1

Intentional Systems

I wish to examine the concept of a system whose behavior can be—at least sometimes—explained and predicted by relying on ascriptions to the system of beliefs and desires (and hopes, fears, intentions, hunches, . . .). I will call such systems *intentional systems*, and such explanations and predictions intentional explanations and predictions, in virtue of the intentionality of the idioms of belief and desire (and hope, fear, intention, hunch, . . .).[1]

I used to insist on capitalizing "intentional" wherever I meant to be using Brentano's notion of *intentionality*, in order to distinguish this technical term from its cousin, e.g., "an intentional shove", but the technical term is now in much greater currency, and since almost everyone else who uses the term seems content to risk this confusion, I have decided, with some trepidation, to abandon my typographical eccentricity. But let the uninitiated reader beware: "intentional" as it occurs here is *not* the familiar term of layman's English.[2] For me, as for many recent authors, intentionality is primarily a feature of linguistic entities—idioms, contexts—and for my purposes here we can be satisfied that an idiom is intentional if substitution of codesignative terms do not preserve truth or if the "objects" of the idiom are not capturable in the usual way by quantifiers. I discuss this in more detail in *Content and Consciousness.*[3]

I

The first point to make about intentional systems[4] as I have just defined them is that a particular thing is an intentional system only in relation to the strategies of someone who is trying to explain and

predict its behavior. What this amounts to can best be brought out by example. Consider the case of a chess-playing computer, and the different strategies or stances one might adopt as its opponent in trying to predict its moves. There are three different stances of interest to us. First there is the *design stance*. If one knows exactly how the computer is designed (including the impermanent part of its design: its program) one can predict its designed response to any move one makes by following the computation instructions of the program. One's prediction will come true provided only that the computer performs as designed—that is, without breakdown. Different varieties of design-stance predictions can be discerned, but all of them are alike in relying on the notion of *function*, which is purpose-relative or teleological. That is, a design of a system breaks it up into larger or smaller functional parts, and design-stance predictions are generated by assuming that each functional part will function properly. For instance, the radio engineer's schematic wiring diagrams have symbols for each resistor, capacitor, transistor, etc.—*each with its task to perform*—and he can give a design-stance prediction of the behavior of a circuit by assuming that each element performs its task. Thus one can make design-stance predictions of the computer's response at several different levels of abstraction, depending on whether one's design treats as smallest functional elements strategy-generators and consequence-testers, multipliers and dividers, or transistors and switches. (It should be noted that not all diagrams or pictures are designs in this sense, for a diagram may carry no information about the functions—intended or observed—of the elements it depicts.)

We generally adopt the design stance when making predictions about the behavior of mechanical objects, e.g., "As the typewriter carriage approaches the margin, a bell will ring (provided the machine is in working order)," and more simply, "Strike the match and it will light." We also often adopt this stance in predictions involving natural objects: "Heavy pruning will stimulate denser foliage and stronger limbs." The essential feature of the design stance is that we make predictions solely from knowledge or assumptions about the system's functional design, irrespective of the physical constitution or condition of the innards of the particular object.

Second, there is what we may call the *physical stance*. From this stance our predictions are based on the actual physical state of the particular object, and are worked out by applying whatever knowledge we have of the laws of nature. It is from this stance alone that we can predict the malfunction of systems (unless, as sometimes happens these days, a system is *designed* to malfunction after a certain time,

in which case malfunctioning in one sense becomes a part of its proper functioning). Instances of predictions from the physical stance are common enough: "If you turn on the switch you'll get a nasty shock," and, "When the snows come that branch will break right off." One seldom adopts the physical stance in dealing with a computer just because the number of critical variables in the physical constitution of a computer would overwhelm the most prodigious calculator. Significantly, the physical stance is generally reserved for instances of breakdown, where the condition preventing normal operation is generalized and easily locatable, e.g., "Nothing will happen when you type in your questions, because it isn't plugged in," or, "It won't work with all that flood water in it." Attempting to give a physical account or prediction of the chess-playing computer would be a pointless and herculean labor, but it would work in principle. One could predict the response it would make in a chess game by tracing out the effects of the input energies all the way through the computer until once more type was pressed against paper and a response was printed. (Because of the digital nature of computers, quantum-level indeterminacies, if such there be, will cancel out rather than accumulate, unless of course a radium "randomizer" or other amplifier of quantum effects is built into the computer).

The best chess-playing computers these days are practically inaccessible to prediction from either the design stance or the physical stance; they have become too complex for even their own designers to view from the design stance. A man's best hope of defeating such a machine in a chess match is to predict its responses by figuring out as best he can what the best or most rational move would be, given the rules and goals of chess. That is, one assumes not only (1) that the machine will function as designed, but (2) that the design is optimal as well, that the computer will "choose" the most rational move. Predictions made on these assumptions may well fail if either assumption proves unwarranted in the particular case, but still this *means* of prediction may impress us as the most fruitful one to adopt in dealing with a particular system. Put another way, when one can no longer hope to beat the machine by utilizing one's knowledge of physics or programming to anticipate its responses, one may still be able to avoid defeat by treating the machine rather like an intelligent human opponent.

We must look more closely at this strategy. A prediction relying on the assumption of the system's rationality is relative to a number of things. First, rationality here so far means nothing more than optimal design relative to a goal or optimally weighted hierarchy of goals

(checkmate, winning pieces, defense, etc., in the case of chess) and a set of constraints (the rules and starting position). Prediction itself is, moreover, relative to the nature and extent of the information the system has at the time about the field of endeavor. The question one asks in framing a prediction of this sort is: What is the most rational thing for the computer to do, given goals $x,y,z, \ldots$, constraints $a,b,c, \ldots$ and information (including misinformation, if any) about the present state of affairs $p,q,r, \ldots$? In predicting the computer's response to my chess move, my assessment of the computer's most rational move may depend, for instance, not only on my assumption that the computer has information about the present disposition of all the pieces, but also on whether I believe the computer has information about my inability to see four moves ahead, the relative powers of knights and bishops, and my weakness for knight-bishop exchanges. In the end I may not be able to frame a very good prediction, if I am unable to determine with any accuracy what information and goals the computer has, or if the information and goals I take to be given do not dictate any one best move, or if I simply am not so good as the computer is at generating an optimal move from this given. Such predictions then are very precarious; not only are they relative to a set of postulates about goals, constraints, and information, and not only do they hinge on determining an optimal response in situations where we may have no clear criteria for what is optimal, but also they are vulnerable to short-circuit falsifications that are in principle unpredictable from this stance. Just as design-stance predictions are vulnerable to malfunctions (by depending on the assumption of no malfunction), so these predictions are vulnerable to design weaknesses and lapses (by depending on the assumption of optimal design). It is a measure of the success of contemporary program designers that these precarious predictions turn out to be true with enough regularity to make the method useful.

The dénouement of this extended example should now be obvious: this third stance, with its assumption of rationality, is the *intentional stance*; the predictions one makes from it are intentional predictions; one is viewing the computer as an intentional system. One predicts behavior in such a case by ascribing to the system *the possession of certain information* and supposing it to be *directed by certain goals*, and then by working out the most reasonable or appropriate action on the basis of these ascriptions and suppositions. It is a small step to calling the information possessed the computer's *beliefs*, its goals and subgoals its *desires*. What I mean by saying that this is a small step, is that the notion of possession of information or misinformation is

just as intentional a notion as that of belief. The "possession" at issue is hardly the bland and innocent notion of storage one might suppose; it is, and must be, "epistemic possession"—an analogue of belief. Consider: the Frenchman who possesses the *Encyclopedia Britannica* but knows no English might be said to "possess" the information in it, but if there is such a sense of possession, it is not strong enough to serve as the sort of possession the computer must be supposed to enjoy, relative to the information it *uses* in "choosing" a chess move. In a similar way, the goals of a goal-directed computer must be specified intentionally, just like desires.

Lingering doubts about whether the chess-playing computer *really* has beliefs and desires are misplaced; for the definition of intentional systems I have given does not say that intentional systems *really* have beliefs and desires, but that one can explain and predict their behavior by *ascribing* beliefs and desires to them, and whether one calls what one ascribes to the computer beliefs or belief-analogues or information complexes or intentional whatnots makes no difference to the nature of the calculation one makes on the basis of the ascriptions. One will arrive at the same predictions whether one forthrightly thinks in terms of the computer's beliefs and desires, or in terms of the computer's information-store and goal-specifications. The inescapable and interesting fact is that for the best chess-playing computers of today, intentional explanation and prediction of their behavior is not only common, but works when no other sort of prediction of their behavior is manageable. We do quite successfully treat these computers as intentional systems, and we do this independently of any considerations about what substance they are composed of, their origin, their position or lack of position in the community of moral agents, their consciousness or self-consciousness, or the determinacy or indeterminacy of their operations. The decision to adopt the strategy is pragmatic, and is not intrinsically right or wrong. One can always refuse to adopt the intentional stance toward the computer, and accept its checkmates. One can switch stances at will without involving oneself in any inconsistencies or inhumanities, adopting the intentional stance in one's role as opponent, the design stance in one's role as redesigner, and the physical stance in one's role as repairman.

This celebration of our chess-playing computer is not intended to imply that it is a completely adequate model or simulation of Mind, or intelligent human or animal activity; nor am I saying that the attitude we adopt toward this computer is precisely the same that we adopt toward a creature we deem to be conscious and rational. All that has been claimed is that on occasion, a purely physical system can be so

complex, and yet so organized, that we find it convenient, explanatory, pragmatically necessary for prediction, to treat it as if it had beliefs and desires and was rational. The chess-playing computer is just that, a machine for playing chess, which no man or animal is; and hence its "rationality" is pinched and artificial.

Perhaps we could straightforwardly expand the chess-playing computer into a more faithful model of human rationality, and perhaps not. I prefer to pursue a more fundamental line of inquiry first.

When should we expect the tactic of adopting the intentional stance to pay off? Whenever we have reason to suppose the assumption of optimal design is warranted, and doubt the practicality of prediction from the design or physical stance. Suppose we travel to a distant planet and find it inhabited by things moving about its surface, multiplying, decaying, apparently reacting to events in the environment, but otherwise as unlike human beings as you please. Can we make intentional predictions and explanations of their behavior? If we have reason to suppose that a process of natural selection has been in effect, then we can be assured that the populations we observe have been selected in virtue of their design: they will respond to at least some of the more common event-types in this environment in ways that are normally appropriate—that is, conducive to propagation of the species.* Once we have tentatively identified the perils and succors of the environment (relative to the constitution of the inhabitants, not ours), we shall be able to estimate which goals and which weighting of goals will be optimal relative to the creatures' *needs* (for survival and propagation), which sorts of information about the environment will be *useful* in guiding goal-directed activity, and which activities will be appropriate given the environmental circumstances. Having doped out these conditions (which will always be subject to revision) we can proceed at once to ascribe beliefs and desires to the creatures. Their behavior will "manifest" their beliefs by being seen as the actions which, given the creatures' desires, would be appropriate to such beliefs as would be appropriate to the environmental stimulation. Desires, in turn, will be "manifested" in behavior as those appropriate desires (given the needs of the creature) to which the actions of the creature would be appropriate, given the creature's beliefs. The circularity of these interlocking specifications is no accident. Ascriptions of beliefs and desires must be interdependent, and the only points of anchorage

*Note that what is *directly* selected, the gene, is a diagram and not a design; it is selected, however, because it happens to ensure that its bearer has a certain (functional) design. This was pointed out to me by Woodruff.

are the demonstrable needs for survival, the regularities of behavior, and the assumption, grounded in faith in natural selection, of optimal design. Once one has ascribed beliefs and desires, however, one can at once set about predicting behavior on their basis, and if evolution has done its job—as it must over the long run—our predictions will be reliable enough to be useful.

It might at first seem that this tactic unjustifiably imposes human categories and attributes (belief, desire, and so forth) on these alien entities. It is a sort of anthropomorphizing, to be sure, but it is conceptually innocent anthropomorphizing. We do not have to suppose these creatures share with us any peculiarly human inclinations, attitudes, hopes, foibles, pleasures, or outlooks; their actions may not include running, jumping, hiding, eating, sleeping, listening, or copulating. All we transport from our world to theirs are the categories of rationality, perception (information input by some "sense" modality or modalities—perhaps radar or cosmic radiation), and action. The question of whether we can expect them to share any of our beliefs or desires is tricky, but there are a few points that can be made at this time; in virtue of their rationality they can be supposed to share our belief in logical truths,* and we cannot suppose that they normally desire their own destruction, for instance.

II

When one deals with a system—be it man, machine, or alien creature—by explaining and predicting its behavior by citing its beliefs and desires, one has what might be called a "theory of behavior" for the system. Let us see how such intentional theories of behavior relate to other putative theories of behavior.

One fact so obvious that it is easily overlooked is that our "commonsense" explanations and predictions of the behavior of both men and animals are intentional. We start by assuming rationality. We do not *expect* new acquaintances to react irrationally to particular topics or eventualities, but when they do we learn to adjust our strategies accordingly, just as, with a chess-playing computer, one sets out with a high regard for its rationality and adjusts one's estimate downward wherever performance reveals flaws. The presumption of rationality is so strongly entrenched in our inference habits that when our predic-

*Cf. Quine's argument about the necessity of "discovering" our logical connectives in any language we can translate in *Word and Object* (Cambridge, Mass.: MIT, 1960), Section 13. More will be said in defense of this below.

tions prove false, we at first cast about for adjustments in the information-possession conditions (he must not have heard, he must not know English, he must not have seen x, been aware that y, etc.) or goal weightings, before questioning the rationality of the system as a whole. In extreme cases personalities may prove to be so unpredictable from the intentional stance that we abandon it, and if we have accumulated a lot of evidence in the meanwhile about the nature of response patterns in the individual, we may find that a species of design stance can be effectively adopted. This is the fundamentally different attitude we occasionally adopt toward the insane. To watch an asylum attendant manipulate an obsessively countersuggestive patient, for instance, is to watch something radically unlike normal interpersonal relations.

Our prediction of animal behavior by "common sense" is also intentional. Whether or not sentimental folk go overboard when they talk to their dogs or fill their cats' heads with schemes and worries, even the most hardboiled among us predict animals' behavior intentionally. If we observe a mouse in a situation where it can see a cat waiting at one mousehole and cheese at another, we know which way the mouse will go, providing it is not deranged; our prediction is not based on our familiarity with maze-experiments or any assumptions about the sort of special training the mouse has been through. We suppose the mouse can see the cat and the cheese, and hence has beliefs (belief-analogues, intentional whatnots) to the effect that there is a cat to the left, cheese to the right, and we ascribe to the mouse also the desire to eat the cheese and the desire to avoid the cat (subsumed, appropriately enough, under the more general desires to eat and to avoid peril); so we predict that the mouse will do what is appropriate to such beliefs and desires, namely, go to the right in order to get the cheese and avoid the cat. Whatever academic allegiances or theoretical predilections we may have, we would be astonished if, in the general run, mice and other animals falsified such intentional predictions of their behavior. Indeed, experimental psychologists of every school would have a hard time devising experimental situations to support their various theories without the help of their intentional expectations of how the test animals will respond to circumstances.

Earlier I alleged that even creatures from another planet would share with us our beliefs in logical truths; light can be shed on this claim by asking whether mice and other animals, in virtue of being intentional systems, also believe the truths of logic. There is something bizarre in the picture of a dog or mouse cogitating a list of tautologies, but we can avoid that picture. The assumption that something is an intentional

system is the assumption that it is rational; that is, one gets nowhere with the assumption that entity x has beliefs $p,q,r, \ldots$ unless one also supposes that x believes what follows from $p,q,r, \ldots$; otherwise there is no way of ruling out the prediction that x will, in the face of its beliefs $p,q,r, \ldots$ do something utterly stupid, and, if we cannot rule out *that* prediction, we will have acquired no predictive power at all. So whether or not the animal is said to *believe* the *truths* of logic, it must be supposed to *follow* the *rules* of logic. Surely our mouse follows or believes in *modus ponens*, for we ascribed to it the beliefs: (a) *there is a cat to the left*, and (b) *if there is a cat to the left, I had better not go left*, and our prediction relied on the mouse's ability to get to the conclusion. In general there is a trade-off between rules and truths; we can suppose x to have an inference rule taking A to B or we can give x the belief in the "theorem": *if A then B*. As far as our predictions are concerned, we are free to ascribe to the mouse either a few inference rules and belief in many logical propositions, or many inference rules and few if any logical beliefs.* We can even take a patently nonlogical belief like (b) and recast it as an inference rule taking (a) to the desired conclusion.

Will all logical truths appear among the beliefs of any intentional system? If the system were ideally or perfectly rational, all logical truths would appear, but any actual intentional system will be imperfect, and so not all logical truths must be ascribed as beliefs to any system. Moreover, not all the inference rules of an actual intentional system may be valid; not all its inference-licensing beliefs may be truths of logic. Experience may indicate where the shortcomings lie in any particular system. If we found an imperfectly rational creature whose allegiance to *modus ponens*, say, varied with the subject matter, we could characterize that by excluding *modus ponens* as a rule and ascribing in its stead a set of nonlogical inference rules covering the *modus ponens* step for each subject matter where the rule was followed. Not surprisingly, as we discover more and more imperfections (as we banish more and more logical truths from the creature's beliefs), our efforts at intentional prediction become more and more cumbersome and undecidable, for we can no longer count on the beliefs, desires, and actions going together that *ought* to go together. Eventually we end up, following this process, by predicting from the

*Accepting the argument of Lewis Carroll, in "What the Tortoise Said to Achilles", *Mind* (1895), reprinted in I. M. Copi and J. A. Gould, *Readings on Logic* (New York: MacMillan, 1964), we cannot allow all the rules for a system to be replaced by beliefs, for this would generate an infinite and unproductive nesting of distinct beliefs about what can be inferred from what.

design stance; we end up, that is, dropping the assumption of rationality.*

This migration from common-sense intentional explanations and predictions to more reliable design-stance explanations and predictions that is forced on us when we discover that our subjects are imperfectly rational is, independently of any such discovery, the proper direction for theory builders to take whenever possible. In the end, we want to be able to explain the intelligence of man, or beast, in terms of his design, and this in turn in terms of the natural selection of this design; so whenever we stop in our explanations at the intentional level we have left over an unexplained instance of intelligence or rationality. This comes out vividly if we look at theory building from the vantage point of economics.

Any time a theory builder proposes to call any event, state, structure, etc., in any system (say the brain of an organism) a *signal* or *message* or *command* or otherwise endows it with content, he *takes out a loan* of intelligence. He implicitly posits along with his signals, messages, or commands, something that can serve as a signal-*reader*, message-*understander*, or *commander*, else his "signals" will be for naught, will decay unreceived, uncomprehended. This loan must be repaid eventually by finding and analyzing away these readers or comprehenders; for, failing this, the theory will have among its elements unanalyzed man-analogues endowed with enough intelligence to read the signals, etc., and thus the theory will *postpone* answering the major question: what makes for intelligence? The intentionality of all such talk of signals and commands reminds us that rationality is being taken for granted, and in this way shows us where a theory is incomplete. It is this feature that, to my mind, puts a premium on the yet unfinished task of devising a rigorous definition of intentionality, for if we can lay claim to a purely formal criterion of intentional discourse, we will have what amounts to a medium of exchange for assessing theories of behavior. Intentionality *abstracts* from the inessential details of the various forms intelligence-loans can take (e.g., signal-readers, volition-emitters, librarians in the corridors of memory, egos and superegos) and serves as a reliable means of detecting exactly where a theory is *in the red* relative to the task of explaining intelligence; wherever a theory relies on a formulation bearing the logical marks of intentionality, there a little man is concealed.

*This paragraph owes much to discussion with John Vickers, whose paper "Judgment and Belief", in K. Lambert, *The Logical Way of Doing Things* (New Haven, Conn.: Yale, 1969), goes beyond the remarks here by considering the problems of the relative strength or weighting of beliefs and desires.

This insufficiency of intentional explanation from the point of view of psychology has been widely felt and as widely misconceived. The most influential misgivings, expressed in the behaviorism of Skinner and Quine, can be succinctly characterized in terms of our economic metaphor. Skinner's and Quine's adamant prohibitions of intentional idioms at all levels of theory is the analogue of rock-ribbed New England conservatism: no deficit spending when building a theory! In Quine's case, the abhorrence of loans is due mainly to his fear that they can never be repaid, whereas Skinner stresses rather that what is borrowed is worthless to begin with. Skinner's suspicion is that intentionally couched claims are empirically vacuous, in the sense that they are altogether too easy to accommodate to the data, like the *virtus dormitiva* Molière's doctor ascribes to the sleeping powder (see Chapter 4 for a more detailed discussion of these issues). Questions can be begged on a temporary basis, however, permitting a mode of prediction and explanation not totally vacuous. Consider the following intentional prediction: if I were to ask a thousand American mathematicians how much seven times five is, more than nine hundred would respond by saying that it was thirty-five. (I have allowed for a few to mis-hear my question, a few others to be obstreperous, a few to make slips of the tongue.) If you doubt the prediction, you can test it; I would bet good money on it. It seems to have empirical content because it can, in a fashion, be tested, and yet it is unsatisfactory as a prediction of an empirical theory of psychology. It works, of course, because of the contingent, empirical—but evolution-guaranteed—fact that men in general are well enough designed both to get the answer right and to want to get it right. It will hold with as few exceptions for any group of Martians with whom we are able to converse, for it is not a prediction just of *human* psychology, but of the "psychology" of intentional systems generally.

Deciding on the basis of available empirical evidence that something is a piece of copper or a lichen permits one to make predictions based on the empirical theories dealing with copper and lichens, but deciding on the basis of available evidence that something is (may be treated as) an intentional system permits predictions having a normative or logical basis rather than an empirical one, and hence the success of an intentional prediction, based as it is on no particular picture of the system's design, cannot be construed to confirm or disconfirm any particular pictures of the system's design.

Skinner's reaction to this has been to try to frame predictions purely in non-intentional language, by predicting bodily responses to physical stimuli, but to date this has not provided him with the alternative

mode of prediction and explanation he has sought, as perhaps an extremely cursory review can indicate. To provide a setting for non-intentional prediction of behavior, he invented the Skinner box, in which the rewarded behavior of the occupant—say, a rat—is a highly restricted and stereotypic bodily motion—usually pressing a bar with the front paws.

The claim that is then made is that once the animal has been trained, a law-like relationship is discovered to hold between non-intentionally characterized events: controlling stimuli and bar-pressing responses. A regularity is discovered to hold, to be sure, but the fact that it is between non-intentionally defined events is due to a property of the Skinner box and not of the occupant. For let us turn our prediction about mathematicians into a Skinnerian prediction: strap a mathematician in a Skinner box so he can move only his head; display in front of him a card on which appear the marks: "How much is seven times five?"; move into the range of his head-motions two buttons, over one of which is the mark "35" and over the other "34"; place electrodes on the soles of his feet and give him a few quick shocks; the controlling stimulus is then to be the sound: "Answer now!" I predict that in a statistically significant number of cases, even *before* training trials to condition the man to press button "35" with his forehead, he will do this when given the controlling stimulus. Is this a satisfactory scientific prediction just because it eschews the intentional vocabulary? No, it is an intentional prediction disguised by so restricting the environment that only one bodily motion is available to fulfill the intentional *action* that anyone would prescribe as appropriate to the circumstances of perception, belief, desire. That it is action, not merely motion, that is predicted can also be seen in the case of subjects less intelligent than mathematicians. Suppose a mouse were trained, in a Skinner box with a food reward, to take exactly four steps forward and press a bar with its nose; if Skinner's laws truly held between stimuli and responses defined in terms of bodily motion, were we to move the bar an inch farther away, so four steps did not reach it, Skinner would have to predict that the mouse would jab its nose into the empty air rather than take a fifth step.

A variation of Skinnerian theory designed to meet this objection acknowledges that the trained response one predicts is not truly captured in a description of skeletal motion alone, but rather in a description of an environmental effect achieved: the bar going down, the "35" button being depressed. This will also not do. Suppose we could in fact train a man or animal to achieve an environmental effect, as this theory proposes. Suppose, for instance, we train a man to push a but-

ton under the longer of two displays, such as drawings or simple designs, that is, we reward him when he pushes the button under the longer of two pictures of pencils, or cigars, etc. The miraculous consequence of this theory, were it correct, would be that if, after training him on simple views, we were to present him with the Müller-Lyer arrow-head illusion, he would be immune to it, for *ex hypothesi* he has been trained to achieve an *actual* environmental effect (choosing the display that *is* longer), not a *perceived* or *believed* environmental effect (choosing the display that *seems* longer). The reliable prediction, again, is the intentional one.*

Skinner's experimental design is supposed to eliminate the intentional, but it merely masks it. Skinner's non-intentional predictions work to the extent they do, not because Skinner has truly found non-intentional behavioral laws, but because the highly reliable intentional predictions underlying his experimental situations (the rat desires food and believes it will get food by pressing the bar—something for which it has been given good evidence—so it will press the bar) are disguised by leaving virtually no room in the environment for more than one bodily motion to be the appropriate action and by leaving virtually no room in the environment for discrepancy to arise between the subject's beliefs and the reality.

Where, then, should we look for a satisfactory theory of behavior? Intentional theory is vacuous as psychology because it presupposes and does not explain rationality or intelligence. The apparent successes of Skinnerian behaviorism, however, rely on hidden intentional predictions. Skinner is right in recognizing that intentionality can be no *foundation* for psychology, and right also to look for purely mechanistic regularities in the activities of his subjects, but there is little reason to suppose they will lie on the surface in gross behavior—except, as we have seen, when we put an artificial straitjacket on an intentional regularity. Rather, we will find whatever mechanistic regularities there are in the functioning of internal systems whose design approaches the optimal (relative to some ends). In seeking knowledge of internal design our most promising tactic is to take out intelligence-loans, endow peripheral and internal events with content, and then look for mechanisms that will function appropriately with such "messages" so that we can pay back the loans. This tactic is hardly untried. Research in artificial intelligence, which has produced, among other things, the

*R. L. Gregory, *Eye and Brain* (London: World University Library, 1966): p. 137, reports that pigeons and fish given just this training are, not surprisingly, susceptible to visual illusions of length.

chess-playing computer, proceeds by working from an intentionally characterized problem (how to get the computer to consider the right sorts of information, make the right decisions) to a design-stance solution—an approximation of optimal design. Psychophysicists and neurophysiologists who routinely describe events in terms of the transmission of information within the nervous system are similarly borrowing intentional capital—even if they are often inclined to ignore or disavow their debts.

Finally, it should not be supposed that, just because intentional theory is vacuous as psychology, in virtue of its assumption of rationality, it is vacuous from all points of view. Game theory, for example, is inescapably intentional,[5] but as a formal normative theory and not a psychology this is nothing amiss. Game-theoretical predictions applied to human subjects achieve their accuracy in virtue of the evolutionary guarantee that man is well designed as a game player, a special case of rationality. Similarly, economics, the social science of greatest predictive power today, is not a psychological theory and presupposes what psychology must explain. Economic explanation and prediction is intentional (although some is disguised) and succeeds to the extent that it does because individual men are in general good approximations of of the optimal operator in the marketplace.

III

The concept of an intentional system is a relatively uncluttered and unmetaphysical notion, abstracted as it is from questions of the composition, constitution, consciousness, morality, or divinity of the entities falling under it. Thus, for example, it is much easier to decide whether a machine can be an intentional system than it is to decide whether a machine can *really* think, or be conscious, or morally responsible. This simplicity makes it ideal as a source of order and organization in philosophical analyses of "mental" concepts. Whatever else a person might be—embodied mind or soul, self-conscious moral agent, "emergent" form of intelligence—he is an intentional system, and whatever follows just from being an intentional system is thus true of a person. It is interesting to see just how much of what we hold to be the case about persons or their minds follows directly from their being intentional systems. To revert for a moment to the economic metaphor, the guiding or challenging question that defines work in the philosophy of mind is this: are there mental treasures that cannot be purchased with intentional coin? If not, a considerable unification of science can be foreseen in outline. Of special importance for such an

examination is the subclass of intentional systems that have language, that can communicate; for these provide a framework for a theory of consciousness. In *Content and Consciousness*, part II, and in parts III and IV of this volume I have attempted to elaborate such a theory; here I would like to consider its implications for the analysis of the concept of belief. What will be true of human believers just in virtue of their being intentional systems with the capacity to communicate?

Just as not all intentional systems currently known to us can fly or swim, so not all intentional systems can talk, but those which can do this raise special problems and opportunities when we come to ascribe beliefs and desires to them. That is a massive understatement; without the talking intentional systems, of course, there would be no ascribing beliefs, no theorizing, no assuming rationality, no predicting. The capacity for language is without doubt the crowning achievement of evolution, an achievement that feeds on itself to produce ever more versatile and subtle rational systems, but still it can be looked at as an adaptation which is subject to the same conditions of environmental utility as any other behavioral talent. When it is looked at in this way several striking facts emerge. One of the most pervasive features of evolutionary histories is the interdependence of distinct organs and capacities in a species. Advanced eyes and other distance receptors are of no utility to an organism unless it develops advanced means of locomotion; the talents of a predator will not accrue to a species that does not evolve a carnivore's digestive system. The capacities of belief and communication have prerequisites of their own. We have already seen that there is no point in ascribing beliefs to a system unless the beliefs ascribed are in general appropriate to the environment, and the system responds appropriately to the beliefs. An eccentric expression of this would be: the capacity to believe would have no survival value unless it were a capacity to believe truths. What is eccentric and potentially misleading about this is that it hints at the picture of a species "trying on" a faculty giving rise to beliefs most of which were false, having its inutility demonstrated, and abandoning it. A species might "experiment" by mutation in any number of inefficacious systems, but none of these systems would deserve to be called belief systems precisely because of their defects, their nonrationality, and hence a false belief system is a conceptual impossibility. To borrow an example from a short story by MacDonald Harris, a soluble fish is an evolutionary impossibility, but a system for false beliefs cannot even be given a coherent description. The same evolutionary bias in favor of truth prunes the capacity to communicate as it develops; a capacity for false communication would not be a capacity for communication at all, but

just an emission proclivity of no systematic value to the species. The faculty of communication would not gain ground in evolution unless it was by and large the faculty of transmitting true beliefs, which means only: the faculty of altering other members of the species in the direction of more optimal design.

This provides a foundation for explaining a feature of belief that philosophers have recently been at some pains to account for.[6] The concept of belief seems to have a normative cast to it that is most difficult to capture. One way of putting it might be that an avowal like "I believe that *p*" seems to imply in some fashion: "One ought to believe that *p*." This way of putting it has flaws, however, for we must then account for the fact that "I believe that *p*" seems to have normative force that "He believes that *p*", said of me, does not. Moreover, saying that one ought to believe this or that suggests that belief is voluntary, a view with notorious difficulties.[7] So long as one tries to capture the normative element by expressing it in the form of moral or pragmatic injunctions to believers, such as "one ought to believe the truth" and "one ought to act in accordance with one's beliefs", dilemmas arise. How, for instance, is one to follow the advice to believe the truth? Could one abandon one's sloppy habit of believing falsehoods? If the advice is taken to mean: believe only what you have convincing evidence for, it is the vacuous advice: believe only what you believe to be true. If alternatively it is taken to mean: believe only what is in fact the truth, it is an injunction we are powerless to obey.

The normative element of belief finds its home not in such injunctions but in the preconditions for the ascription of belief, what Phillips Griffiths calls "the general conditions for the possibility of application of the concept". For the concept of belief to find application, two conditions, we have seen, must be met: (1) In general, normally, more often than not, if *x* believes *p*, *p* is true. (2) In general, normally, more often than not, if *x* avows that *p*, he believes *p* [and, by (1), *p* is true]. Were these conditions not met, we would not have rational, communicating systems; we would not have believers or belief-avowers. The norm for belief is evidential well-foundedness (assuring truth in the long run), and the norm for avowal of belief is accuracy (which includes sincerity). These two norms determine pragmatic implications of our utterances. If I assert that *p* (or that I believe that *p*—it makes no difference), I assume the burden of defending my assertion on two fronts: I can be asked for evidence for the truth of *p*, and I can be asked for behavioral evidence that I do in fact believe *p*.[8] I do not need to examine my own behavior in order to be in a position to avow my belief that *p*, but if my sincerity or self-knowledge is challenged, this

is where I must turn to defend my assertion. But again, challenges on either point must be the exception rather than the rule if belief is to have a place among our concepts.

Another way of looking at the importance of this predominance of the normal is to consider the well-known circle of implications between beliefs and desires (or intentions) that prevent non-intentional behavioral definitions of intentional terms. A man's standing under a tree is a behavioral indicator of his belief that it is raining, but only on the assumption that he desires to stay dry, and if we then look for evidence that he wants to stay dry, his standing under the tree will do, but only on the assumption that he believes the tree will shelter him; if we ask him if he believes the tree will shelter him, his positive response is confirming evidence only on the assumption that he desires to tell us the truth, and so forth *ad infinitum*. It is this apparently vicious circle that turned Quine against the intentional (and foiled Tolman's efforts at operational definition of intentional terms), but if it is true that in any particular case a man's saying that *p* is evidence of his belief only conditionally, we can be assured that in the long run and in general the circle is broken; a man's assertions are, unconditionally, indicative of his beliefs, as are his actions in general. We get around the "privacy" of beliefs and desires by recognizing that in general anyone's beliefs and desires must be those he "ought to have" given the circumstances.

These two interdependent norms of belief, one favoring the truth and rationality of belief, the other favoring accuracy of avowal, normally complement each other, but on occasion can give rise to conflict. This is the "problem of incorrigibility". If rationality is the mother of intention, we still must wean intentional systems from the criteria that give them life, and set them up on their own. Less figuratively, if we are to make use of the concept of an intentional system in particular instances, at some point we must cease *testing* the assumption of the system's rationality, adopt the intentional stance, and grant without further ado that the system is qualified for beliefs and desires. For mute animals—and chess-playing computers—this manifests itself in a tolerance for less than optimal performance. We continue to ascribe beliefs to the mouse, and explain its actions in terms of them, after we have tricked it into some stupid belief. This tolerance has its limits of course, and the less felicitous the behavior—especially the less adaptable the behavior—the more hedged are our ascriptions. For instance, we are inclined to say of the duckling that "imprints" on the first moving thing it sees upon emerging from its shell that it "believes" the thing is its mother, whom it follows around, but we emphasize

the scare-quotes around "believes". For intentional systems that can communicate—persons for instance—the tolerance takes the form of the convention that a man is incorrigible or a special authority about his own beliefs. This convention is "justified" by the fact that evolution does guarantee that our second norm is followed. What better source could there be of a system's beliefs than its avowals? Conflict arises, however, whenever a person falls short of perfect rationality, and avows beliefs that either are strongly disconfirmed by the available empirical evidence or are self-contradictory or contradict other avowals he has made. If we lean on the myth that a man is perfectly rational, we must find his avowals less than authoritative: "You *can't* mean—understand—what you're saying!"; if we lean on his "right" as a speaking intentional system to have his word accepted, we grant him an irrational set of beliefs. Neither position provides a stable resting place; for, as we saw earlier, intentional explanation and prediction cannot be accommodated either to breakdown or to less than optimal design, so there is no coherent intentional description of such an impasse.*

Can any other considerations be brought to bear in such an instance to provide us with justification for one ascription of beliefs rather than another? Where should one look for such considerations? The Phenomenologist will be inclined to suppose that individual introspection will provide us a sort of data not available to the outsider adopting the intentional stance; but how would such data get used? Let the introspector amass as much inside information as you please; he must then communicate it to us, and what are we to make of his communications? We can suppose that they are incorrigible (barring corrigible verbal errors, slips of the tongue, and so forth), but we do not need Phenomenology to give us that option, for it amounts to the decision to lean on the accuracy-of-avowal norm at the expense of the rationality norm. If, alternatively, we demand certain standards of consistency and rationality of his utterances before we accept them as authoritative, what standards will we adopt? If we demand perfect rationality, we have simply flown to the other norm at the expense of the norm of accuracy of avowal. If we try to fix minimum standards at something less than perfection, what will guide our choice? Not

*Hintikka takes this bull by the horns. His epistemic logic is acknowledged to hold only for the ideally rational believer; were we to apply this logic to persons in the actual world in other than a normative way, thus making its implications *authoritative* about actual belief, the authority of persons would have to go by the board. Thus his rule A.CBB* (*Knowledge and Belief*, pp. 24–26), roughly that if one believes p one believes that one believes p, cannot be understood, as it is tempting to suppose, as a version of the incorrigibility thesis.

Phenomenological data, for the choice we make will determine what is to count as Phenomenological data. Not neurophysiological data either, for whether we interpret a bit of neural structure to be endowed with a particular belief content hinges on our having granted that the neural system under examination has met the standards of rationality for being an intentional system, an assumption jeopardized by the impasse we are trying to resolve. That is, one might have a theory about an individual's neurology that permitted one to "read off" or predict the propositions to which he would assent, but whether one's theory had uncovered his *beliefs*, or merely a set of assent-inducers, would depend on how consistent, reasonable, true we found the set of propositions.

John Vickers has suggested to me a way of looking at this question. Consider a set T of transformations that take beliefs into beliefs. The problem is to determine the set T_s for each intentional system S, so that if we know that S believes p, we will be able to determine other things that S believes by seeing what the transformations of p are for T_s. If S were ideally rational, every valid transformation would be in T_s; S would believe every logical consequence of every belief (and, ideally, S would have no false beliefs). Now we know that no actual intentional system will be ideally rational; so we must suppose any actual system will have a T with less in it. But we also know that, to qualify as an intentional system at all, S must have a T with some integrity; T cannot be empty. What rationale could we have, however, for fixing some set between the extremes and calling it *the* set for belief (for S, for earthlings, or for ten-year-old girls)? This is another way of asking whether we could replace Hintikka's normative theory of belief with an empirical theory of belief, and, if so, what evidence we would use. "Actually," one is tempted to say, "people do believe contradictions on occasion, as their utterances demonstrate; so any adequate logic of belief or analysis of the concept of belief must accommodate this fact." But any attempt to *legitimize* human fallibility in a theory of belief by fixing a permissible level of error would be like adding one more rule to chess: an Official Tolerance Rule to the effect that any game of chess containing no more than k moves that are illegal relative to the other rules of the game is a legal game of chess. Suppose we discovered that, in a particular large population of poor chess-players, each game on average contained three illegal moves undetected by either opponent. Would we claim that these people *actually* play a different game from ours, a game with an Official Tolerance Rule with k fixed at 3? This would be to confuse the norm they follow with what gets by in their world. We could claim in a similar vein that people *actually* believe, say, all synonymous or

intentionally isomorphic consequences of their beliefs, but not all their logical consequences, but of course the occasions when a man resists assenting to a logical consequence of some avowal of his are unstable cases; he comes in for criticism and cannot appeal in his own defense to any canon absolving him from believing nonsynonymous consequences. If one wants to get away from norms and predict and and explain the "actual, empirical" behavior of the poor chess-players, one stops talking of their *chess moves* and starts talking of their proclivities to move pieces of wood or ivory about on checkered boards; if one wants to predict and explain the "actual, empirical" behavior of believers, one must similarly cease talking of belief, and descend to the design stance or physical stance for one's account.

The concept of an intentional system explicated in these pages is made to bear a heavy load. It has been used here to form a bridge connecting the intentional domain (which includes our "common-sense" world of persons and actions, game theory, and the "neural signals" of the biologist) to the non-intentional domain of the physical sciences. That is a lot to expect of one concept, but nothing less than Brentano himself expected when, in a day of less fragmented science, he proposed intentionality as the mark that sunders the universe in the most fundamental way: dividing the mental from the physical.