BEHAVIORAL AND BRAIN SCIENCES (1988) 11, 495–546
*Printed in the United States of America*

# Précis of *The Intentional Stance*

**Daniel C. Dennett**
*Center for Cognitive Studies, Tufts University, Medford, Mass. 02155*

**Abstract:** The intentional stance is the strategy of prediction and explanation that attributes beliefs, desires, and other "intentional" states to systems – living and nonliving – and predicts future behavior from what it would be rational for an agent to do, given those beliefs and desires. Any system whose performance can be thus predicted and explained is an *intentional system*, whatever its innards. The strategy of treating parts of the world as intentional systems is the foundation of "folk psychology," but is also exploited (and is virtually unavoidable) in artificial intelligence and cognitive science more generally, as well as in evolutionary theory. An analysis of the role of the intentional stance and its presuppositions supports a naturalistic theory of mental states and events, their *content* or *intentionality*, and the relation between "mentalistic" levels of explanation and neurophysiological or mechanistic levels of explanation. As such, the analysis of the intentional stance grounds a theory of the mind and its relation to the body.

**Keywords:** artificial intelligence; belief attribution; cognition; evolution; explanation; intentionality; materialism; mental representation; mental states; rationality

*The Intentional Stance* gathers together, reorganizes, and extends my essays on intentional systems (and the intentional stance from which they are visible), presenting the first comprehensive and uniform expression of my theory, together with detailed replies to the criticisms it has provoked. After an introductory chapter, six previously published essays are reprinted, detailing various parts and implications of the theory of intentional systems, each followed by an essay of "Reflections," reconsidering and extending the claims and arguments of the earlier work. The next two chapters further develop the theory, and the final chapter traces the history of philosophical theorizing on these topics during the last thirty years, placing my view in the context of the views of Quine, Sellars, Chisholm, Putnam, Davidson, Fodor, Bennett, Stich, and others, and showing that there has in fact been a growing consensus about a number of fundamental points. Philosophy does make progress, and the progress discerned is on issues of concern to theorists in other disciplines working on the nature of the mind. The chapters and reflections are here summarized in the order in which they appear.

## 1. Setting off on the right foot

Although some theorists would rather not talk about the mind – would prefer to avoid all "mentalistic" theorizing – the various abstemious alternatives in the social sciences and biology have now shown their futility, so there is a need for an account – a philosopher's account – of the presuppositions and implications of mentalistic talk.

Not everything has a mind, but the dividing line is obscure: Do insects have minds? Do any nonhuman animals? Could a robot have a mind? The "common sense" answers to these questions are unreliable; even the most compelling intuitions may be swept aside by a successful, broadly supported theory, so there is little to be gained from a search for indubitable first principles. One must start somewhere, however, and my tactical choice is to begin with the objective, materialistic, third-person world of the physical sciences, and see what can be made from that perspective of the traditional (and deeply intuitive) notions of mind.

Others would insist that this is already to make a fundamental error – an error of scientism, roughly. Nagel (1986), for instance, has claimed that anyone who starts with the objective, third-person point of view is doomed to leave out or distort something essential about our minds, our subjective experience. There is no way to join debate between these two perspectives; Nagel's arguments beg the question against mine by accepting as simply obvious and in need of no support various intuitions I am prepared to dismiss from my theoretical vantage point. In the end, since these are tactical choices, the proof must be in the results: Does my third-person perspective fail to do justice to the phenomena (as we think we understand them), or does it provide explanations not only for the primary phenomena, but also for the presence and potency of the contrary intuitions that it dismisses?

Starting, then, from the third-person point of view, what first strikes us is that the predictive strategies and methods that work relatively well when the topic is germination or magnetism fail to yield much when applied to the behavior of people and other higher animals. For these phenomena, we all rely on what may be called *folk psychology*, which can best be understood in a parallel with *folk physics*, the everyday, largely subliminal lore we rely on to guide our expectations about spilled liquids, garden swings, and other nonliving macroscopic objects and events. It may well be that parts of both folk physics and folk psychology are innate, or at least aided by innate perceptual or dispositional biases, but much of both must be learned from experience. The expectations generated by these folk theories are for the most part effortless, uniform, and reliable. When a blind person fails to react to something right before his eyes, for instance, this can startle us, so compelling is our normal

© 1988 Cambridge University Press 0140-525X/88 $5.00+.00

expectation that people come to believe the truth about what is happening right in front of their eyes.

The intuitions of folk physics are not all to be trusted; gyroscopes and siphons, for instance, are counterintuitive. Similarly, the intuitions of folk psychology should not be deemed immune to revision or dismissal by theoretical advances, no matter how obvious they may now seem to us. Until recently, however, philosophers (and other theorists) have tended to endow our everyday intuitions about our minds with an implausible certainty or incontrovertibility. This invulnerable status must be denied to the putative truths of folk psychology, while at the same time we acknowledge the very great powers of organization and expectation we derive from it. The task that lies ahead is to describe accurately and then explain the power of folk psychology.

## 2. True believers: The intentional strategy and why it works

Sometimes attributions of belief appear entirely objective and unproblematic, and sometimes they appear beset with subjectivity and infected with cultural relativism. Catering to these two families of cases are two apparently antithetical theoretical options: *realism*, the view that beliefs are objective things in the head which could be discovered and whose identities could be confirmed, in principle, by physiological psychology; and *interpretationism*, the view that attributing a belief is a highly relativistic undertaking, more like asserting that someone is immoral, or has style, or talent, or would make a good wife – "it all depends on what you're interested in."

It is a common mistake to see these alternatives as mutually exclusive and exhaustive. Warding off the enthusiastic proponents of this mistake is a recurrent task in the book. My thesis is that while belief is a perfectly objective phenomenon (which apparently makes me a realist), it can be discerned only from the point of view of someone who adopts a certain predictive strategy, *the intentional stance* (which apparently makes me an interpretationist). In order to demonstrate this stance-dependent realism I first describe the stance, and then argue that any object – whatever its innards – that is reliably and voluminously predictable from the stance is in the fullest sense of the word a believer. *What it is* to be a true believer, to have beliefs and desires, is to be an *intentional system. How the stance works:* Suppose astrology actually worked. Then we would have the dual task of first describing how the *astrological stance* worked in action (predicting the futures of astrological systems) and then explaining why the astrological stance or strategy worked when it did. Other strategies of prediction have better credentials. In the *physical stance*, one predicts the behavior of a physical system (of stars, tides, volcanoes, DNA molecules) by exploiting information about the physical constitution of the system, and the laws of physics. (Laplace's omniscient calculator is the ultimate adopter of the physical stance.) In the *design stance*, one predicts the behavior of a system by assuming that it has a certain design (is composed of elements with functions) and that it will behave as it is designed to behave under various circumstances. The design stance can often safely ignore details of the physical implementation of the various imputed functions; there are many levels of design stance, from highly detailed to highly abstract. (Predicting the behavior of a computer from an analysis of the source code of its program is more abstract than predicting its behavior from an analysis of the states of its flip-flops, but both are design stance predictions dependent on assumptions that every element, however physically implemented, will perform its proper function.)

Finally, there is the *intentional stance*, in which one treats the system whose behavior is to be predicted as a rational agent; one attributes to the system the beliefs and desires it ought to have, given its place in the world and its purpose, and then predicts that it will act to further its goals in the light of its beliefs.

Generally, the beliefs any system ought to have are true beliefs about its circumstances, and the desires any system ought to have are those that directly or instrumentally aim to secure whatever the system needs to preserve itself and further any other projects it has. We all strive to believe the True and desire the Good, but the rules for belief and desire attribution must allow for the deflective effects of engineering shortcuts, design approximations, uncorrected malfunctions, and in particular the distorting effects of the *verbal expression* of belief and desire by those intentional systems capable of language.

The intentional stance is ubiquitous; we all adopt it every day to make sense of the behavior of our conspecifics, but also that of other animals, some artifacts (e.g., chess-playing computers), and even plants, on occasion. The definition of an intentional system is so forgiving that some have wondered whether it is vacuous. Consider a lectern. What prevents its being construed as an intentional system with the belief that it is currently located at the center of the civilized world and the desire to remain at that center? Given that belief and desire, it is rational for it to stay put; it does, so apparently it is predictable from the intentional stance, and hence it is an intentional system. If it is, everything is. But the lectern is disqualified because, first, there is no predictive leverage gained by adopting the intentional stance, and second, there is no principle governing the ad hoc attribution of the belief and desire.

The power of the intentional strategy is often concealed by theorists' concentration on cases which yield dubious results. Consider predicting moves in a chess game. What makes chess interesting is the *un*predictability of one's opponent's moves, it seems, but in fact even when the intentional stance fails to distinguish a single move as best (and hence predicted), it can narrow down the set of legal (and hence equipossible) moves to a typically much smaller subset of relatively likely moves, a large predictive edge not otherwise readily obtainable. Similarly, although the intentional stance does not permit such fine-grained predictions as the exact purchase and sell decisions of stock traders, or the exact sequence of words a politician will utter, it can yield reliable and high-probability expectations regarding more generic action types: The trader *will not buy utilities today* and the politician *will side with the unions against his party*. It is this neutrality with regard to fine-grained details of implementation that permits the intentional stance to be exploited in complex cases involving chaining predictions.

The example of the Martian predictor using the physical stance in a prediction contest with an Earthling using

the intentional stance shows that the objective pattern viewed from the intentional stance is indiscernible to the Martian, who might be able to make laborious and accurate predictions, à la Laplace, but would have no explanation of how the Earthling was caused to make the same true predictions without access to the same microcausal information. The objective presence of the pattern relied on by the Earthling (with whatever imperfections) does not rule out the presence of other patterns, each discernible from a different stance.

The example of the thermostat shows that it is theoretically perspicuous to include even the simplest homeostats as intentional systems, since there are no sharp discontinuities between them, lower animals, and fully fledged (human) believers. In particular, it gets matters backwards to attempt to make a divide between intentional systems that harbor internal (mental) representations and those that do not: It is not that we attribute (or should attribute) beliefs and desires only to things in which we find internal representations, but rather that when we discover some object for which the intentional strategy works, we endeavor to interpret some of its internal states or processes as internal representations. What makes some internal feature of a thing a representation could only be its role in regulating the behavior of an intentional system.

Finally, *why* does the intentional strategy work? First, because we are close enough approximations of optimal cognitive design (i.e., rationality). This says nothing, however, about what the ultimately mechanical details of that design are, which is an independent question. We should not jump to the conclusion that the internal machinery of an intentional system and the strategy that predicts its behavior *coincide* – that is, we should not conclude that the language-of-thought hypothesis is true (a topic that is discussed further in later chapters).

In the "Reflections," the example of Conway's "Life world" (Gardiner 1970) illustrates and reinforces the claim that my view is both a mild form of "realism" and stance-dependent. There are patterns of "behavior" definable in the state transitions of configuration in the Life world that are describable only from the intentional stance. Further, when there are imperfections in these intentional stance patterns, there are no "deeper facts" to resolve the outstanding questions of belief attribution. This claim is related to Quine's (1960) thesis of the indeterminacy of radical translation and to Parfit's (1984) account of "empty questions" about personal identity.

## 3. Three kinds of intentional psychology

This chapter extends the analysis of the role of the intentional stance in folk psychology and in cognitive psychology.

What do we have in common if we both believe that cats eat fish? Such a question can call for two different kinds of answers: a conceptual answer (giving the essence, in effect) or a causal answer (the "reductive," "scientific" answer). What do all magnets have in common? The conceptual answer is that all magnets attract iron; the causal answer goes on to give the microphysical account of what underlies that capacity. Some philosophers, in their eagerness to ground a proper scientific theory of belief, have confounded the parallel questions about what believers have in common. To answer the first question, we must analyze the concept of belief in its source context: folk psychology. What we discover is that folk psychology is best seen not as a sketch of internal processes, but as an idealized, abstract, instrumentalistic calculus of prediction.

At the heart of folk psychology is the intentional stance. There is some evidence that we are not as rational as we have thought we are, but however rational we are, the myth of our rational agenthood structures and organizes our attributions of belief and desire. [See also Cohen: "Can Human Irrationality Be Experimentally Demonstrated? *BBS* 4(3) 1981.] So folk psychology is idealized in that it produces its predictions by calculating in a normative system. It is abstract in that the beliefs and desires it attributes need not be presumed to be independent, distinguishable intervening states of an internal behavior-causing system. It is instrumentalistic in a way the most ardent realist should permit: People have beliefs in the same way they and all other physical objects have centers of gravity. Reichenbach (1938) distinguished two sorts of referents for theoretical terms: *illata* – posited theoretical entities – and *abstracta* – calculation-bound entities or logical constructs (such as the center of gravity or the equator). Beliefs and desires are properly seen as *abstracta*. This view contrasts most sharply with views such as Fodor's (1975; 1981; 1987).

The example of how various people come to learn about Jacques shooting his uncle in Trafalgar Square shows that what is in common in such cases need not be anything describable at any less abstract and instrumentalistic level than the intentional stance. Other examples show that beliefs make poor *illata* in any case – rather like Anaxagoras's "seeds," they have problematic identity conditions and are prone to create infinite regresses. But we have reason to want to cite beliefs in causal explanations (as many philosophers have insisted), and we can do this without their being viewed as *illata*, but only by doing some violence to folk psychology.

The ordinary notion of belief is pulled in two directions. If we want to have good theoretical entities, good *illata*, or good logical constructs, good *abstracta*, we will have to jettison some of the ordinary freight of the concepts of belief and desire. So I propose a divorce of two theories from their source in folk psychology: (1) pure *intentional system theory* (idealizing, holistic, abstract, instrumentalistic) and (2) *sub-personal cognitive psychology* (a concrete, microlevel theory of the realization of the intentional systems whose behavior is specified by the first theory).

*Intentional system theory* is thus a competence theory. Like kinematics, and unlike dynamics, it proceeds via assumptions of perfection (no friction, no stretching or bending), and is useful in providing an abstract and relatively noncommittal set of specs for the design of the concrete mechanism. *Subpersonal cognitive psychology* is then a performance theory. In particular, its task is to show how brains, which as physical mechanisms can only be syntactic engines, nevertheless succeed in implementing (or better: approximating) the semantic engines specified by intentional system theory. This task cannot be executed within the strictures of what Putnam (1975) has called methodological solipsism.

Thus the best way to view the relation of psychology and its phenomena to the rest of science is not, as philosophical tradition would have it, to try to find a reducing vocabulary in a lower, microtheoretical science (such as neuroscience), so that we can complete the formulae:

$x$ believes that $p$ iff $Px$

where "$P$" is some predicate of neuroscience or other physical science. The best way to view the relation is as analogous to the relation of a Turing machine specification to the underlying concrete mechanism. One cannot say what two realizations of Turing machine $k$ have in common in any vocabulary "lower" than Turing machine talk, but one can have an engineering theory of how Turing machines can be realized in mechanisms.

Thus the claim that all mental phenomena exhibit intentionality, which is traditionally taken as an irreducibility thesis (Brentano's Thesis), can be viewed, through the lens of the intentional stance, as a sort of reductionist thesis, in exactly the same way Church's Thesis is viewed as reductionist: Church's Thesis reduces the irremediably intuitive and informal concept of effective procedure to the crisply defined concept of Turing-computability. Analogously, my claim that every mental phenomenon alluded to in folk psychology is *intentional-system-characterizable* would reduce a domain whose boundaries are fixed (at best) by mutual intuition to a clearly defined domain of entities whose principles of organization are relatively formal and systematic, and entirely general.

In the "Reflections," after brief comments on Searle's (1980; 1982) version of my claim that "you can't get the semantics from the syntax," operationalism, and "core beliefs," I address the problems occasioned by my calling my view "instrumentalist," which has proven in some regards a misnomer. One can consider the intentional stance to be a special variety of design stance, positing homunculi pro tempore. This makes intentional system theory literally a black box theory, but it is not thereby a behaviorist theory. In fact, it is consonant with Marr's (1982) strategic vision of three levels in cognitive theory: the computational level (which is roughly equivalent to the intentional stance and to Newell's (1982) "knowledge level"), the algorithmic level, and the hardware level.

The role of the intentional stance in specifying a competence is illustrated by the example of the Newfie-joke-getter, which requires a mechanism-independent, but still counterfactual-supporting, description from the intentional stance. The example of the hand calculator illustrates the way an idealized competence model yields to a partial specification of a performance model as imperfections and shortcuts are discovered. Finally, in response to various objections, there is a discussion of the use of the intentional stance to predict chess moves, and particularly to predict errors by one's opponent.

## 4. Making sense of ourselves

Nominally a reply to Stich (1981) this chapter concentrates on the rationality assumption of the intentional stance, which has been the target of considerable criticism.

First, according to Stich (and others), the intentional stance cannot account for (predict, explain) mistakes or failures of rationality. This claim is examined via Stich's example (1981) of the lemonade seller who gives the wrong change; I claim that the indeterminacy of attribution encountered by the intentional stance is real and ineliminable, even in such mundane cases.

Second, Stich (and others) have criticized me for not giving a specific definition of rationality, but this is a misguided objection; the concept of rationality exploited in the intentional stance is, and should be, the pretheoretical concept of cognitive excellence (whatever that comes to in particular circumstances), rather than any one fixed canon such as deductive closure, and whatever-evolution-has-given-us. We must allow for the discovery that deductive consistency or completeness is not always a cognitive virtue, and that evolution does not always provide us with the best, most rational equipment or habits of thought.

Third, Stich contrasts my theory of belief attribution (roughly: x believes what x ought to believe) with his own (roughly; x believes what I would believe in x's circumstances), but this contrast is shown to be evanescent; it is not clear that these methods would ever yield different predictions, given what Stich must acknowledge in the specification of circumstances.

In the "Reflections," the example of the lemonade seller is explored in more depth to show that the advance of subpersonal cognitive psychology would not eliminate the indeterminacy of the intentional stance, would not be capable of rendering objective verdicts about those belief attributions that Stich supposes would have to be true or false.

A section on "frog psychology" shows how two well-known positions tug at mine in opposite directions, while agreeing about something I reject. Critics on both sides see a wide gulf between frogs (for instance) and human beings: Fodor (1975; 1986), Dretske (1985; 1986) and other "Realists" insist that we human beings, unlike frogs, *really* have beliefs and desires; Stich (1983), Churchland & Churchland (1981) and other "eliminativists" accept the implied contrast between us and frogs, and with it they accept the Realists' vision of what *really having a belief* would be. Strictly speaking, they agree, frogs have no beliefs; but, strictly speaking, neither do we! There are no such things as beliefs. I agree that if beliefs had to be what the Realists think they are, there wouldn't be any beliefs, for frogs or any of us. I am not tempted, as both groups of critics are, by the contrast between us and frogs.

The illusions of Realism are engendered in the literature by a misplaced concentration on verbalized beliefs, or what I call (in a special technical sense) *opinions*. An examination of Quine's (1956) famous example of Ralph believing Ortcutt to be a spy exhibits the misbegotten reliance on linguistic forms, which lends spurious support to the language-of-thought hypothesis, the view that is analogous to the idea that a head cold is composed of a large set of sneezes, some of which escape. A confusion between thinking (in the sense of thinking specific, verbalized, or partly verbalized thoughts) and believing is one of the prime sources of this illusion.

## 5. Beyond belief

This chapter is a demolitional tour of that philosophical black hole: propositional attitude psychology. CAUTION: Reading this chapter, and the literature discussed in it, may be hazardous to your health.

Philosophical tradition views beliefs as *propositional attitudes*, but this tradition has less secure foundations than its orthodoxy would suggest. In the first section, the traditional doctrines of propositional attitudes are presented, and the persistent difficulties noted. It is standard to suppose that belief attributions and their kin are analyzable into three variable components:

*x* [subject] believes [attitude] that *p* [proposition]

but there has been no uniformity on the nature of the third component, propositions. There are three distinct views defended in the literature. Propositions are:

(1) like sentences (composed of something like symbols, with a syntax)

(2) sets of possible worlds

(3) ordered collections of objects and properties in the world.

This disunity is due to the fact that, by tradition, propositions are supposed to meet three conditions. Propositions are:

(a) truth-value bearers

(b) intensions (in Carnap's sense, as extension-determiners)

(c) "graspable" by the mind

Nothing can meet all three of these conditions simultaneously, as a number of arguments (by Kaplan 1980, Perry 1977; 1979, Putnam 1975, Stich 1983, and others) purport to show.

The more or less standard retreat in the face of these arguments is to what I call *sentential attitude psychology*, one version or another of the language-of-thought hypothesis. Four routes to this destination are described:

(1) Sentences in the head are what are physically "grasped" when a proposition (an abstract object) is entertained.

(2) Beliefs about dogs and cats must be composed of something – not dogs and cats, of course, but *symbols* for dogs and cats; these symbols must be arranged in sentence-like structures.

(3) Sentences in the head can account for the phenomenon of referential opacity[1] in a straightforward way.

(4) Following Kaplan (1980) a distinction between content and character can be drawn, but then sentences in the head are needed to be the entities with character.

One hopes, via this strategy, to specify the organismic contribution to the determination of an organism's propositional attitudes, but the language-of-thought hypothesis runs into difficulties. First, it distinguishes too finegrainedly between psychological states; two individuals with somewhat different languages of thought could never be counted as having the same belief, for instance. Second, it presupposes that one could in principle discover the "syntax" of a language of thought (the system governing the composition of representations from basic elements or terms) prior to discovering any of its "semantic" properties (what the representations were *about*), and this presupposition is shown to be dubious on several counts. Third, and most important, it presupposes that semantic contributions to a system can always be cast in "verbal" form, as a message or premise added to the system, but there is good reason to believe that some important varieties of contentful effect cannot be captured as things "said", in effect, to the system.

What is needed is an intermediate specification of psychological traits, independent of any hypotheses about the medium of internal representation, and at the same time independent of the particular facts of the external environment of the organism. This can be provided by what I call *notional attitude psychology*. Notional attitudes are indirectly characterizable via a theorist's fiction – the "notional world" which is the environment for which the organism, as currently constituted, is ideally suited. Various versions of this idea of notional worlds are seen to operate in disparate schools of thought: Husserlian (1931) Phenomenology, artificial intelligence, Quine's (1960) doctrine of indeterminacy of radical translation, and the interpretation of fiction by literary theorists.

One of the implications of notional attitude psychology that many philosophers find strongly counterintuitive can be expressed in terms of Putnam's (1975) Twin Earth thought experiment: If I and my Doppelgänger switch places, I would immediately have beliefs (and other propositional attitudes) about the things on Twin Earth – in advance of any particular causal commerce with them, contrary to the various versions of the causal theory of reference.

The issues raised by such wildly unrealistic philosophical thought experiments are better explored in a mundane, if unlikely, scenario: the practical joke described in the Ballad of Shakey's Pizza Parlor. Tom, while eating a pizza in one Shakey's Pizza Parlor, is drugged by his friends and moved to another, practically indistinguishable Shakey's Pizza Parlor some miles away, where he awakes, unaware of the switch. He makes a wager about the existence of his initials carved in the men's room door, and loses the wager, of course. But which door did he "have in mind"? The difficulties encountered in trying to give a proper description of Tom's psychological state at the time of the wager expose the problem with several technical concepts traditionally honored by philosophers. First, Tom must be seen as having had a dual set of *propositional attitudes* (double entry bookkeeping, one might say) in spite of the presumably unitary nature of his internal psychological state. And worse, even if we grant the dual set of propositional attitudes, there is an unresolved problem about how to impose another technical distinction on the case: the putative distinction between *de re* and *de dicto* propositional attitudes.

This distinction has so far defied uncontroversial definition, in spite of much attention, but it can be illustrated by an example. Bill believes that the captain of the Soviet ice hockey team (whoever he is) is a man; Bill also believes that his own father is a man. The latter belief, arising as it does out of a relatively intimate experiential acquaintance, is said to be *de re* (literally: of or about the thing) whereas the former belief, arising presumably as a highly probable inference from some general beliefs and not from any direct experience of that stalwart Russian is said to be *de dicto* (literally: of or about the *dictum* or proposition). Intuitively compelling as many examples make this distinction appear, it has engendered a host of problems and should not be assumed to be either safe or salvageable.

The history of the distinction is traced from the work of Chisholm (1956; 1966) and Quine (1956; 1960) and others, and it is shown that there has been a confounding of several independent factors. There is on the one hand a distinction between relational and notional attribution-claims, and on the other hand, between general and specific beliefs. Once these are properly distinguished, the various attempts to isolate criteria for *de re* belief can be seen to fail: Kaplan's (1968) vividness, the lifting of substitution-restrictions, the availability of direct ostension or demonstrative reference, the putative failure of definite descriptions to capture a putative special mode of reference.

The conclusion drawn is that although it is possible to isolate and individuate a subset of beliefs meeting one or another causal or metaphysical condition, it will not be a theoretically interesting subset from the point of view of psychology. Alternatively, one can draw a distinction within the confines of notional attitude psychology that is psychologically perspicuous, but which fails to line up with the tradition on *de re* and *de dicto* – so much the worse for that tradition.

Viewed from another perspective, the conclusion is that we must abandon what Evans (1980) calls Russell's Principle: *It is not possible to make a judgment about an object without knowing what object you are making a judgment about.* Once this is done, one can discern a host of interesting, but largely ignored, distinctions in need of further analysis: different ways of thinking of things; the difference between episodic thinking and believing; the difference between explicit and virtual representation; the difference between linguistically infected beliefs (opinions) and what we might call animal beliefs; the difference between artifactual or transparently notional objects and other notional objects. Once good philosophical accounts of these are at hand, the artifactual (and dismissible) status of the doctrine of *de re* and *de dicto* will be clearer.

In the "Reflections," I summarize and expand on the four major conclusions of the chapter:

(1) At this time there is no stable, received view of propositions or propositional attitudes on which one can rely.

(2) The well-trodden path to "language-of-thought psychology" is a route to fantasyland.

(3) Russell's Principle must be abandoned.

(4) There is no stable, coherent view of the so-called *de re/de dicto* distinction.

This leaves something of a vacuum, since if there is one thing cognitive science seems to need from philosophers, it is a theory of propositions to provide the foundation for the ubiquitous use of propositional attitude talk. After all, statements of propositional attitude are central to the intentional stance. One may continue to rely on such talk, I argue, so long as one gives up a certain misguided hope of realism with regard to propositions. Propositions are abstract objects, but they are *more like dollars than numbers.* There are no real, natural, universal units of either economic value or semantic information.

## 6. Styles of mental representation

In *The Concept of Mind* (1949), Ryle attacked a vision he called the intellectualist myth, which bears a strong resemblance to the currently applauded cognitivist program. Contemporary cognitivists have been right to shrug off much of Ryle's celebrated attack as misguided, but there remains a central Rylean criticism that has been underappreciated. Ryle saw that at bottom all cognition (all knowledge, belief, thinking, inference, etc.) had to be grounded in varieties of know-how that are only tacit – present not in virtue of any explicitly represented rules. The open empirical question is how much explicit representation rests on this tacit foundation. The issue has been confused in the literature because of a failure to distinguish between several different styles of representation.

I distinguish *explicit* representation from *implicit* representation as follows: Something is represented implicitly if it is *implied* by something that is represented explicitly by the system. Some, but not all, that is implicitly represented by a system may be *potentially explicitly* represented. So implicit representation depends on explicit representation. But explicit representation depends (as Ryle saw) on what we may call *tacit* representation. Otherwise we would be faced with an infinite regress of explicit representations being "understood" in virtue of the system's consultation of further explicit rules, and so forth. These various distinctions are all independent of any distinction between *conscious* and *unconscious* representation. For instance, Chomsky (1980) has claimed that the rules of grammar are explicitly but entirely unconsciously represented in speakers. [See also Chomsky: "Rules and Representations" *BBS* 3(1) 1980; and Stabler: "How Are Grammars Represented" *BBS* 6(3) 1983.]

This idea of explicit unconscious representation of rules is ubiquitous in theoretical discussions in cognitive science, but it is neither as clear nor as inevitable as its orthodoxy would suggest. The idea can be clarified, and alternatives discerned, by considering a parallel distinction between three relations bridge players might have to the rule of thumb, "third hand high." The first, dim-witted player has memorized the rule, and consciously and deliberately consults it during play, first determining by a serial inspection that its conditions are met and then letting his action be dictated by it; the second, "intuitive" player typically plays third hand high (when this is appropriate) but has never entertained, considered, heard, or formulated the rule explicitly; the third player is both intuitively and explicitly aware of the rule, but unlike the first, understands the rule and when and why its use is appropriate. The first player, the stupid, uncomprehending rule-follower, is the best model for the subpersonal explicit rule-following of cognitive science, since the threatened homuncular regress must be terminated by the cumulative diminution to zero of "understanding" in the subsystems.

But we know a priori that any such regress must be terminated by subsystems with merely tacitly represented know-how, so the empirical question that remains is whether the design principle of unconscious-but-explicit-rule-following ever gets used. Could Ryle turn out to be right about the "size" of the largest merely tacit knowers: whole people? Could virtually all the backstage know-how be merely tacit in the organization of the system?

This question is explored via the example of the hand

calculator, which is designed to honor the rules of arithmetic, but explicitly represents no general arithmetical rules or facts of arithmetic. The design process that creates such things as hand calculators and automatic elevators begins, no doubt, with an explicit representation of the rules to be obeyed, but terminates, typically, in a "hard-wired" design which guarantees the desired behavior without in any way representing the rules or principles that specify that behavior. Moreover, such purely tacit systems need not be fixed or hard-wired. An elevator, for instance, can operate according to one set of rules on weekdays and a different set of rules on the weekends without either set of rules being explicitly represented. All that is needed is a clock that will switch control between two different tacit systems.

This transient tacit representation is a design option of open-ended power. But as the number of possible different states (each with its distinctive set of tacitly represented rules) grows, this profligacy demands that the design rely in one way or another on economies achieved via multiple use of resources. Changing state by editing rather than replacing leads to the creation of functional roles that are analogous to variables with different substitution instances – and we are headed back in the direction of a language of thought. But there is a difference: The "semantics" of such a system are entirely internal, "referring", if you like, to memory addresses, internal operations, and so forth. The external, or real-world, semantics of a state in such a system is still determinable only by the globally defined role of the state.

In the "Reflections," the language-of-thought hypothesis is reconsidered. After more than ten years of being at the center of attention in cognitive science, it has been neither vindicated nor utterly discredited. It would have been vindicated if sententialist models of central thinking had been developed and partially confirmed, but instead, such models that have been proposed appear hopelessly brittle, inefficient, and unversatile. As Fodor (1983) himself acknowledges, the language-of-thought hypothesis has provided us with no good account of nonperipheral cognition.

The idea of a language of thought would have been entirely discredited if a clear alternative to it had been formulated and confirmed, but although connectionism (McClelland & Rumelhart 1986) has been offered in this role, it has not yet been shown to live up to its promise.

Connectionist networks are characterized, typically, by the following factors:

(1) distributed memory and processing
(2) no central control or highest level executive
(3) no complex message-passing
(4) a reliance on statistical properties of ensembles
(5) relatively mindless making and unmaking of connections or solutions

Perhaps most important, from the point of view of the philosophical theory of meaning, is the fact that the level at which connectionist models are computational is not the level of their real-world, or external, semantics. This is what most sharply distinguishes them as an alternative to the language-of-thought hypothesis. But we have yet to see just what can be constructed using these new and surprisingly powerful connectionist fabrics. At present all the demonstrations are made possible by (over)simplifying assumptions, and serious grounds for skepticism abound.

But suppose connectionism triumphed; what would be the implications for folk psychology? According to the eliminativists (Stich 1983; Churchland 1981), the triumph of connectionism would show that there are no such things as beliefs at all. Stich would replace folk psychology with his "syntactic psychology" whereas Churchland would do just the opposite: cling to the contentfulness or intentionality of inner something-or-others but drop the presupposition that these inner vehicles of meaning behaved, syntactically, the way Realists have supposed beliefs to behave. Belief dis-contented versus content dis-believed. There is something to be said for both options, but something more to be said for their combination: Stich, with his syntactic theory, can be recommending more or less the right way to go with opinions (linguistically infected states) whereas Churchland can be recommending the right course to take about animal belief (the internal, behavior-guiding states that are not "sentential" at all).

But I recommend yet another, less radical approach: The vindication of connectionism would not undercut the modest "instrumentalistic" objectivism of the intentional stance. Are there beliefs? Of course. We already know that there are robust, reliable patterns in which all behaviorally normal people participate – the patterns we traditionally describe in terms of belief and desire and the other terms of folk psychology. What spread around the world on July 20, 1969? The belief that a man had stepped on the moon. In no two people was the effect causally or syntactically the same, and yet the claim that they all had nothing in common is false and obviously so. There are indefinitely many ways of reliably distinguishing those who have the belief in question from those who have not, and there would be a high correlation among the methods. This is something science should not and could not turn its back on.

This completes the presentation of the basic theory of intentional systems and its defense against rival philosophical positions, and sets the stage for the application of these ideas to further domains.

## 7. Intentional systems in cognitive ethology: The "Panglossian paradigm" defended

Because this chapter first appeared in *BBS* (Dennett 1983a), it will not be summarized here. The following abstract appeared with the original article:

> Ethologists and others studying animal behavior in a "cognitive" spirit are in need of a descriptive language and method that are neither anachronistically bound by behaviorist scruples nor prematurely committed to *particular* "information-processing models." Just such an interim descriptive method can be found in *intentional system theory*. The use of intentional system theory is illustrated with the case of the apparently communicative behavior of vervet monkeys. A way of using the theory to generate data – including usable, testable "anecdotal" data – is sketched. The underlying assumptions of this approach can be seen to ally it directly with "adaptationist" theorizing in evolutionary biology, which has recently come under attack from

Stephen Gould and Richard Lewontin, who castigate it as the "Panglossian paradigm." Their arguments, which are strongly analogous to B. F. Skinner's arguments against "mentalism," point to certain pitfalls that attend the careless exercise of such "Panglossian" thinking (and rival varieties of thinking as well), but do not constitute a fundamental objection to either adaptationist theorizing or its cousin, intentional system theory.

In the "Reflections," I first give an account of my own later observation of Robert Seyfarth and Dorothy Cheney observing the vervet monkeys in Kenya. I think the most important thing I learned from actually watching the vervets is that they live in a world in which secrets are virtually impossible. So it is a rare occasion indeed when one vervet is in a position to learn something that it alone knows *and knows that it alone knows*, and the opportunity for such secrets is, I claim, a necessary condition for the development of a sophisticated (truly Gricean) capacity for communication. But even if this killjoy verdict is correct – the vervets are further from having a human language than one might have thought – the intentional stance is still the best strategy to adopt to describe (and eventually explain) their behavior.

My defense of the Panglossian paradigm against Gould and Lewontin (1979) in the original *BBS* target article struck some commentators as a digression, but in fact it introduced a central theme in my analysis of the intentional stance. The most important parallel I wished to draw was this: Psychologists can't do their work without the rationality assumption of the intentional stance, and biologists can't do their work without the optimality assumptions of adaptationist thinking. The suspicion aroused among both psychologists and biologists by the use of these assumptions suggests a common source: the underestimation of the role of functional interpretation (and its underlying assumptions) in these sciences. This came out clearly in Ghiselin's (1983) commentary. He recommended rejecting "such teleology altogether. Instead of asking, What is good? we ask, What has happened? The new question does everything we could expect the old one to do, and a lot more besides" (p. 363). But I show that this is an illusion exactly parallel to Skinner's (1964; 1971) familiar claim that the question, What is the history of reinforcement? is a vast improvement over What does this person believe, want, intend? We cannot hope to answer either sort of historical question with a "pure" investigation. [See also *BBS* special issue on the work of B. F. Skinner, *BBS* 7(4) 1984]

The use of optimality assumptions leads theorists in both fields to striking discoveries largely by yielding predictions that are (surprisingly) false. Kahneman and Tversky's (1983) discovery of a curious local irrationality in the way people tend to think about money provides a good illustration in psychology. It is only the contrast between what people do and what one would pretheoretically assume to be the rational course that opens up testable hypotheses about heretofore ignored design factors, hidden costs of cognition.

I think I have shown my alignment of ideologues to yield a harvest of insight, but several authors have championed roughly the opposite match-up: Gould and Lewontin should be placed not with Skinner but with Chomsky! They have a point: Chomsky's extreme nativism, tantamount to the denial that there is any such thing as learning, can be lined up with the extreme structuralism of some evolutionary theorists, tantamount to the denial that there is any such thing as adaptation. (This is not so much the *best* of all possible worlds; it is the *only* possible world.) As I argued in a commentary on Chomsky in *BBS* (1980), when asked how the innate, constraining structures got there, Chomsky can pass the buck to biology, but to whom can the "structuralist" evolutionary theorists pass the buck? What accounts for the constraining *Baupläne* relative to which adaptations are only so much fine tuning? One can say that the biosphere is just built that way: end of explanation. But one can always resist this and go on to ask, one more time, why? It is these "why" questions that are asked by both adaptationists and adopters of the intentional stance, and it is the puritanical distaste for this teleological and functional thinking that unites – in this regard – Skinner and Chomsky, Lewontin and Ghiselin.

So I stick to my guns:

(1) Adaptationist thinking in biology is precisely as unavoidable, as wise, as fruitful – and as risky – as mentalist thinking in psychology and cognitive science generally.

(2) Proper adaptationist thinking just *is* adopting the intentional stance in evolutionary thinking – uncovering the "free-floating rationales" of designs in nature.

The role of the intentional stance in interpreting evolutionary history is brought out more clearly by the exploration of a question that has typically been ignored by Darwinians. The theory of natural selection shows how every feature of the natural world *can* be the product of a blind, unforesightful, nonteleological process of differential reproduction, but it could not show that all features of the natural world *are* such products, since some of them aren't: The short legs of dachshunds and the thick skins of tomatoes (to say nothing of the more recent and spectacular products of genetic engineering) were the deliberately and foresightedly sought outcomes of conscious design processes. Now can such special cases be distinguished in retrospective analysis? Are there any telltale signs of artificial selection or conscious design? I argue that there are no such foolproof marks of either natural or artificial selection but that this is no "threat" to orthodox neo-Darwinian theory. We can see that the most extreme cases – suppose conscious designers "sign" their creations by including trademark messages or user's manuals in the junk DNA of their created species – reveal the kinship between psychology and biology most strikingly: Reading those messages, like the less radical, more familiar task of giving functional interpretations of the designs found in organisms, is always an exercise in radical interpretation.

## 8. Evolution, error, and intentionality

This chapter reveals a heretofore unrecognized division among theorists into two camps: those who believe in one form or another of *original intentionality* (the Bad Guys) and those who don't (the Good Guys). In the former camp belong Fodor, Searle (1980), Kripke (1982), Dretske (1985; 1986), Burge (1979; 1986), Chisholm (1976), Nagel (1974; 1986), Popper and Eccles (1977). In the latter:

myself, the Churchlands, Haugeland (1981), Millikan (1984), Rorty (1982), Stalnaker (1984), our distinguished predecessors in philosophy, Quine (1960) and Sellars (1954; 1963; 1974), and just about everyone in AI.

The doctrine of original intentionality is the claim that whereas some of our artifacts may have intentionality derived from us, we have original (or intrinsic) intentionality, utterly underived. Aristotle said that God is the Unmoved Mover, and this doctrine announces that we are Unmeant Meaners. The curious implications of this doctrine can be brought out via a series of thought experiments.

Consider an encyclopedia. It has derived intentionality. It contains information, but only insofar as it is a device designed and intended for our use. If we "automate" it on a computer, and equip it with a natural-language-processing front end so that it "answers" our questions it is still just a tool, and whatever meaning or aboutness we vest in it is still derived. A chess-playing computer has slightly more autonomy (since it is designed to try to defeat us), but still – according to this line of reasoning – since it is our tool or toy, its intentionality is just derived. The principle that emerges for those who believe in original intentionality is that no artifact, no computer or robot, could ever be a truly autonomous agent with the sort of original intentionality we enjoy.

See where this doctrine leads when we confront the problem of *error*, a central theme in recent theoretical explorations of meaning in cognitive science. Consider the device on vending machines that detects United States quarters (I shall call it a two-bitser). Not being foolproof, the two-bitser can be provoked into "errors" – accepting slugs and rejecting legal quarters. That a two-bitser will be reliably fooled by a particular kind of slug is just as much a matter of the laws of physics and the principles of its design as that it will reliably accept genuine quarters, but the former incidents count as errors only because of the intentions of the two-bitser's creators and designers. This relativity to the user's intentions is the hallmark of derived intentionality. And it follows that if we transport the two-bitser to Panama and use it to detect Panamanian quarter-balboas, what counts as an error will shift accordingly; when it accepts a United States quarter in Panama, that will count as an error, a "misperception" or misrepresentation. This shift is harmless, say the believers in original intentionality, since states of artifacts like the two-bitser don't *really* mean anything at all.

Now everyone agrees that moving the two-bitser to Panama does not change anything intrinsic about its internal states, but it does change the external context, and hence changes the (derived) meaning of those internal states. There are no "deeper facts" about what those states mean. I want to insist, however, that the same principle applies to us and all other cognitive creatures, and it is just here that we part company. To focus the disagreement, consider a parallel case, which echoes much of the recent discussion by philosophers. Suppose some human being, Jones, looks out the window and thereupon goes into the state of thinking he sees a horse (the example follows Fodor, 1987). According to believers in original or intrinsic intentionality, it is not just a matter of interpretation that this is the meaning of his internal state, as can be seen if we imagine the planet Twin Earth, where there are schmorses (which are well-nigh indistinguishable from horses) in place of horses. If we whisk Jones to Twin Earth and confront him with a schmorse, what state does he go into? Does he think he sees a horse or does he think he sees a schmorse? Anyone who thinks that, however hard it may be to determine exactly which state he is in, he is *really* in one or the other believes in original intentionality.

Another thought experiment provides grounds for abandoning the idea of original intentionality. Suppose you decided to build a robot to preserve your body (suspended in a cryogenic life-support capsule) for hundreds of years. Such a survival machine, being an artifact, would have only derived intentionality: Its raison d'être is to preserve you indefinitely, and the design of all its machinery is to be interpreted relative to that purpose. Anything that counts as a blunder or mistake or error or malfunction can only do so relative to this interpretation. But then the conclusion forced upon us is that our own intentionality is of the same sort, for, as Dawkins (1976) has put it, we (and all organisms) are survival machines designed to prolong the futures of our genes. Our interests as we conceive them and the interests of our genes may well diverge, even though were it not for our genes' interests, we would not exist. So our intentionality is derived from the intentionality of our "selfish" genes; *they* are the Unmeant Meaners, not us!

This vision provides a satisfying answer to the question of whence came our own intentionality, but it seems to leave us with a problem: Our genes, it seems, are at best only metaphorically intentional. How could the literal depend in this fashion on the metaphorical?

The answer lies in seeing that the process of natural selection is the source of all functional design, and hence, when considered from the intentional stance, the ultimate source of meaning. There is no representation ("mental" or otherwise) in the process of natural selection, but that process is nevertheless amenable to principled explanation from the intentional stance, in terms of the ("free-floating") rationales of the designs "selected."

There is nothing unorthodox about this interpretation of the theory of natural selection, but it is resisted rather subliminally by theorists who believe in original intentionality. Tracing the recent discussions of the problem of error (or the disjunction problem) by Dretske (1986), Fodor (1987), Burge (1986), and Kripke (1982) reveals that all of them are tempted by, but ultimately reject, the appeal to the theory of natural selection (what Fodor disarmingly calls "vulgar Darwinism"), because they dimly recognize that if they accept the Darwinian picture, they must abandon something they find irresistible: the intuition that, in our case, unlike that of artifacts (and perhaps, lower animals), there always are "deeper facts" about what our innermost thoughts *really* mean – and (at least some of them also hold) we each have "privileged access" to these deeper facts.

This clash of intuitions is not restricted to philosophers. We can see the uneasiness, for instance, in the microbiologists' frank use on the one hand of intentional terms to describe and explain the activity of macromolecules, while feeling the puritanical urge to renounce all talk of function and purpose on the other. Unless there can be perfect *determinacy* of function, they suspect, they had better not appeal to function at all. In Gould's (1980)

recent discussions (e.g., of the panda's thumb), we find the same resistance to the moral that was welcome in the case of the two-bitser: Just as it can change its function, and hence the meaning of its internal states, by changing contexts, so the panda's wrist bone can be pressed into service as a thumb. But is there no answer to the question of what it *really* is for, what it *really* means? Gould sees a "paradox" here, for he refuses to accept the adaptationist (and intentional stance) methodology of using optimality assumptions ("reading function off prowess") as the touchstone of functional interpretation. He still hankers, like the believers in original intentionality, for some deeper facts, but there are none. There is no ultimate user's manual in which the *real* functions and *real* meanings of biological artifacts are officially represented. There is no more bedrock for what we might call original functionality than there is for its cognitivistic scion, original intentionality. You can't have realism about meanings without realism about functions.

## 9. Fast thinking

The ever-compelling idea of intrinsic or original intentionality has one final prop holding it up, which is removed in this chapter. Searle's (1980) Chinese-room argument originally appeared in *BBS* and has provoked much debate. There is no point in reviewing one more time the story of the Chinese room and its competing diagnoses, since, as Searle (forthcoming) himself acknowledges, the story is not itself the argument, but just an intuition pump designed to "remind us" of the obvious truth of the following argument:

1. Programs are purely formal (i.e., syntactical).
2. Syntax is neither equivalent to nor sufficient by itself for semantics.
3. Minds have mental contents (i.e., semantic contents).

Conclusion: Having a program – any program at all – is neither sufficient for nor equivalent to having a mind.

Searle challenges his opponents to show explicitly what is wrong with this argument and I do just that, concentrating on the conclusion, which, for all its apparent simplicity and straightforwardness, is subtly ambiguous. (I eventually show that all three premises contain errors, but I have learned from my long exposure to the debate about the Chinese room that those who side with Searle are not really interested in the argument; they are so sure the conclusion is correct that dispute about Searle's path to it strikes them as mere academic caviling.) My diagnosis of this extreme and intensely felt confidence in Searle's conclusion is that it is due partly to the fervent desire that it be true, and partly to an illusion: They are mistaking Searle's conclusion for a much more defensible near neighbor. Compare the following claims:

(S) No computer program by itself could ever be sufficient to produce what an organic human brain, with its particular causal powers, demonstrably can produce: mental phenomena with intentional content.

(D) There is no way an electronic digital computer could be programmed so that it could produce what an organic human brain, with its particular causal powers, demonstrably can produce: control of the swift, intelligent, intentional activity exhibited by normal human beings.

As the initials suggest, Searle endorses (S) as a version of his conclusion, whereas I offer an argument for (D). The two may seem at first to be equivalent, but when we see how my argument for (D) runs, we see that (S), given what Searle means by it, is incoherent. I am not convinced that (D) is true, but I take it to be a coherent empirical claim for which there is something interesting to be said.

My argument for (D) involves repeated exploitation of the trade-off between speed and architectural simplicity. (1) *In principle*, any computable function can be computed by a two-dimensional universal Turing machine, but in practice, three dimensions provides much greater computational speed. (2) In the same spirit, a von Neumann machine can compute anything computable, but massively parallel architectures can compute some functions with huge time savings. Is it possible that an *organic* parallel architecture can compute some functions faster than any inorganic realization of the same architecture? Yes, if – a big "if" – the function each neuronal node computes is a hard-to-compute function involving atomic-level information-processing by macromolecules. For, as Monod (1971) once observed, individual macromolecules can in principle compute functions, and the speed at which they exercise their "cybernetic power" cannot be duplicated in any larger assemblage of atoms (such as an electronic microchip circuit!) There are some serious objections to this proposal – for which no good positive evidence has yet been adduced, to my knowledge – but if we grant that this is even a remote physical possibility, then it could indeed turn out that (D) was true, for the entirely unmysterious reason that no electronic computer could reproduce an organic brain's real-time virtuosity.

And speed *is* of the essence, contrary to the impression typically created by Searle's intuition pumps (especially his imagined computer made of beer cans tied together with string). Actual AI models typically run too slowly, on their von Neumann machines, to compete with brains as real-time controllers in the real world, but that is of only practical – as opposed to theoretical – import. We must be careful here to distinguish between the genuine theoretical issue dividing the connectionists from the traditional AI theorists, and Searle's issue. Connectionism is still what Searle would call "strong AI," since it supposes that some connectionist program, in virtue of its enhanced *computational* power (which boils down to speed, since any connectionist program can be inefficiently computed by a von Neumann machine), is the "right program" – the program that is sufficient for mentality. [See also Smolensky: "The Proper Treatment of Connectionism" *BBS* 11(1) 1988.]

Searle, then, cannot adopt (D) as what he has been claiming. What, in contrast, can he be asserting in (S)? The "causal powers" Searle has invoked are sharply distinguished by him (as "bottom-up causal powers") from the "control powers" of a program. According to Searle, "control powers by themselves are irrelevant" (personal communication, p. 334 *Stance*). This doctrine then has a peculiar implication. Searle insists that it is an empirical question whether anything other than organic brains have these "bottom-up causal powers": "Simply imagine right now that your head is opened up and inside is found not neurons but something else, say, silicon chips. There are no purely logical constraints that exclude

any particular type of substance in advance" (forthcoming, see *Stance*, p. 333). But although such a discovery would settle the empirical question for *you*, Searle must admit that it would not settle it for any other observer! It is just a very obvious empirical fact, Searle claims, that organic brains can produce intentionality, but one wonders how he can have established this empirical fact. Perhaps left-handers' brains, for instance, only mimic the control powers of brains that produce genuine intentionality! Asking the left-handers if they have minds is no help, of course, since their brains may be just Chinese rooms.

This reveals that the causal powers Searle champions are mysterious indeed. Searle proclaims that somehow – and he has nothing to say about the details – the biochemistry of the human brain ensures that no human beings are zombies. This is reassuring but mystifying. How does the biochemistry create such a happy effect? By a wondrous causal power indeed; it is the very same causal power Descartes imputed to immaterial souls, and Searle has made it no less wondrous or mysterious – or incoherent in the end – by assuring us that it is all somehow a matter of biochemistry.

## 10. Midterm examination: Compare and contrast

The final chapter of *The Intentional Stance* leaves the close-quarters trench warfare of analysis, criticism, and rebuttal, and attempts a third-person, bird's-eye view of what has happened to theories of intentionality during the last quarter century. Happily, this perspective reveals that in spite of ongoing controversy about some matters, agreement has predominated about the most important points, and progress can be discerned. In fact, many of the most salient disagreements appear to be the amplified products of minor differences of judgment or taste, or tactical overstatement. Philosophers have been gradually homing in on a temperate, intermediate position, first outlined by Sellars (1954) and Quine (1960), with details added by Davidson, (1970; 1973; 1975) Dennett (1969; 1978), Stich (1981), Bennett (1976), and Putnam (1974; 1975; 1978; 1981). Fodor is the most notable iconoclast, but his valiant attempt to resist the gravitational attractions of this community of thinkers has led him to defend an increasingly narrow and inhospitable terrain. The potential consensus resisted by Fodor (and some others) agrees with Quine that the "dramatic idiom" of intentional attribution must be taken seriously, but not *too* seriously, treating it always as what I call a "heuristic overlay" (Dennett 1969) or a "stance."

NOTE

1. A context in a sentence is referentially opaque if replacing a term within it by another term that refers to exactly the same thing can turn the sentence from true to false. Sentences that attribute beliefs or other propositional attitudes to people are typically referentially opaque.