I take it that item 4 follows from items 1–3. Items 5–7, on the other hand, need to be justified *in practice*. What must be shown is that it is, in fact, productive to conduct psychological research along the lines they recommend. Much of the material in later chapters of this book will be concerned to show precisely that. Hence, the discussion will become more intimately involved with empirical findings, and with their interpretations, as we go along.

This chapter, however is primarily concerned with items 1 and 2. I shall argue that, quite independent of one's assumptions about the *details* of psychological theories of cognition, their general structure presupposes underlying computational processes and a representational system in which such processes are carried out. It is often quite familiar facts which, in the first instance, constrain one's models of the mental life, and this chapter is mostly a meditation on a number of these. I shall, in short, discuss some kinds of theories which, I think, most cognitive psychologists would accept in outline, however much they might disagree about specifics. I want to show how, in every case, these theories presuppose the existence and exploitation of a representational system of some complexity in which mental processes are carried out. I commence with theories of choice.

I take it to be self-evident that organisms often believe the behavior they produce to be behavior of a certain kind and that it is often part of the explanation of the way that an organism behaves to advert to the beliefs it has about the kind of behavior it produces.[1] This being assumed, the following model seems overwhelmingly plausible as an account of how at least some behavior is decided on.

8. The agent finds himself in a certain situation (S).

9. The agent believes that a certain set of behavioral options (B_1, B_2, B_n) are available to him in S; i.e., given S, B_1 through B_n are the things the agent believes that he can do.

10. The probable consequence of performing each of B_1 through B_n are predicted; i.e., the agent computes a set of hypotheticals of roughly the form if B_i is performed in S, then, with a certain probability, C_i. Which such hypotheticals are computed and which probabilities are assigned will, of course, depend on what the organism knows or believes about situations like S. (It will also depend upon other variables which are, from the point of view of the present model, merely noisy: time pressure, the amount of computation space available to the organism, etc.)

11. A preference ordering is assigned to the consequences.

[1] I am not supposing that this is, in any technical sense, a *necessary* truth. But I do think it is the kind of proposition that it would be silly to try to confirm (or confute) by doing experiments. One can (just barely) imagine a situation in which it would be reasonable to abandon the practice of appealing to an organism's beliefs in attempts to account for its behavior: either because such appeals had been shown to be internally incoherent or because an alternative theoretical apparatus had been shown to provide better explanations. As things stand, however, no such incoherence has been demonstrated (the operationalist literature to the contrary notwithstanding) and no one has the slightest idea what an alternative theoretical option would be like (the behaviorist literature to the contrary notwithstanding). It is a methodological principle I shall adhere to scrupulously in what follows that if one has no alternative but to assume that P, then one has no alternative but to assume that P.

12. The organism's choice of behavior is determined as a function of the preferences and the probabilities assigned.

Two caveats. First, this is not a theory but a theory schema. No predictions about what particular organisms will choose to do on particular occasions are forthcoming until one supplies values for the variables; e.g., until one knows how *S* is described, which behavioral options are considered, what consequences the exploitation of the options are believed to lead to, what preference ordering the organism assigns to these consequences and what trade-off between probability and preferability the organism accepts. This is to say that, here as elsewhere, a serious theory of the way an organism behaves presupposes extensive information about what the organism knows and values. Items 8–12 do not purport to give such a theory, but only to identify some of the variables in terms of which one would have to be articulated.

Second, it is obvious that the model is highly idealized. We do not always contemplate each (or, indeed, any) of the behavioral options we believe to be available to us in a given situation. Nor do we always assess our options in the light of what we take to be their likely consequences. (Existentialists, I'm told, make a point of never doing so.) But these kinds of departures from the facts do not impugn the model. The most they show is that the behaviors we produce aren't always in rational correspondence with the beliefs we hold. It is sufficient for my point, however, that some agents are rational to some extent some of the time, and that when they are, and to the extent that they are, processes like the ones mentioned by items 8–12 mediate the relation between what the agent believes and what he does.[2]

Insofar as we accept that this model applies in a given case, we also accept the kinds of explanations that it licenses. For example, given the model, we may explain the fact that organism *a* produced behavior *B* by showing:

13. That *a* believed himself to be in situation *S*.

[2] It is not, of course, a sufficient condition for the rationality of behavior that processes like items 8–12 should be implicated in its production. For example, behaviors so mediated will generally be *ir*rational if the beliefs involved in item 10 are superstitious, or if the preferences involved in item 11 are perverse, or if the computations involved in items 9–12 are grossly unsound. Nor, so far as I can see, do items 8–12 propose *logically* necessary conditions upon the rationality of behavior. To revert to the idiom of the introduction, the conceptual story about what makes behavior rational presumably requires a certain kind of correspondence between behavior and belief but doesn't care about the character of the processes whereby that correspondence is effected; it is, I suppose, logically possible that angels are rational by reflex. The claim for items 8–12, then, is just that they—or something reasonably like them—are *empirically* necessary for bringing about a rational correspondence between the beliefs and the behaviors of sublunary creatures. The short way of saying this is that items 8–12 propose a (schematic) psychological theory.

14. That a believed that producing behavior of the type B_i in S would probably lead to consequence C_i.
15. That C_i was a (or the) highly valued consequence for a.
16. That a believed and intended B to be behavior of the B_i type.

The point to notice is that it is built into this pattern of explanation that agents sometimes take their behavior to be behavior of a certain kind; in the present case, it is part of the explanation of a's behavior that he believed it to be of the B_i kind, since it is behavior of that kind for which highly valued consequences are predicted. To put it briefly, the explanation fails to *be* a (full) explanation of a's behavior unless that behavior was B_i and a believed it to be so.

Items 13–16 might, of course, *contribute* to an explanation of behavior even where B is *not* produced and where the actual behavior is *not* taken by the agent to be B_i behavior. 'Will nobody pat my hiccup?' cried the eponymous Reverend Spooner. We assume that what goes in for B_i is a structural description of the sentence type 'Will nobody pick my hat up?' and that the disparity between the behavior produced and a token of that type is attributable to what the networks call a temporary mechanical failure. In such cases, our confidence that we know what behavior the agent intended often rests upon three beliefs:

17. That items 14 and 15 are true under the proposed substitution for B_i.
18. That items 14 and 15 would be false if we were instead to substitute a description of the type of which the observed behavior was in fact a token. (In the present example, it is plausibly assumed that Spooner would have set no positive utility upon the production of a token of the type 'Will nobody pat my hiccup?'; why on earth should he want to say *that*?)
19. That it is plausible to hypothesize mechanisms of the sort whose operations would account for the respects in which the observed and the intended behaviors differ. (In the present case, mechanisms of metathesis.)

It is notorious that if 'psychodynamic' explanations of behavior are true, the mechanisms envisaged by item 19 may themselves be of practically fathomless complexity. My present point, in any event, is that not only accounts of observed behavior, but also attributions of thwarted behavioral intentions, may intimately presuppose the applicability of some such explanatory schema as items 8–12.

I am laboring these very obvious remarks because I think that their immediate consequences are of profound significance for the construction of cognitive theories in general: viz., that this sort of explanation can go through only if we assume that agents have means for representing their behaviors to themselves; indeed, means for representing their behaviors as having certain properties and not having others. In the present case, it is essential to the explanation that the agent intends and believes the behavior

he produced to be behavior of a certain kind (viz., of the kind associated with relatively highly valued consequences in *S*) and not of some other kind (viz., not of the kind associated with relatively low-valued consequences in *S*). Give this up, and one gives up the possibility of explaining the behavior of the agent by reference to his beliefs and preferences.

The moral I want to draw, then, is that certain kinds of very central patterns of psychological explanation presuppose the availability, to the behaving organism, of some sort of representational system. I have emphasized, for purposes of exposition, the significance of the organism's representation of its own behavior in the explanation of its considered actions. But, once made, the point is seen to be ubiquitous. It was, for example, implicit in the model that the organism has available means for representing not only its behavioral options but also: the probable consequence of acting on those options, a preference ordering defined over those consequences and, of course, the original situation in which it finds itself. To use this sort of model is, then, to presuppose that the agent has access to a representational system of very considerable richness. For, according to the model, deciding is a computational process; the act the agent performs is the consequence of computations defined over representations of possible actions. No representations, no computations. No computations, no model.

I might as well have said that the model presupposes a language. For, a little prodding will show that the representational system assumed by items 8–12 must share a number of the characteristic features of real languages. This is a point to which I shall return at considerable length in Chapters 2 and 3. Suffice it to point out here just two of the properties that the putative system of representations must have in common with languages properly so-called (e.g., with natural languages).

In the first place, an infinity of distinct representations must belong to the system. The argument here is precisely analogous to the argument for the nonfiniteness of natural languages: Just as, in the latter case, there is no upper bound to the complexity of a sentence that can be used to make a statement, so in the former case, there is no upper bound to the complexity of the representation that may be required to specify the behavioral options available to the agent, or the situation in which he finds himself, or the consequences of acting one way or another.

This is not, of course, to argue that the *practical* possibilities are *literally* infinite. Just as there is a longest-sentence-that-anyone-can-utter, so there must be a most-complex-situation-that-anyone-can-act-upon. The infinite capacity of the representational system is thus an idealization, but it is not an *arbitrary* idealization. In both cases, the essential point is the organism's ability to deal with *novel* stimulations. Thus, we infer the productivity of natural languages from the speaker/hearer's ability to produce/understand sentences on which he was not specifically trained. Precisely the same argument infers the productivity of the internal representational

system from the agent's ability to calculate the behavioral options appropriate to a kind of situation he has never before encountered.

But productivity isn't the only important property common to natural languages and whatever system of representation is exploited in deciding what to do. It is evident, for example, that the notion that the agent can represent to himself salient aspects of the situations in which he finds himself presupposes that such familiar semantic properties as truth and reference are exhibited by formulae in the representational system.[3] We have been supposing that, underlying the capacity for reasoned action, there must be a capacity for the description of real and possible states of affairs. But the notions of description, truth, and reference are inseparable: Roughly, '*D*' describes what '*a*' refers to iff ('*Da*' is true iff *a* is *D*).

A similar line of thought shows that mechanisms for expressing intensional properties will have to be available to the representational system. In particular, calculated action presupposes decisions between possible (but) nonactual outcomes. So, the representational system recruited for the calculations must distinguish between possible, nonactual states of affairs. Whether one ought to do this by defining preference orderings over propositions (as traditional treatments of intensionality would suggest) or over possible worlds (in the manner of model-theoretic approaches to semantics) is a question I won't even attempt to deal with. My present point is just that *some* such mechanism must be available to the representational system, and for reasons quite parallel to those that lead us to think that some such mechanisms are available to natural languages.

I have assumed so far in this discussion that anyone reasonable will accept that something like items 8–12 is essential to a theory of the psychology of choice; what I have been doing is just spinning out some of the implications of that assumption. But, notoriously, the assumption isn't true. Behaviorists, for example, don't accept that deciding is a computational process, so behavioristic accounts of action can make do without postulating a system of internal representations. I don't propose to raise the general question of the adequacy of such accounts; it seems to me a dead issue. Suffice it to remark that, in light of our discussion, some of the standard criticisms can be deepened.

It is a point often made against behaviorists that they seek a prima facie implausible reduction of calculated actions to habits. The intended criticism is usually that insofar as actions are viewed simply as trained responses to environmental inputs the productivity of behavior is rendered unintelligible.

[3] I use the term 'formulae' without prejudice for whatever the vehicles of internal representation may turn out to be. At this point in the discussion it is left open that they might be images, or semaphore signals, or sentences of Japanese. Much of the discussion in succeeding chapters will concern what is known about the character of internal representations and what can be inferred about it from what is known of other things.

(For elaboration, see Chomsky, 1959.) But this is not the only thing wrong with construing calculated behaviors as species of conditioned responses. What everyone knows, but the behaviorist's methodology won't allow him to admit, is that at least some actions are choices from among a range of options contemplated by the agent. The behaviorist cannot admit this because he is committed to describing actions as the effects of environmental causes. Since only *actual* states of affairs can be causes, the-possibility-that-*P* cannot be among the determinants of a response. But nor, however, can *contemplations* of possibilities since, though they are presumably real events on any rational ontology, they are not *environmental* events in the behaviorist's proprietary sense of that notion. Looked at either way, the behaviorist is methodologically committed to denying what would seem to be self-evident: that we sometimes act the way we do because that seems the best way to act given what we take to be the options. In short, the behaviorist requires us to view considered behaviors as responses to actual inputs, when what we want to do is view them as responses to possible outcomes.

It is, conversely, one of the great advantages of computational theories of action that they allow us to acknowledge what everybody knows: that deciding what to do often involves considering what might turn out to be the case. To assume a representational system which can distinguish among (viz., assign different representations to) distinct possible states of affairs is precisely to permit oneself to view the behavior that is actually produced as a choice from among those options that the agent regards as 'live'. It is worth emphasizing that the behaviorist literature offers no grounds for rejecting this immensely plausible treatment except the reiterated assertion that it is, somehow, 'unscientific'. So far as I can tell, however, this amounts only to the (correct) observation that one cannot both say what it is plausible to say about actions and adhere to a behavioristic methodology. So much the worse for the methodology.

It will have occurred to the reader that what I am proposing to do is resurrect the traditional notion that there is a 'language of thought' and that characterizing that language is a good part of what a theory of the mind needs to do. This is a view to which, it seems to me, much of the current psychological work on cognition bears a curious and mildly schizoid relation. On the one hand, it seems to be implicit in almost every kind of explanation that cognitive psychologists accept since, as I remarked above, most such explanations treat behavior as the outcome of computation, and computation presupposes a medium in which to compute. But, on the other hand, the assumption of such a medium is relatively rarely made explicit, and the pressing question to which it leads—what properties does the system of internal representations have—is only occasionally taken as the object of sustained research.

I propose, as we go along, to consider a variety of types of evidence

that may bear upon the answer to that question. Before doing so, however, I want to explore two more lines of argument which seem to lead, with a fair show of inevitability, to the postulation of a language of thought as a precondition for any sort of serious theory construction in cognitive psychology. My point will be that not only considered action, but also learning and perception, must surely be viewed as based upon computational processes; and, once again, no computation without representation.

Let us first consider the phenomenon that psychologists sometimes call 'concept learning'. I want to concentrate on concept learning not only because it provides a useful illustration of our main thesis (cognitive processes are computational processes and hence presuppose a representational system) but also because the analysis of concept learning bears on a variety of issues that will arise in later chapters.

To begin with, then, concept learning is one of those processes in which what the organism knows is altered as a consequence of its experiences; in particular, as a consequence of its interactions with the environment. But, of course, not *every* case of an environmentally determined alteration in knowledge would count as learning; *a fortiori,* not all such cases count as *concept* learning. So, for example, aphasia is often environmentally induced, but catching aphasia isn't a learning experience. Similarly, if we could somehow induce knowledge of Latin by swallowing blue pills, I suppose that that would be acquiring Latin without learning it. Similarly, imprinting (see Thorpe, 1963) alters what the organism knows as a consequence of its experiences, but is only marginally a learning process if it is a learning process at all. A general theory of concept learning is, at best, *not* a general theory of how experience affects knowledge.

There are, moreover, kinds of *learning* that very probably aren't kinds of concept learning.[4] Rote learning is a plausible example (e.g., the learning of a list of nonsense syllables. However, see Young, 1968). So is what one might call 'sensory learning' (learning what a steak tastes like, learning what middle C sounds like played on an oboe, and so forth). Very roughly, and just by way of marking out the area of our concern, what distinguishes rote learning and sensory learning from concept learning is that, in the latter cases, what is *remembered of* an experience typically exhausts what is *learned from* that experience. Whereas concept learning somehow 'goes beyond' the experiential data. But what does *that* mean?

I think that what concept learning situations have in common is fundamentally this: The experiences which occasion the learning in such situations (under their theoretically relevant descriptions) stand in a *confirma-*

[4] I regard this as an empirical issue; whether it's true depends on what, in fact, goes on in the various learnings processes. It *might* turn out that the mechanism of concept learning is the general learning mechanism, but it would be a surprise if that were true and I want explicitly not to be committed to the assumption that it is. We badly need—and have not got—an empirically defensible taxonomy of kinds of learning.

tion relation to what is learned (under *its* theoretically relevant description). A short way of saying this is that concept learning is essentially a process of hypothesis formation and confirmation.[5] The best way to see that this is so is to consider the experimental paradigm in terms of which the concept learning 'construct' is, as one used to say, 'operationally defined'

In the typical experimental situation, the subject (human or infra-human) is faced with the task of determining the environmental conditions under which a designated response is appropriate, and learning is manifested by *S*'s increasing tendency, over time or trials, to produce the designated response when, and only when, those conditions obtain. The logic of the experimental paradigm requires, first, that there be an 'error signal' (e.g., reinforcement or punishment or both) which indicates whether the designated response has been appropriately performed and, second, that there be some 'criterial property' of the experimentally manipulated stimuli such that the character of the error signal is a function of the occurrence of the designated response together with the presence or absence of that property. Thus, in a simple experiment of this kind, *S* might be asked to sort stimulus cards into piles, where the figures on the cards exhibit any combination of the properties red and black with square and circular, but where the only correct (e.g., rewarded) sorting is the one which groups red circles with black squares. In such a case, the 'designated response' is sorting into the positive pile and the 'criterial property' is *red circle or black square*.

It is possible to use this sort of experimental setup to study the rate of learning as a function of any of a large number of variables: e.g., the character of the criterial property; *S*'s ability to report the property in terms of which he is sorting; the character of the error signal; the character of the relation (temporal, statistical, etc.) between occurrences of the error signal and instantiations of the criterial property; the character of the subject population (age, species, intelligence, motivation, or whatever); and so on. Much of the experimental psychology of learning in the last thirty years has been concerned with ringing changes on the values of these variables; the paradigm has been central to the work of psychologists who have as little else in common as, say, Skinner and Vygotsky.[6]

[5] This analysis of concept learning is in general agreement with such sources as Bruner, Goodnow, and Austin (1956), as is the emphasis upon the inferential character of the computations that underlie success in concept learning situations.

[6] Though Skinner would not, perhaps, like to see it put this way. Part of the radical behaviorist analysis of learning is the attempt to reduce concept learning to 'discrimination learning'; i.e., to insist that *what* the organism learns in the concept learning situation is *to produce the designated response*. It seems clear, however, that the reduction ought to go the other way around: The concept learning paradigm and the discrimination learning paradigm *are* the same, but in neither is the existence of a designated response more than a convenience to the experimenter; all it does is

My present point is that there is only one kind of theory that has ever been proposed for concept learning—indeed, there would seem to be only one kind of theory that is conceivable—and this theory is incoherent unless there is a language of thought. In this respect, the analysis of concept learning is like the analysis of considered choice; we cannot begin to make sense of the phenomena unless we are willing to view them as computational and we cannot begin to make sense of the view that they are computational unless we are willing to assume a representational system of considerable power in which the computations are carried out.

Notice, to begin with, that at any given trial t and in respect of any given property P, the organism's experience in the concept learning paradigm is appropriately represented as a data matrix in which the rows represent trials and the columns represent the performance of the designated response, the presence or absence of P, and the character of the error signal.[7] Thus:

---

provide a regimented procedure whereby S can indicate which sorting he believes to be the right one at a given stage in the learning process.

This is, I take it, not a methodological but an empirical claim. It is clear on several grounds that concept learning (in the sense of learning which categorization of the stimuli is the right one) can, and usually does, proceed in the absence of specific designated responses—indeed, in the absence of any response at all. Nature addicts learn, I'm told, to distinguish oaks from pine trees, and many of them probably do so without being explicitly taught what the distinguishing criteria are. This is true concept learning, but there is no distinctive response that even nature addicts tend to make when and only when they see an oak.

There is, in fact, plenty of experimental evidence on this point. Tolman (1932) showed that what a rat learns when it learns which turning is rewarded in a T-maze is *not* specific to the response system that it uses to make the turn. Brewer (to be published), in a recent survey of the literature on conditioning in human beings, argues persuasively that the designated response can usually be detached from the criterial stimuli simply by instructing the subject to detach it ('From now on, please do *not* sort the red circles with the black squares'). It is, in short, simply not the case that learning typically consists of establishing connections between specific classes of stimuli and specific classes of responses. What *is* the case is (a) that S can often use what he has learned to effect a correspondence between the occurrence of criterial stimulation and the production of a designated response; (b) that it is often experimentally convenient to require him to do so, thereby providing a simple way for E to determine which properties of the stimuli S believes to be criterial; and (c) that Ss will go along with this arrangement providing that they are adequately motivated to do so. Here as elsewhere, what the subject does is determined by his beliefs together with his preferences.

[7] One might, ideally, want a three-valued matrix since, on any given trial, the organism may not have observed, or may have observed and forgotten, whether the designated response was performed, whether P was present, or what the value of the error signal was. This is the sort of nicety which I shall quite generally ignore. I mention it only to emphasize that it is the organism's internal representation of its experiences (and not the objective facts about them) that is immediately implicated in the causation of its behavior.

TRIAL	DESIGNATED RESPONSE PERFORMED	PROPERTY *P* PRESENT	VALUE OF ERROR SIGNAL
1	yes	yes	minus
2	no	no	minus
3	yes	no	plus

Put this way, it seems clear that the problem the organism faces on trial t is that of choosing a value of P for which, in the ideal case, the last column of the matrix is positive when and only when the first two columns are, and which is such that the matrix will continue to exhibit that correspondence for any (reasonable) value of $t_n > t$. This is the sense in which what is learned in concept learning 'goes beyond' what is given in the experiential data. What the organism has to do in order to perform successfully is to extrapolate a generalization (all the positive stimuli are P-stimuli) on the basis of some instances that conform to the generalization (the first n positive stimuli were P-stimuli). The game is, in short, inductive extrapolation, and inductive extrapolation presupposes (a) a source of inductive hypotheses (in the present case, a range of candidate values of P) and (b) a confirmation metric such that the probability that the organism will accept (e.g., act upon) a given value of P at t is some reasonable function of the distribution of entries in the data matrix for trials prior to t.

There are, of course, many many ways of fleshing out the details of this kind of model. For example, there is plenty of reason to believe that the various values of P are typically tested in a determinate order; indeed, that the choice of P may be very subtly determined by the character of the P-values previously assessed and rejected and by the particular configuration of the data matrix for those values. But, however the details go, what seems entirely clear is that the behavior of the organism will depend upon the confirmation relation between the data and the hypothesis, so that accounts of its behavior will require information about how, in the course of learning, the data and the hypotheses are represented.

Why is this entirely clear? Fundamentally, because one of the distinguishing characteristics of concept learning is the *nonarbitrariness* of the relation between what is learned and the character of the experiences that occasion the learning. (Compare the case of acquiring Latin by taking pills.) That is, what a theory of concept learning has to explain is why it is experiences of xs which are F (and not, say, experiences of xs which are G) that leads the organism, eventually, to the belief that all the xs are F. We *can* explain this if we assume (a) that the organism *represents* the relevant experiences as experiences of xs which are F; (b) that one of the hypotheses that the organism entertains about its environment is the hypothesis that perhaps all xs are F; and (c) that the organism employs, in the fixation of its beliefs, a rule of confirmation which says (*very*

roughly) that all the observed *x*s being *F* is, *ceteris paribus*, grounds for believing that all the *x*s are *F*. To put it mildly, it seems unlikely that any theory radically incompatible with items (a–c) could account for the non-arbitrariness of the relation between what is learned and the experiences that occasion the learning.[8]

In short, concept learning begs for analysis as involving the determination of a confirmation relation between observed and extrapolated reward contingencies, and this is already to commit oneself to a representational system in which the observations and the candidate extrapolations are displayed and the degree of confirmation is computed. There is, however, also a more subtle way in which inductive extrapolation presupposes a representational system, and this point bears considering.

Inductive extrapolation is a form of nondemonstrative inference. For present purposes this means that, at any given trial *t*, there will be indefinitely many nonequivalent values of *P* that are 'compatible' with the data matrix up to *t*. That is, there will be indefinitely many values of *P* such that, on all trials prior to *t*, the designated response is rewarded iff *P* is exhibited by the stimulus, but where each value of *P* 'predicts' a different pairing of responses and rewards on future trials. Clearly, if the organism is to extrapolate from its experiences, it will need some way of choosing between these indefinitely many values of *P*. Equally clearly, that choice cannot be made on the basis of the data available up to *t* since the choice that needs to be made is precisely among hypotheses all of which predict the *same* data up to *t*.

This is a familiar situation in discussions of inductive inference in

[8] I have purposely been stressing the analogies between the theory of inductive confirmation and the theory of the fixation of belief. But I do *not* intend to endorse the view (which examples like item (c) might suggest) that the confirmation of universal hypotheses in science is normally a process of simple generalization from instances. For that matter, I do not intend to endorse the view, embodied in the program of 'inductive logic', that confirmation is normally reconstructable as a 'formal' relation between hypotheses and data. On the contrary, it appears that the level of confirmation of a scientific hypothesis is frequently sensitive to a variety of *in*formal considerations concerning the overall economy, plausibility, persuasiveness and productivity of the theory in which the hypothesis is embedded, to say nothing of the existence of competing theories.

It may well be that the fixation of belief is also sensitive to these sorts of 'global' considerations. Even so, however, the prospects for a formal theory of belief seem to me considerably better than the prospects for an inductive logic. To formalize the relation of inductive confirmation, we should have to provide a theory which picks the *best* hypothesis (the hypothesis that *ought* to be believed), given the available evidence. Whereas, to formalize the fixation of belief, we need only develop a theory which, given the evidence, picks the hypothesis that the organism *does* believe. To the extent that this *cannot* be done, we cannot view learning as a computational process; and it is, for better or for worse, the working assumption of this book that computational accounts of organisms will not break down.

the philosophy of science. The classic argument is due to Goodman (1965), who pointed out that, for any fixed set of observations of green emeralds, both the hypothesis that all emeralds are green and the hypothesis that all emeralds are *grue* will be compatible with the data. (One way of defining a *grue*-predicate is: An emerald is *grue* iff it is ((in the data sample and green) or (not in the data sample and blue)). It is part of Goodman's point, however, that there are indefinitely many ways of constructing predicates which share the counterinductive properties that *grue* exhibits.) Since both hypotheses are compatible with the data, the principle that distinguishes between them must appeal to something other than observations of green emeralds.

The way out of this puzzle is to assume that candidate extrapolations of the data receive an a priori ordering under a *simplicity metric*, and that that metric prefers 'all *x*s are green' to 'all *x*s are *grue*' as the extrapolation of any body of data compatible with both.[9] In the present case this means that the decision that a given value of *P* is confirmed relative to a given data matrix must be determined not only by the distribution of entries in the matrix, but also by the relative simplicity of *P*. This conclusion seems to be irresistible, given the nondemonstrative character of the extrapolations involved in concept learning. It has, however, immediate consequences for the general claim that theories of concept learning are incoherent unless they presuppose that a representational system is available to the organism.

The point is that, so far as anyone can tell, simplicity metrics must be sensitive to the *form* of the hypotheses that they apply to, i.e., to their syntax and vocabulary.[10] That is, so far as anyone can tell, we can get an a priori ordering of hypotheses only if we take account of the way in which the hypotheses are expressed. We need such an ordering if we are to provide a coherent account of the order in which values of *P* are selected in the concept learning situation. But this means that a theory of concept

[9] I take it that this is common ground among philosophers of science. Where they disagree is on how to characterize the difference between predicates like *grue* (which the simplicity metric doesn't like) and predicates like green (which it does); and also, on how to justify adopting a simplicity metric which discriminates that way.

[10] Notions like entrenchment, for example, are defined over the *predicates* of a science. If 'green' is more entrenched than 'grue', that is presumably because there are laws expressed in terms of the former but no laws expressed in terms of the latter. (For discussion, see Goodman, 1965.) One could, of course, try to avoid this conclusion by defining simplicity, entrenchment, and related notions for *properties* (rather than for predicates). But even if that *could* be done it would seem to be a step in the wrong direction: Insofar as one wants psychological processes to turn out to be *computational* processes, one wants the rules of computation to apply formally to the objects in their domains. Once again: my goal in this book is not to *demonstrate* that psychological processes are computational, but to work out the consequences of assuming that they are.

learning will have to be sensitive to the way that the organism represents its hypotheses. But the notion of the organism representing its hypotheses in one way or another (e.g., in one or another vocabulary or syntax) just *is* the notion of the organism possessing a representational system.

In fact, this argument states the case too weakly. In the formalization of scientific inference a simplicity metric distinguishes between hypotheses that are compatible with the data but make different predictions for *un*observed cases. Our point, thus far, has been that the corresponding remarks presumably hold in the special case where the hypotheses are *P*-values and the data are the observed values of the error signal. There is, however, a respect in which the case of scientific inference differs from the extrapolations involved in concept learning. A simplicity metric used in the evaluation of scientific theories is presumably *not* required to distinguish between *equivalent* hypotheses. To put it the other way around, two hypotheses are identical, for the purposes of formalizing scientific inferences, if they predict the same extrapolations of the data matrix and are equally complex. Pairs of hypotheses that are identical in this sense, but differ in formulation, are said to be 'notational variants' of the same theory.

There is ample evidence, however, that the a priori ordering of *P*-values exploited in concept learning *does* distinguish between hypotheses that are, in this sense, notational variants of each other; i.e., the ordering of *P*-values imposes *stronger* constraints upon the form of a hypothesis than simplicity metrics do.

It is, for example, a standard finding that *S*s prefer affirmative conjunctive representations of the data matrix to negative or disjunctive representations. (See Bruner et al. 1956.) Thus, subjects in the concept learning task will typically find it easier to learn to sort all the red triangles together than to learn to sort together all things that *aren't* triangles or all the things that are either triangles or red. Yet, affirmative conjunctive hypotheses are interdefinable with negative disjunctive hypotheses; the subject who is choosing all and only red triangles as instances of positive stimuli is ipso facto choosing all and only things that are (not triangles or not red) as instances of the negative stimuli.[11] What makes the difference in the subject's performance is which of these choices he takes himself to be making; i.e., the way he represents the choices. *S*s who report an affirmative conjunctive hypothesis typically learn faster than those who don't.[12] This is

[11] The point is, of course, that 'choosing' is opaque in the first occurrence and transparent in the second. Perhaps it's not surprising that what is chosen opaquely is chosen under a representation.

[12] For example, Wason and Johnson-Laird (1972) describe an experiment in which *S*s were, in effect, presented with data matrices and required to articulate the appropriate extrapolations. The basic prediction, which was confirmed, was that "concepts which were essentially conjunctive in form would be easier to formulate than con-

thoroughly intelligible on the assumption that the same hypothesis can receive different internal representations and that the subject's a priori preferences are sensitive to such differences. But it doesn't seem to be intelligible on any other account.

We have been considering some of the ways in which viewing the concept learning task as essentially involving inductive extrapolation commits one to postulating a representational system in which the relevant inductions are carried through. I think it is worth emphasizing that no alternative view of concept learning has ever been proposed, though there are alternative vocabularies for formulating the view just discussed. For example, many psychologists use the notion of habit strength (or strength of association) where I have used the notion of degree of confirmation of a hypothesis. But once it has been recognized that any such construct must be defined over candidate extrapolations of a data matrix (and not over S-R pairings; see footnote 6) the residual issue is entirely terminological. A theory which determines how habit strength varies as a function of reinforcement (or which determines strength of association as a function of frequency of association, etc.) just *is* an inductive logic, where the confirmation function is articulated by whatever laws of reinforcement/association are assumed.

Similarly, some psychologists would prefer to speak of a theory of attention where I have spoken of a theory which determines the order in which *P*-values are tested. But again the issue is just terminological. A theory which determines what the organism is attending to at t thereby predicts the stimulus parameter that is extrapolated at t. It must therefore be sensitive to whatever properties of the data matrix, and of the previously contemplated hypotheses, affect the order in which *P*-values are tested, and to whatever a priori ordering of *P*-values determines their relative complexity. Whether or not one *calls* this a theory of attention,

---

cepts which were essentially disjunctive in form, and that whenever a component was negated there would be a slight increase in difficulty" (p. 70). They note that the order of difficulty that they obtained by asking the subject to state the relevant generalization "conforms to the order obtained when subjects have to *learn* concepts in the conventional manner" (p. 72), i.e., in the concept learning task. The point to notice is that, since conjunction is interdefinable with negation and disjunction, no concept is, *strictly speaking*, essentially conjunctive or essentially disjunctive. Strictly speaking, concepts don't *have* forms, though representations of concepts do. What Wason and Johnson-Laird mean by a conjunctive concept is, as they are careful to point out, just one which can be expressed by a (relatively) economical formula *in the representational system that the subject is using* (in the present case, in English). What the experiment really shows, then, is that the employment of such a representation facilitates the subject's performance; hence that formulations of a hypothesis which are, in the sense described above, mere notational variants of one another, may nevertheless be differentially available as extrapolations of a data matrix.

the function of the construct is precisely to predict what extrapolations of the data matrix the organism will try and in what order it will try them.

Finally, there are psychologists who prefer to describe the organism as 'sampling' the properties of the stimulus rather than as constructing hypotheses about which such properties are criterial for sorting. But the notion of a property is proprietary in the former kind of theory. In the non-proprietary sense of 'property', every stimulus has an infinity of properties an infinite subset of which are never sampled. The properties that *are* sampled, on the other hand, are of necessity a selection from those that the organism is capable of internally representing. Given that, talking about sampling the properties of the stimulus and talking about projecting hypotheses about those properties are two ways of making the same point.

To summarize: So far as anyone knows, concept learning is essentially inductive extrapolation, so a theory of concept learning will have to exhibit the characteristic features of theories of induction. In particular, concept learning presupposes a format for representing the experiential data, a source of hypotheses for predicting future data, and a metric which determines the level of confirmation that a given body of data bestows upon a given hypothesis. No one, so far as I know, has ever doubted this, though I suppose many psychologists have failed to realize what it was that they weren't doubting. But to accept that learning which 'goes beyond the data' involves inductive inference is to commit oneself to a language in which the inductions are carried out, since (a) an inductive argument is warranted only insofar as the observation statements which constitute its premises confirm the hypothesis which constitutes its conclusion; (b) whether this confirmation relation holds between premises and conclusion depends, at least in part, upon the *form* of the premises and conclusion; and (c) the notion of 'form' is defined only for 'linguistic' objects; viz. for representations.

I shall close this chapter by pointing out that the same kinds of morals emerge when one begins to think about the structure of theories of perception.

To begin with, there is an obvious analogy between theories of concept learning of the kind I have just been discussing and classical theories of perception in the empiricist vein. According to the latter, perception is essentially a matter of problem solving, where the form of the problem is to predict the character of future sensory experience given the character of past and current sensations as data. Conceived this way, models of perception have the same general structure as models of concept learning: One needs a canonical form for the representation of the data, one needs a source of hypotheses for the extrapolation of the data, and one needs a confirmation metric to select among the hypotheses.

Since some of the empiricists took their project to be the formalization of perceptual *arguments*—viz., of those arguments whose cogency justifies our knowledge claims about objects of perception—they developed fairly explicit doctrines about the kinds of representations that mediate perceptual

inferences. It is possible (and it is in the spirit of much of the empiricist tradition) to regard such doctrines as implying theories of the computational processes that underlie perceptual integration. It is notorious, however, that in a number of respects empiricist accounts of perceptual inferences make dubious psychology when so construed. For example, the premises of perceptual inferences were sometimes presumed to be represented in a 'sense datum' language whose formulae were supposed to have some extremely peculiar properties: E.g. that sense datum statements are somehow incorrigible, that all empirical statements have a unique decomposition into sense datum statements; that each sense datum statement is logically independent of any of the rest, and so on.

For many of the empiricists, the defining feature of this data language was supposed to be that its referring expressions could refer only to qualia; If sense datum statements were curious, that was because qualia were curiouser. Conversely, the language in which perceptual hypotheses are couched was identified with 'physical object language', thereby making the distinction between what is sensed and what is perceived coextensive with the distinction between qualia and things. Redescriptions of sensory fields in physical object terms could mediate the prediction of future sensations because, on this view, to accept a description of one's experiences in a physical object language is logically to commit oneself to (at least hypothetical) statements about experiences yet to come. Roughly, sense datum statements provide inductive support for physical object statements, and physical object statements entail statements about further sensations. One thus accepts an 'inductive risk' in inferring from sensations to perceptions, and the problem posed to the perceiver is that of behaving rationally in face of this risk. That is, given a description of experience couched in the sensation language, he must somehow choose that *re*-description in physical object terms which the experiences best confirm. Only by doing so can he be rationally assured that most of the expectations about future or hypothetical experiences to which his perceptual judgements commit him are likely to be true.

If, in short, I describe my current experience in terms of color patches, textures, smells, sounds, and so forth, I do not commit myself to predictions about the character of my prior or future experiences. But if I describe it in terms of tables and chairs and their logical kin then I *am* so committed since nothing can be a table or chair unless it performs in a reasonably table-or-chair-wise fashion across time. So, if I claim that what I see is a table, I am (implicitly) going bond for its past and future behavior; in particular, I am issuing guarantees about the sensations it will, or would, provide. So the story goes.

It is widely known that this account of perception has taken a terrific drubbing at the hands of epistemologists and Gestalt psychologists. It is hard, these days, to imagine what it would be like for the formulae of a

representational system to be privileged in the way that formulae in the sense datum language were supposed to be. Nor is it easy to imagine a way of characterizing qualia which would make it turn out that one's perceptual information is all mediated by the sensing of them. Nor does it seem pointful to deny that what one sees are typically *things*; not, in any event, if the alternative is that what one sees are typically color patches and their edges.

This line of criticism is too well known to bear repeating here. I think that it is clearly cogent. But I think, nevertheless, that the core of the empiricist theory of perception is inevitable. In particular, the following claims about the psychology of perception seem to me to be almost certainly true and entirely in the spirit of empiricist theorizing:

1. Perception typically involves hypothesis formation and confirmation.
2. The sensory data which confirm a given perceptual hypothesis are typically internally represented in a vocabulary that is impoverished compared to the vocabulary in which the hypotheses themselves are couched.

Before I say why I think these aspects of the empiricist treatment of perception are right, I want to say something brief about where I think the empiricists went wrong.

I am reading the typical empiricist theory of perception as doing double duty: as an account of the justification of perceptual beliefs and as a psychology of the integration of percepts. I think it is clear that many of the empiricists took their views this way. But it is also pretty clear that when a conflict arose between what the psychology required and what the epistemology appeared to, it was the demands of the latter that shaped the theory.

For example, the claim of incorrigibility for sense datum statements was not responsive to any particular psychological insight, but rather to the presumed need to isolate inductive risk at some epistemic level other than the one at which the data are specified. The idea was, roughly, that we could not know physical object statements to be true unless we were certain of the data for those statements, and we could not be *certain* of the data statements if it is possible that some of them are false. Certainty is, as it were, inherited upward from the data to the perceptual judgments they support. Similarly, experiences of qualia have to be conscious events because the statements which such experiences confirm are the premises for arguments whose conclusions are the physical object statements we explicitly believe. If such arguments are to be our justification for believing such statements, their premises had better be available for us to cite.

This is, very probably, mostly muddle. Justification is a far more pragmatic notion than the empiricist analysis suggests. In particular, there is no reason why the direction of all justificatory arguments should be upward from epistemologically unassailable premises. Why should not one

of my physical object statements be justified by appeal to another, and that by appeal to a third, and so on? What justificatory argument requires is not that some beliefs be unquestionable but at most that some of them be (de facto) unquestioned. What *can't* be done is to justify all my beliefs *at once*. Well, what can't be done can't be done.

But while I think that the notion of *the* direction of justification is largely confused, the notion that there is a direction of information flow *in perception* is almost certainly well taken, though the arguments are empirical rather than conceptual.

To begin with, it seems clear that causal interactions between the organism and its environment must contribute to the etiology of anything one would want to call *perceptual* knowledge. Insofar as this is right, there is a good deal of empirical information available about the character of these interactions.

So far as anybody knows, any information that the organism gets about its environment as a result of such interactions must be mediated by the activity of one or another *sensory mechanism*. By a sensory mechanism, I mean one which responds to *physical properties* of environmental events. By a physical property I mean one designated by a natural kind term in some (ideally completed) physical science (for the notion of a natural kind term, see the second part of the introduction). What *mediated by* comes to will take some explaining, but as a first approximation I mean that the operation of a sensory mechanism in responding to a physical property of an environmental event is an empirically necessary condition for the organism's perception of *any* property of that environmental event.

Suppose, for example, that we think of a sensory mechanism as represented by a characteristic function, such that the value of the function is 1 in any case where the mechanism is excited and 0 otherwise. Then, so far as anyone knows, we can develop a theory which predicts the values of that function across time only if we take into account the physical properties of inputs to the mechanism. And we can predict the perceptual analysis that the organism will assign a given environmental event only if we know which physical properties of that event the sensory mechanisms of the organism have responded to. (Thus, for example, to predict the state of excitation of the human auditory system, we need information about the spectrum analysis of impinging wave forms. And to predict the sentence type to which an utterance token will be perceptually assigned, we must know at least which auditory properties of the utterance have been detected.)

I want to stress that this is an *empirical* fact even though it is not a *surprising* fact. We can imagine an organism (say an angel or a clairvoyant) whose perceptual knowledge is *not* mediated by the operation of sensory mechanisms; only, so far as we know, there are no such organisms, or, if there are any, psychologists have yet to find them. For all the

known cases, perception is dependent upon the operation of mechanisms whose states of excitation can be predicted from physical descriptions of their input and not in any other way.

Viewed in terms of information flow, this means that a sensory mechanism operates to associate token physical excitations (as input) with token physical descriptions (as output); i.e., a sensory mechanism is a device which says 'yes' when excited by stimuli exhibiting certain specified values of physical parameters and 'no' otherwise.[13] In particular, it does not care about any property that environmental events *fail* to share so long as the events have the relevant physical properties in common, and it does not care about nonphysical properties that environmental events have in common so long as they fail to share the relevant physical properties. In this sense, the excitation of a sensory mechanism encodes the presence of a physical property. (If the auditory system is a mechanism whose states of excitation are specific to the values of frequency, amplitude, etc., of causally impinging environmental events, then one might as well think of the output of the system as an encoded description of the environment in terms of those values. Indeed, one had better think of it this way if one intends to represent the integration of auditory percepts as a *computational* process.) But if this is true, and if it is also true that whatever perceptual information the organism has about its environment is mediated by the operation of its sensory mechanisms, it follows that perceptual analyses must somehow be responsive to the information about values of physical parameters of environmental events that the sensory mechanisms provide.[14]

[13] For purposes of exposition, I am ignoring the (serious) empirical possibility that some or all sensory mechanisms have output values between 0 and 1. Problems about the 'digitalness' of the various stages of cognitive processing are at issue here; but, though these problems are interesting and important, they don't affect the larger issues. Suffice is to say that the question is not just whether the outputs of sensory mechanisms are continuous under physical description, but rather whether intermediate values of excitation carry information that is used in later stages of processing. I don't know what the answer to this question is, and I don't mean to preclude the possibility that the answer is different for different sensory modalities.

[14] It bears emphasizing that the present account of sensory systems, like most of the psychological theorizing in this chapter, is highly idealized. Thus, "from the physical point of view the sensory receptors are transducers, that is, they convert the particular form of energy to which each is attuned into the electrical energy of the nerve impulse." (Loewenstein, 1960). But, of course, it does not follow that the sensors are *perfect* transducers, viz., that their output is predictable *just* from a determination of the impinging physical energies. On the contrary, there is evidence that any or all of the following variables may contribute to such determinations.

i. Cells in sensory systems exhibit a characteristic cycle of inhibition and heightened sensitivity consequent upon each firing. The effects of impinging stimuli are thus not independent of the effects of prior stimulations unless the interstimulus interval is large compared to the time course of this cycle.

That, I suppose, *is* the problem of perception insofar as the problem of perception is a problem in psychology. For though the information provided by causal interactions between the environment and the organism is information about physical properties in the *first* instance, in the *last* instance it may (of course) be information about any property the organism can perceive the environment to have. To a first approximation, the outputs of sensory mechanisms are appropriately viewed as physical descriptions, but perceptual judgments need not be articulated in the vocabulary of such descriptions. Typically they *are* not: A paradigm perceptual judgment is, 'There's a robin on the lawn' or 'I see by the clock that it's time for tea'.

It is, I take it, an empirical question whether psychological processes are computational processes. But if they are, then what must go on in perception is that a description of the environment that is *not* couched in a vocabulary whose terms designate values of physical variables is somehow computed on the basis of a description that *is* couched in such a vocabulary. Presumably this is possible because the perceptual analysis of an event is determined not just by sensory information but also by such background knowledge as the organism brings to the task. The computational processes in perception are mainly those involved in the integration of these two kinds of information. I take it that that is what is left of the classical empiricist view that perception involves the (nondemonstrative) inference from descriptions couched in a relatively impoverished language to conclusions couched in a relatively unimpoverished one.

Almost nothing is left of the empiricist epistemology. For example,

---

ii. Cells on the sensory periphery may be so interconnected that the excitation of any of them inhibits the firing of the others. Such mutual 'lateral' inhibition of sensory elements is usually interpreted as a 'sharpening' mechanism; perhaps part of an overall system of analog-to-digital conversion. (See Ratliff, 1961.)

iii. At any distance 'back' from the periphery of the sensory system one is likely to find 'logic' elements whose firing may be thought of as coding Boolean functions of the primary transducer information. (See Letvin et al., 1961, Capranica, 1965.)

iv. There may be central 'centripetal' tuning of the response characteristics of the peripheral transducers, in which case the output of such transducers may vary according to the motivational, attentional, etc. state of the organism.

v. Cells in the sensory system exhibit 'sponaneous' activity; viz., firing which is *not* contingent upon stimulus inputs.

A sensory transducer may thus diverge, in all these respects, from the ideal mechanisms contemplated in the text; nor do I wish to claim that this list is complete. But for all that, the main point holds: Insofar as the environment *does* contribute to the etiology of sensory information, it is presumably only under physical description that the uniformities in its contribution are revealed. Equivalently for these purposes: Insofar as the activity of sensory mechanisms encodes information about the state of the environment, it is the physical state of the environment that is thus encoded.

the perceptually pertinent description of sensory information is not given in the theory-free language of qualia but rather in the theory-laden language of values of physical parameters. (This is a way of saying what I said above: that, so far as anyone knows, the only way of providing a reasonably compact account of the characteristic function for a sensory mechanism is by taking its inputs under physical description.) Hence, there is no reason to believe that the organism cannot be mistaken about what sensory descriptions apply in any given case. For that matter, there is no reason to believe that organisms are usually conscious of the sensory analyses that they impose.

This distinction—between the notion of a sensory mechanism as the source of a mosaic of conscious experiences out of which percepts are constructed (e.g., by associative processes) and the notion of the sensors as transducers of such environmental information as affects perceptual integration—is now standard in the psychological literature. It is stressed even by such psychologists as Gibson (1966), whose approach to perception is not, on the whole, sympathetic to the sort of computational views of psychology with which I am primarily concerned. For Gibson, perception involves the detection of invariant (typically relational) properties of impinging stimulus arrays. He apparently assumes that any percept can be identified with such an invariant if only the relevant property is sufficiently abstractly described.[15] But, though Gibson denies that percepts are constructed from conscious sensory data, he does apparently hold that the presence of the relevant stimulus invariant must be inferred from the information output by sensory transducers.

> I will distinguish the input to the nervous system that evokes conscious sensation from the input that evokes perception. For

[15] The status of the claim that there are stimulus invariants corresponding to precepts is unclear. On one way of reading it it would seem to be a necessary truth: Since 'perceive' is a success verb, there must be at least one invariant feature of all situations in which someone perceives a thing to be of type *t*; viz., the presence of a thing of type *t*. On the other hand, it is a very strong *empirical* claim that, for any type of thing that can be perceived, there exists a set of *physical* properties such that the detection of those properties is plausibly identified with the perception of a thing of that type. This latter requires that the distinction between things of type *t* and everything else *is a physical distinction*, and, as we saw in the introduction, that conclusion does *not* follow just from the premise that *t*-type objects are physical objects.

The issue is whether there are physical kinds corresponding to perceptual kinds and that, as we have been saying all along, is an empirical issue. My impression of the literature is that the correspondence fails more often than it holds; that perception cannot, in general, be thought of as the categorization of *physical* invariants, however abstractly such invariants may be described. (For a discussion of the empirical situation in the field of speech perception, cf. Fodor et al., 1974.)

> it is surely a fact that *detecting* something can sometimes occur without the accompaniment of sense impressions. An example is the visual detection of one thing behind another. But this does not mean that perception can occur without stimulation of receptors; it only means that organs of perception are sometimes stimulated in such a way that they are not specified in consciousness. Perception cannot be without input; it can only be so if that means without awareness of the visual, auditory, or other quality of the input. An example of this is the 'obstacle sense' of the blind, which is felt as 'facial vision' but is actually auditory echo detection. The blind man 'senses' the wall in front of him without realizing what sense has been stimulated. In short there can be sensationless perception, but not informationless perception. (p. 2)

Thus, even for psychologists who think of perceptual distinctions as distinctions between (abstract) stimulus invariants, the problem of how such invariants are themselves detected needs to be solved; and it appears that solving it requires postulating the same sorts of inferences from inputs that empiricist theories assumed. The difference is mainly that contemporary psychologists do not assume that the computations, or the data over which they are defined, must be consciously accessible.[16]

It is worth emphasizing that the claim that the outputs of sensory mechanisms are, in general, not consciously accessible is supposed to be an empirical result rather than a truth of epistemology. There is, for example, quite good empirical evidence that an early representation of a speech sig-

[16] Gibson sometimes writes as though the problem of how the (presumed) stimulus invariants are detected could be avoided by distinguishing between the stimulus for the *sensory transducers* (viz., physical energies) and the stimulus for the *perceptual organs* (viz., abstract invariants). But this way trivialization lies. If one is allowed to use the notion of a stimulus so as to distinguish the input to the retina (light energy) from the input to the optic system (patterns of light energy which exhibit invariancies relevant, e.g., to the explanation of perceptual constancies), why not also talk about the stimulus for the *whole organism* (viz., perceptibles)? Thus, the answer to 'How do we perceive bottles?' would go: 'It is necessary and sufficient for the perception of a bottle that one detect the present of the stimulus invariant *bottle*'. The trouble with this answer (which, by the way, has a curiously Rylean sound to my ears) is, of course, that the problem of how one detects the relevant stimulus invariant is the *same* problem as how one perceives a bottle, so no ground has been gained overall.

What this shows, I think, is not that the psychological problem of perception is a muddle, but that *stating* the problem requires choosing (and motivating) a proprietary vocabulary for the representation of inputs. I have argued that the vocabulary of values of physical parameters *is* appropriate on the plausible assumption that sensory transducers detect values of physical parameters and that all perceptual knowledge is mediated by the activity of sensory transducers.

nal must specify its formant relations.[17] Yet speaker/hearers have no conscious access to formant structure and, for that matter, very little conscious access to any other acoustic property of speech. It is, in fact, very probably a general truth that, of the various redescriptions of the input that underlie perceptual analyses, the degree of conscious accessibility of a representation is pretty well predicted by the abstractness of its relation to what the sensors specify. This is the kind of point that such philosophers as Cassirer have had in mind when they remark that we 'hear through' an utterance of a sentence to its meaning; one is much better at reporting the syntactic type of which an utterance is a token than at reporting the acoustic properties of the token, and one is much better at reporting those syntactic features which affect meaning than those which don't. One might put it that one does not hear the formant relations in utterances of sentences even though one does hear the linguistic relations and the formant structure (*inter alia*) causally determines *which* linguistic relations one hears. Of course, which descriptions are consciously accessible is to some extent labile. Artists and phoneticians learn consciously to note properties of their sensory experience to which the layman is blind and deaf. This fact is by no means uninteresting; some of its consequences for the theory of internal representation will be pursued in Chapter 4.

Where we have gotten to is that the etiology of perceptual analyses involves a series of redescriptions of the environment, and that the initial description in this series specifies perceptually relevant physical properties of the environment. Perception must involve hypothesis formation and confirmation because the organism must somehow manage to infer the appropriate task-relevant description of the environment *from* its physical description together with whatever background information about the structure of the environment it has available. Notoriously, this inference is nondemonstrative: There is typically no *conceptual* connection between a perceptual category and its sensory indicants; an indefinite number of perceptual analyses will, in principle, be compatible with any given specification of a sensory input.[18] On this account, then, perceptual integrations are most plausibly viewed as species of inferences-to-the-best-explanation, the computational problem in perceptual integration being that of choosing the best hypothesis about the distal source of proximal stimulations.

There is, in short, an enormous problem about how to relate the conditions for applying physical descriptions to the conditions for applying

[17] I have been assuming that the representations of an environmental event that are assigned in the course of perceptual analysis are computed serially. Actually, a weaker assumption will do: viz., that at least *some* information about physical parameters normally 'gets in' before any higher-level representations are computed. I don't suppose this is a claim that any psychologist would wish to deny.

[18] Hence the possibility of perceptual illusions. For a discussion of perception that runs along the lines I have endorsed, see Gregory (1966) or Teuber (1960).

such descriptions as 'time for tea'. My present point is that the computational capacities of the organism must constitute a solution to such problems insofar as its perceptual judgments are (a) mediated by sensory information, and (b) true.

It is time to draw the moral, which will by now sound familiar. If one accepts, even in rough outline, the kind of approach to perception just surveyed, then one is committed to the view that perceptual processes involve computing a series of redescriptions of impinging environmental stimuli. But this is to acknowledge that perception presupposes a representational system; indeed, a representational system rich enough to distinguish between the members of sets of properties all of which are exhibited by the same event. If, for example, *e* is a token of a sentence type, and if understanding/perceptually analyzing *e* requires determining which sentence type it is a token of (see the first part of Chapter 3), then on the current view of understanding/perceptually analyzing, a series of representations of *e* will have to be computed. And this series will have to include, and distinguish between, representations which specify the acoustic, phonological, morphological, and syntactic properties of the token. It will have to include all these representations because, so far as anybody knows, each is essential for determining the type/token relation for utterances of sentences. It will have to distinguish among them because, so far as anyone knows, properties of sentences that are defined over any one of these kinds of representation will, ipso facto, be undefined for any of the others.

We are back to our old point that psychological processes are typically computational and computation presupposes a medium for representing the structures over which the computational operations are defined. Instead of further reiterating this point, however, I shall close this part of the discussion by making explicit two assumptions that the argument depends upon.

I have claimed that the only available models for deciding, concept learning, and perceiving all treat these phenomena as computational and hence presuppose that the organism has access to a language in which the computations are carried through. But, of course, this argument requires taking the models literally as at least schemata for explanations of the phenomena. In particular, it requires assuming that if such a model attributes a state to an organism, then insofar as we accept the model we are ontologically committed to the state. Now many philosophers do not like to play the game this way. They are willing to accept computational accounts of cognitive processes if only for lack of viable theoretical alternatives. But the models are accepted only as *façons de parler*, some reductionist program having previously been endorsed.

As I remarked in the introduction, I cannot prove that it is impossible to get the force of computational psychological theories in some framework which treats mental states as (e.g.) behavioral dispositions. But I think it is

fair to say that no one has ever given any reason to believe that it is possible, and the program seems increasingly hopeless as empirical research reveals how complex the mental structures of organisms, and the interactions of such structures, really are. I have assumed that one oughtn't to eat the cake unless one is prepared to bite the bullet. If our psychological theories commit us to a language of thought, we had better take the commitment seriously and find out what the language of thought is like.

My second point is that, while I have argued for a language of thought, what I have really shown is at best that there is a language of computation; for thinking is something that *organisms* do. But the sorts of data processes I have been discussing, though they may well go on in the nervous systems of organisms, are presumably not, in the most direct sense, attributable to the organisms themselves.

There is, obviously, a horribly difficult problem about what determines what a person (as distinct from his body, or parts of his body) did. Many philosophers care terrifically about drawing this distinction, and so they should: It can be crucial in such contexts as the assessment of legal or moral responsibility. It can also be crucial where the goal is phenomenology: i.e., the systematic characterization of the *conscious* states of the organism.[19] But whatever relevance the distinction between states of the organism and states of its nervous system may have for *some* purposes, there is no particular reason to suppose that it is relevant to the purposes of cognitive psychology.

What cognitive psychologists typically try to do is to characterize the etiology of behavior in terms of a series of transformations of information. See the second part of Chapter 2, where this notion will be spelled out at length; but, roughly speaking, information is said to be available to the organism when the neural event which encodes it is one of the causal determinants of the behavior of the organism. 'Behavior' is itself construed broadly (and intuitively) to include, say, thinking and dreaming but not accelerating when you fall down the stairs.

If one has these ends in view, it turns out (again on empirical rather than conceptual grounds) that the ordinary distinction between what the

[19] It is, of course, quite unclear whether the latter undertaking can be carried through in any very revealing way. That will depend upon whether there *are* generalizations which hold (just) for conscious mental states, and that depends in turn on whether the conscious states of an organism have more in common with one another than with the *un*conscious states of the nervous system of the organism. It is, in this sense, an open question whether conscious psychological states provide a natural domain for a theory, just as it is an open question whether, say, all the objects in Minnesota provide a natural domain for a theory. One can't have theories of everything under every description, and which descriptions of which things can be generalized is not usually a question that can be settled a priori. I should have thought that, since Freud, the burden of proof has shifted to those who maintain that the conscious states (of human beings) do form a theoretical domain.

organism does, knows, thinks, and dreams, and what happens to and in its nervous system, does not seem to be frightfully important. The natural kinds, for purposes of theory construction, appear to include some things that the organism does, some things that happen in the nervous system of the organism, and some things that happen in its environment. It is simply no good for philosophers to urge that, since this sort of theory does not draw the usual distinctions, the theory *must* be a muddle. It cannot be an objection to a theory that there are some distinctions it does not make; if it were, it would be an objection to every theory. (Aristotelians thought that it was an argument *against* the Galelean mechanics that it did not distinguish between sublunary and heavenly bodies; i.e., that its generalizations were defined for both. This line of argument is now widely held to have been ill-advised.)

In short, the states of the organism postulated in theories of cognition would not count as states of the organism for purposes of, say, a theory of legal or moral responsibility. But so what? What matters is that they should count as states of the organism for *some* useful purpose. In particular, what matters is that they should count as states of the organism for purposes of constructing psychological theories that are true.

To put this point the other way around, if psychological theories fail to draw the usual distinctions between some of the things that happen to organisms and some of the things that organisms do, that does *not* imply that psychologists are committed to denying that there are such distinctions or that they should be drawn for some purposes or other. Nor does it imply that psychologists are (somehow, and whatever precisely this may mean) committed to 'redrawing the logical geography' of our ordinary mental concepts. What *is* implied (and all that is implied) is just that the distinction between actions and happenings isn't a *psychological* distinction. Lots of very fine distinctions, after all, are not.[20]

[20] These remarks connect, in obvious ways, with the ones that concluded the introduction: The various intellectual disciplines typically cross-classify one another's subject matter.