Reset version of an essay published in *Philosophy and AI: Essays at the Interface* (Robert Cummins and John Pollock, eds., 1991), pp. 39-77, itself a revision of CMU Computer Science report CMU-CS-82-149. Reprinted May 1995.

The Foundations of Psychology

A logico-computational inquiry into the concept of mind*

Jon Doyle
Laboratory for Computer Science
Massachusetts Institute of Technology
545 Technology Square
Cambridge, Massachusetts 02139

for Gerald Jay Sussman,
whose life is contagious

Abstract

We compare certain trains of thought in philosophy of mind and artificial intelligence which lead to a remarkable convergence of ideas. Demands from philosophy that psychological theories have predictive power join with demands from artificial intelligence that machines adaptively maintain their own mental state to suggest a conception of minds as narrowly realized psychological theories. We use this conclusion both to clarify the domains of study and scientific aims of psychology and artificial intelligence, and to suggest some methodological principles for constructing intelligent machines.

*This paper is revised from one dated February 18, 1982 and written while the author was with Carnegie Mellon University.

Preface

If today one asks "What is the brain?," one receives volumes of material from the neuroscientists. This material may not be very complete, nor terribly revealing about the operation of the brain, but it is a beginning one can hope to continue. On the other hand, if one asks "What is the mind?" or "What is thinking?" one receives today little information more certain than that supplied by the quaint theories of antiquity. On these latter questions one finds not just volumes of material, but volumes of material for each of many diverse theories of mind. Philosophers alone supply dozens, but they are not exceptional, for so do the psychologists, linguists, decision-theorists, artificial intelligence researchers, and novelists. These and other answers are in turn assumed by the social sciences, each of which depends on a conception of man for its formulation, and by moral thinkers in philosophy, theology, and politics for similar reasons. What's a body to think?

Is it not a scandal that such common notions as mind and thought enter so prominently in our understanding of the world, yet find so little definiteness in the sciences they underlie? How are the sciences to look for facts, to measure, analyze, or construct minds without a clear notion of the object of their study? Without a clear conception of mind, how can the sciences tell if they are studying minds or if they have accidentally drifted to studying something else instead? How, indeed!

In spite of the great wealth of theories extant, my purpose here is to present yet another, one which underlies many of my recent studies.[1] The concept of mind is ultimately a philosophical or metaphysical subject, and as a consequence my inquiry begins in largely philosophical terms. I approach the subject from backgrounds in mathematics and artificial intelligence, and use of mathematical, logical, and computational methods soon join philosophical methods in my investigations. I believe the use of both philosophical and mathematical tools necessary to this project. If successful, these results will inform researchers in artificial intelligence of the logical foundations and implications of their techniques, and will inform philosophical and mathematical logicians of both psychological applications of their theories and possible areas for new mathematical developments. I permit myself the hope that even the philosophers, if they examine what I have written without prejudice, will find in it something of use to them.

The best of tools are powerless, however, if the nature of the investigation is not clear. Sadly, there is some confusion about the aims of artificial intelligence and the other cognitive sciences. One finds views of artificial intelligence ranging from "frontier applications of computers," to "knowledge engineering," to "making machines act more intelligently," to "making machines simulate human thought," to "making intelligent machines." A glance at introductions to cognitive science reveals scant definition beyond a melding of artificial intelligence and psychology, together with hopes that maybe bits of philosophy, linguistics, education, neurosciences, etc., will make their way into the field. This is a recipe for a stew, not a statement of the scientific aims of a field. Artificial intelligence and cognitive science

[1] [Doyle 1983a, 1983b, 1983c, 1983d, 1988]

may in fact be or turn out to be all these things, but these must be consequences of the pursuit of yet-unarticulated scientific aims rather than the aims themselves.

Clarity about scientific aims is especially important for the present work, for our investigations will not make much sense or will be misconstrued without a proper understanding of their intended contribution. The title of this work is *The Foundations of Psychology*, yet one will find virtually nothing from modern psychological theories here. This paper is not so much about the modern discipline of Psychology as it is about all possible organizations for minds, i.e., psychologies. These are the objects of study of what I call *rational psychology*, which, in analogy with rational mechanics, is the conceptual investigation of psychology by means of the most fit mathematical concepts.[2] The disciplines of psychology and artificial intelligence are simply particular sub-disciplines of cognitive science with special interests to pursue. I share this last conclusion with the views mentioned above, but the former conclusion, upon which to me the latter is based, seems much less widely held, if others hold it at all. To avoid misconstrual of these investigations, the remainder of this introduction sets out the assumed scientific aims of rational psychology, psychology, and artificial intelligence.

Rational psychology is the study of all possible minds. It may not be possible to set out in advance a definite class containing all possible minds, just as biologists have had to abandon all definitions of "living things" and adopt an accommodating approach to newly discovered life forms. Nevertheless, this paper formulates a definition of what minds are as an initial foundation for their study. To jump ahead of ourselves, we view minds as narrowly realized theories, so that minds are not natural objects but theory-relative instead. The theories realized as minds are psychologies (not to be confused with the discipline Psychology), ways of viewing the organizations of minds. This definition can hardly mean much now, but we do assume some definite range of objects as minds. The task of rational psychology is to discover classifications of minds so that each mind can be uniquely characterized in terms of the system of classifications, and so that identically characterized minds are isomorphic in some natural sense. In other words, the aim of rational psychology is to characterize the equivalence classes of possible minds. Some classifications of minds will involve the sorts of constituents (e.g., beliefs, pains, etc.) from which mental states are constructed. In addition to internal structural classifications, other classifications involve relations between psychologies and other things. In the case of relating one psychology to another, one has classifications involving homomorphism, embeddability, or compatibility, of one psychology being a form of another. For example, if humans typically realize a general psychology ψ, then a psychology ψ' realized by a particular normal human, will be a more detailed version of the general psychology, and so will admit a homomorphism onto ψ. In the case of relating psychologies to non-psychological entities, the principal question is whether a particular psychology ψ is realizable in entities of class E. For example, the task of the familiar discipline of Psychology is to find psychologies that can be and typically are realized in human beings. One view of the task of artificial intelligence can be phrased as discovering "interesting" (e.g., human-like) psychologies which can be realized in Turing-equivalent machines.

Human beings and Turing-equivalent machines need not exhaust the range of entities in

[2] See [Truesdell 1966] and [Doyle 1983a, 1983b].

which to realize psychologies. Many interesting psychologies may lie beyond the realm of what is realizable given the physics of our universe. Consideration of this possibility may seem strange to those steeped in traditions involving empirical psychology and CHURCH'S thesis, but limiting the scope of rational psychology by the laws of physics is mistaken. The questions of rational psychology have content independent of the particular characteristics of our universe. This can be understood in several steps as follows.

Individual psychologies are particular theories of mind, precise specifications of mental organizations. If a psychology is a formal theory of a mental organization, realizations of this psychology are models of the theory. There is no normative or descriptive content in the notion of psychology itself, just the notion of a theory and its possible models. However, we may view psychologies from both normative and descriptive perspectives. We can ask for the psychologies describing some individual human, for example, and we can say some psychology specifies normal human mental organization. When viewed normatively, psychologies are competence theories in CHOMSKY'S sense, where realizations of the normative psychology are mentally "competent" agents, and realizations of differing psychologies are mentally "incompetent" agents (although their incompetence may involve being stronger as well as weaker in their faculties, unavoidable supercompetence considered a form of incompetence at meeting normative limitations).

CHOMSKY'S idea of competence theories has influenced much work in the cognitive sciences, but the applications of this idea may have been unnecessarily limited in comparison with the more basic notion of normative theory due to the context in which the idea was introduced. In the beginning of *Aspects of Syntax*,[3] CHOMSKY develops his competence-performance distinction as a tool in explaining how a finite mind can use an apparently infinite language, to say how a grammar makes possible the "infinite use of finite means." CHOMSKY motivates the notion of competence with a picture of an idealized speaker as one free of all memory limitations and free of certain computational limitations such as distractions, shifts of attention and interest, and random or characteristic errors. Since CHOMSKY speaks from a linguistic tradition which gathers yes-or-no judgments of grammaticality from subject speakers, it is easy to assume he assumes recursiveness of spoken languages. But to extend the requirement of recursiveness to the idea of competence seems unwarranted.[4] "Finite means" might be finitely axiomatized second-order theories like arithmetic, beyond the pale of recursive enumerability. Once one widens one's interest from human languages to rational psychology, limitations of recursiveness can be viewed as merely a characteristic error for the speakers of a non-recursive language. The aim of a normative or competence theory is to give an ideal against which to measure the performance of supposed practitioners. There is nothing in the notion of ideal *simpliciter* which entails recursiveness or even recursive enumerability. We should be able to consider an ideal consisting of just the true sentences of arithmetic. Of course, we can prove that no Turing-equivalent speaker can achieve this ideal of competence, but that is nothing new in the world. Suppose, for example, we try to formulate weight-lifting competence and performance. It seems plausible to take

[3][Chomsky 1965]

[4][Thomason 1979] discusses some problems with the notion of competence theories for cognitive scientists wedded to the assumption of recursiveness.

the notion of weight-lifting competence to be that one can lift barbells of any weight. Of course, there are many "processing" limitations idiosyncratic to all humans. Moreover, one can prove from physical laws that there are weights humans can never lift. Is this provable physical incompetence somehow specially different from provable mental incompetence? It seems unlikely that CHURCH'S thesis is anything but an empirical fact, just like the ordinary laws of physics. GANDY, for instance, has outlined a proof of CHURCH'S thesis from physical laws.[5] The apparent equivalence of effective computability and recursive enumerability may be nothing more than an amazing coincidence of our universe. But if the limitations of CHURCH'S thesis are empirical rather than logically necessary, there might exist universes in which machines have super-Turing powers of computation, and excluding these universes and their inhabitants from the domain of normative theories is nothing but chauvinism.

By virtue of their use as normative theories, we must conclude that psychologies need not be actually realizable to be of interest to rational psychology. This range of concern frees rational psychology to calmly study controversial assumptions. Rational psychology studies in part the range of psychologies realizable by machines limited to the effectively computable, where what may be computable effectively by machines may vary with the laws of physics. Artificial intelligence will be the main research vehicle for studying mechanical intelligence in our universe. Rational psychology also studies the full range of psychologies realizable by arbitrary physical systems. In some universes, these may include non-mechanical psychologies, and in others, these may be the same as the mechanical psychologies. For example, human psychologies may or may not be mechanically realizable in our universe, but this is no cause for heated debates to impede the progress of rational psychology. Mechanists will relate their studies to the comparable but independent questions of artificial intelligence, and non-mechanists will not be so bothered unless it is to demonstrate the non-effectiveness of some aspects of human psychologies.

In addition to this argument from anti-chauvinism, one might arrive at the proposed scope of rational psychology as one more fruitful scientifically than one limited to our universe. Arbitrarily limiting the range of acceptable psychologies may unnecessarily prevent discovery of important facts about uncontroversial psychologies. The more generally posed problem may admit an easy solution even when the special case presents intractable difficulties. For example, real and complex numbers may be "unreal" in some sense in which integers are not, but some facts about integers and integer functions (e.g., the prime number theorem) are obtained most easily as facts about real and complex numbers and functions, and can be obtained only with extreme difficulty as facts about integers alone. In the same way we may most easily see things about the minds of men by looking, so to speak, to the mind of God. Limiting the domain of rational psychology to the boundaries of our own universe and excluding the trans-computable is analogous to limiting mathematics to the integers and excluding the transfinite. This analogy has a moral. CANTOR led mathematicians into a paradise many will not abandon due to the rich harvest they find there. Other mathematicians stay outside to see what fruit hang on branches reaching over the walls. In comparison, rational psychologist have hardly begun to taste the fruit of their field. Perhaps

[5][Gandy 1980]

our paradise grows only sour grapes, but we must taste them to see.

Introduction

§**1.** After proceeding for centuries without exceptional standards of rigor, philosophy of mind within this century has been the beneficiary of three great boons. The first of these was the development, beginning primarily with FREGE, RUSSELL, and HILBERT, of modern mathematical logic and metamathematics.[6] The second boon was the development, beginning with WUNDT, FREUD, and WATSON, of modern scientific psychology.[7] The third boon was the development, with the advice of TURING and VON NEUMANN, of the general-purpose electronic computer.[8] Mathematical logic has provided the language for precise formulations of problems of mind and their solutions. Scientific psychology has provided many careful observations of human behavior and performance to account for in theories. And the high-speed electronic computer has provided the means for experimentally investigating the consequences of partial and comprehensive theories of mind in ways that were previously infeasible. Together these three developments have stimulated the new field of artificial intelligence. With the intellectual and practical tools at hand, artificial intelligence promises the most precisely formulated and visibly detailed theories of mind yet developed. Where previous centuries served as grounds for battles between imprecise, ill-understood proposals, battles allowing little hope for eventual comprehension of or agreement on the nature of the issues at stake, the new tools permit far greater clarity in formulating the concepts and structures of theories, so that parties to debates at least can agree on the issues, if not on the answers. Properly applied, the new tools ruthlessly expose inadequacies of theories previously hidden behind the general vagueness of formulations of the theories, and this shows up in the aphorism of artificial intelligence which states that about the nature of mind, the first twenty obvious ideas are wrong. More clearly than ever before, the new tools prove that in psychology simple moral conviction, even supported by a mass of successful applications, is not enough. Rigor in formulation must accompany conviction if the field is to visibly progress.

§**2.** Precision in formulating psychological theories leads quickly to a need for clear conceptions of the nature of mind and the nature of thinking. In the past, psychological theories have all too often lacked clarity through vagueness about the object of the theory. Routinely, investigators would propose hypotheses in great detail about some particular component of mind or thought, while merely alluding to an overall structure never made precise. This course holds two dangers: first, that not being closely tied to a definite conception of mind, the component theory will wander off to become an abstract plaything of its creator; and second, that the component theory will become vacuous by pushing all important theoretical burdens onto the vaguely defined theory of the whole. To avoid these dangers, we must strive for as much precision in setting out concepts of mind as in proposing theories of particular

[6] See, for example, [Frege 1884], [Whitehead and Russell 1910], and [Hilbert 1900].
[7] See, for example, [Wundt 1874], [Freud 1895], and [Watson 1914].
[8] See [Goldstein 1972] and [Randell 1975] for histories.

mental components.

If first aim of psychology must be to set out at least skeletal theories of mind, lest any empirical or detailed studies of supposed mental components risk triviality or irrelevance, then the first step toward this aim must be to understand the possible forms of psychological theories, to understand the range of possible conceptions of mind. Care is required in this enterprise, for historically there has been frequent temptation to confuse the theoretical tools used to formulate theories with the theories themselves, and this confuses the aims of psychology with the aims of mathematics, logic, computer science, and other fields. For example, logic is not part of psychology, and neither is psychology a part of logic. Nevertheless, the temptation is frequently great to transfer the methods of logic to the methods of thinking and *vice versa*. Just as Modus Ponens is well established in logic, many non-deductive inferential processes are well established in psychology, and confusing the logical notions of entailment and proof with the psychological notions of argument and thinking is just wrong. Yet numerous students of mind make this confusion, and in consequence many works in artificial intelligence place psychological burdens on logic, faulting it for remaining silent on properly psychological problems, and mistakenly turning away from the clarity of formulation logical tools provide. Complementing this, other works view the purpose of thinking as the production of new knowledge from old, to which logical deduction is well suited if anything is. This transference of purpose oversteps the proper role of logical rigor in psychology. The purpose of thinking is insight, not knowledge. The aim of reasoning is not merely to ensure the truth or drawing of a conclusion, but more fundamentally, to afford insight into the dependence of conclusions upon one another. From certain interdependencies of conclusions, we can tell that one conclusion will be true if others are; and from other interdependencies, that one conclusion will be held by the agent if others are. On the face of it, these are different relations between conclusions. It is not difficult to think of further interesting relations between conclusions, such as confirmation or falsification in inductive reasoning. In fact, each of the disciplines brought to bear in formulating a conception of mind may have its own special interdependencies, but these must not be confused with each other. The separate disciplines must enhance each other, not replace each other.

§**3.** Although I am led to ask these questions in the course of formulating psychological theories, philosophical motives too have prompted me to enquiries of this kind. Our understanding of the questions traditionally raised about the relation of mind and matter, of free will and determinism, and of the limitations of effective computability all must increase from consideration of the nature of psychological theories. We may, if lucky, arrive at answers to these questions, or at formulations of possible positions if not answers. Even if no conclusions result, at least the outlines of the mysteries will be clarified.

§**4.** Surprisingly, starting from these theoretical and philosophical questions about the nature of mind, we are led to formulate the same demand as which had arisen independently in the practice of artificial intelligence, namely that psychological theories be cast as sets of narrowly interpreted self-specifications; for only if a theory is suitably narrow in the range of its references can it be realized in humans or machines; and only if a theory can be used in isolation to reconstruct states of mind after changes does it facilitate the design of artificial agents. Our story is largely a tale of the two theories leading to this remarkable convergence.

Views of certain writers on the nature of mind

Are there thoughts?

§**5.** During most of the history of philosophy, men have been assumed to have thoughts. More often than not, thoughts and their vehicle, the mind, were taken to be different somehow than the ordinary stuff from which our bodies are composed. A high point in this train of thought was the skeptical argument of DESCARTES.[9] By elementary considerations, DESCARTES managed to convince himself that his mind existed. He required substantial additional hypothesis and argument to convince himself that his body and other ordinary things existed as well. Based on these considerations, DESCARTES divided the world into two parts: the realm of matter, from which our bodies are composed, and the realm of an immaterial sort of substance, from which our minds are composed. For a long time after DESCARTES, the principal themes in discussions of mind consisted of speculative theories about the relation and possible interaction of immaterial mind and material body. These theories were hampered by the growth of the physical sciences, for as more and more of the universe came under the domain claimed by physics, it became increasingly difficult to supply a plausible account of how an immaterial object, not subject to physical law, could influence or be influenced by a material object subject to those laws.

§**6.** Fortunately for the growth of a somewhat less speculative psychology, another train of thought was developing in the wings. Long ago, the Epicurians, taken with DEMOCRITUS'S ideas about atoms, proclaimed that all the world was composed of atoms.[10] Men's minds, no less than their bodies, were made up of atoms, and the behavior of minds could be explained like the behavior of bodies, in terms of the mechanical laws governing the motions of atoms. However, the Epicurians proposed no details of these explanations and laws. Instead, they simply claimed the soul was composed of atoms lighter and more mobile than ordinary atoms, so presumably accounting for the rapidity with which thoughts can progress while leaving no outward sign of change visible to observers.

After some delay, the Epicurean programme was taken up by HOBBES, who tried to give some explanation of the workings of the material mind. HOBBES took his models of mechanics from the new physics, which clarified the notions of forces and inertia. With

[9][Descartes 1637]

[10]Much of my knowledge of these fragments of history comes from [Boring 1950] and [Peters and Mace 1967].

these, he explained thought and imagination as the decaying motion of the atoms of the sense organs as they stimulate other atoms and are stimulated by new impressions. In spite of the possibilities of this programme, HOBBES was limited by contemporary ignorance about neurophysiology and non-naive psychological phenomena, so his theory remained largely speculative and general about mental mechanics.

After yet another delay, these ideas received their first detailed treatment. An experimental, scientific psychology developed under the impetus of WUNDT and others, and much information became available about psychological phenomena.[11] However, much of this information still depended on uncritical speculation in the form of introspective evidence. WATSON urged that psychology secure its foundations and either shore up introspection or avoid it altogether in favor of theories of overt and neurophysical behavior.[12] He took the latter course, and his programme, psychological behaviorism, became widely influential, and led to much progress in matters such as neurophysiology, neurochemistry, and neuropsychology.

§**7.** Given this development of ideas, it seems like a small step from psychological behaviorism to philosophical behaviorism. Where psychological behaviorism, in modest formulations at least, seeks simply to determine the laws of mind and thought and their realization in the body, philosophical behaviorism, as championed by RYLE and SKINNER, claims reducibility to overt behavior as well.[13] For the philosophical behaviorist, mental phenomena can not only be realized in terms of neurological hardware, but can be reduced to those functionings of that hardware, in the sense that all theories about belief, desire, inference, and action can be formulated purely in terms of the overt behavior of the body. In this view, mental entities and processes are simply unreal, and their use is as unscientific as phlogiston descriptions of combustion. RYLE, for example, sought descriptions of human behavior strictly in terms of overt acts, dispositions to behave, and changes in dispositions to behave, where dispositions to behave were formulated purely as physical states and physical laws.

Philosophical behaviorism eschewed the mind. Indeed, some of its adherents went so far as to deny any introspective self-awareness in their own cases: not a tactic unknown to philosophy, but one never before practiced on such a grand scale. Cartesian skepticism seems meager compared to that of the philosophical behaviorists.

§**8.** Philosophical behaviorism is implausible, and for very simple reasons. Suppose one attempts to analyze an ascription of belief, e.g., "Fred believes his computer program has a bug" in terms of behavior and dispositions to behave. We can of course guess at predictive generalizations about Fred's behavior given this belief, such as "Fred is disposed to log in and debug his program," but for any of these predictions to have any plausibility, they cannot be formulated simply in terms of behavior and states of the world; they must refer to other

[11] [Wundt 1874]
[12] [Watson 1914]
[13] [Ryle 1949], [Skinner 1957]

of Fred's beliefs and desires that might influence his dispositions, e.g., "Fred believes the computer is down" or "Fred wants to take his time so as to increase his wages." Because predictions of Fred's behavior must refer to his thoughts about things as well as to the actual state of things, philosophical behaviorism is untenable.

§**9.** With Cartesian dualism indefensible, but with its prime alternative inadequate, philosophers developed several replacement theories which identified mental states and processes with the physical states and processes of the brain and connecting parts of the body.[14] These identity theories, as they are known, hoped to analyze beliefs and desires, for example, as physical predicates of brains, thus allowing the objections to philosophical behaviorism to be overcome by phrasing references to beliefs and desires as references to certain sorts of brain states. The identity theorists returned to a position much like that of Hobbes, although unlike Hobbes, they had a stronger physics and neurophysiology to draw upon for formulation and examples.

§**10.** Although the identity theories avoid the obvious problems with philosophical behaviorism, they suffer from inadequacies of their own. Curiously, these inadequacies are made all the sharper by the development of somewhat intelligent machines in artificial intelligence, although the inadequacies can be brought out without using those machines as examples.

Identity theories claim that all statements about mental phenomena are reducible to statements purely about physical brain states, with the implication that perhaps we are better off studying the brain states in psychology and leaving the mind alone. The first objection to this idea, at least to its implication, is that even if such a reduction is possible, it is useless in practice and so even theoretically uninteresting (at least to those who include practical power as a measure of a good theory). Putnam gives as example the analogous case of a description of a hole in a board in terms of the location and momentum of elementary particles.[15] If the hole is round, we cannot insert a square peg of equal cross-section area, and presumably this fact can be explained in terms of the locations of and forces between the particles comprising the board and peg. But such an explanation must be astronomically long, and must involve many details that are in some sense irrelevant to the important facts, namely, the roundness of the hole and squareness of the peg. Dennett gives as another example chess playing computers.[16] Unless one is a chess master, one's best tactic in playing such a machine is to treat it as a rational agent with beliefs, desires, and knowledge of the game, for the actual sequence of computational steps the machine employs to develop its moves is incomprehensible (at least in any reasonable amount of time for individual humans). Dennett describes this observation as the notion of "intentional stance": that no matter what other theories are possible concerning the realization of mental phenomena, we still must use a psychology formulated in mental terms, because mentalist ascriptions are the ones of practical predictive power (for humans, at any rate).

[14] See, for instance, [Place 1956], [Feigl 1958], and [Smart 1959].
[15] [Putnam 1975c]
[16] [Dennett 1978]

The identity theories also fall prey to a stronger criticism: that their supposed reduction of psychological physical laws cannot be done, even in principle, without severely gutting the notion of law. Suppose one has a psychological theory formalized in mental terms, a physical theory formalized in terms of particles and fields, and, as the identity theorists suggest, a set of "bridging laws" or "bridging definitions" connecting the two theories. FODOR points out that because there are presumably many, many ways of realizing minds in matter (just as many physical objects can be used as money), the bridging statements cannot be laws unless they are so wildly disjunctive as to ridicule our ordinary conception of what it means to be a law.[17] FODOR continues by pointing out that if the bridging statements are reasonably informative or concise definitions, then the psychological theory that results cannot aspire to the title of law either — its applicability need not be wide, nor need its conditionals have counterfactual force.

§**11.** These criticisms of identity theories contain the seeds of a theory to replace the identity theories, functionalism. Functionalists contend that mental states are functional states; that is, what makes something a belief or a desire of the agent is the fact that that something plays a certain role in the processes of the agent, independent of any other properties possessed by its particular realizations in brains, computers, or ghosts.[18] Expressions of functionalist psychological theories take the form of physical theory plus a collection of Ramsey sentences: existentially quantified statements asserting the existence of something that bears certain relations to other things which, in light of the criticism of psychological behaviorism, will include some of the other entities asserted to exist by the Ramsey sentences. For example, a ten cent soda machine accepting either nickels or dimes might be described by the two sentences "There exists a state S_0 such that receiving a nickel in S_0 results in a change to state S_5, and receiving a dime in S_0 results in dispensing a soda and remaining in S_0" and "There exists a state S_5 such that receiving a nickel in S_5 results in dispensing a soda and a change to state S_0, and receiving a dime in S_5 results in dispensing a soda, dispensing a nickel, and a change to state S_0." Functionalism thus permits some separation between the development of a psychological theory and the delineation of the ways in which the theory may be realized in humans or machines. However, functionalism is not without problems of its own, and some of these concerning surprises about what things can and cannot be functional states will arise in the following.

Have thoughts content?

§**12.** Even if one accepts functionalism as a hopeful theory of the mental, there are many serious difficulties that must be overcome. The first of these is the nature of thoughts as conceived in functionalist terms. To understand this problem, we must examine the functions

[17] [Fodor 1975]

[18] See, for instance, [Putnam 1975b] and [Fodor 1968].

thoughts play in mental activity.

To begin a functional characterization of the role of particular thoughts in thinking, we can take as a crude approximation all the truisms about rational thought and action. For example, a functionalist psychology might have specifications like the following: (1) "If agent A desires p and believes M a method for achieving p, then A desires to do M, other things being equal," and (2) "If A desires to do M, and believes doing M is better than doing any other action A desires, then A does M." Our lazy programmer Fred might desire to earn a living without the necessity of toil, might believe that being slow to debug his programs a way of doing this, and might in fact be slow in his chores thinking this the easiest way of life. Of course, specifications (1) and (2) are woefully simplistic, and any careful treatment would have to be based on a better theory of rationality and practical reasoning, but the example serves to illustrate the basic idea.

§**13.** In addition to basing functional specifications for the mind on sound theories of thinking and doing, we must take care to ensure that the specifications are suitably narrow in scope. For example, we might consider a specification like ($1'$) "If A desires p, and M is a method for achieving p, then A desires to do M, other things being equal." The problem with ($1'$) is that it puts a measure of omniscience into the specifications, omniscience that offends our notion of what a psychology should involve. While one might give such specifications in ecological theories of how well agents fit into their environments, for psychology proper we eschew any specifications which do not express what PUTNAM terms "psychology in the narrow sense."[19] Psychology in the narrow sense is concerned only with narrowly realized psychological theories, where narrowly realized psychological theories refer only to the agent's personal mental structures, and do not refer to any correspondence or lack of correspondence between the agent's thoughts and its environment, such as the truth of its beliefs or the physical possibility of its desires. Narrowness is a property of realizations or interpretations of psychological theories, rather than a property of the theories themselves, since a particular theory might admit interpretations stepping outside the agent as well as strictly internal interpretations. For example, to return to the ten cent soda machine introduced above, the functional specifications given there were not narrowly realized. Those specifications referred to dispensing of sodas, objects presumably beyond the machine's ken. Ordinary soda machines (at least the ones I have lost money to) only send signals to the dispenser for its jaws to open, so allowing a soda to fall if one is there, but failing to dispense a soda (unbeknownst to the machine) if the soda rack is empty or jammed. To describe the machine's structure narrowly, we must replace references to sodas by references to signals sent from the cashier to the dispenser. FODOR terms such restrictions of narrowness on acceptable interpretations of psychological specifications "methodological solipsism."[20]

§**14.** Adherence to the methodology of psychological solipsism requires that we re-examine our naive psychological theories to excise all externalities. The example specifications (1)

[19][Putnam 1975a]
[20][Fodor 1981]

and (2) given above seem to refer to the content or meaning of the agents beliefs, desires, etc., but notions of meaning or interpretation are usually external relations of the agent to the world, hence the use of such notions in our psychological theory is suspect. We could immediately excise these notions from our psychology, but this would be hasty since there are many reasons why we would like to have some notion of content available for thoughts. The agent must be able to compare or distinguish its thoughts, one from another, lest its psychology be completely trivial and implausible. Some notion of content is also required for comparing thoughts of the agent at different times, and with the thoughts of other agents. We clearly can make some powers of discrimination between thoughts available to the agent simply by relying on a formal syntax of thoughts, but it is less clear whether these discriminations will match or can match the discriminations needed in diachronic and inter-agent comparisons. These latter comparisons are more clearly dependent on interpretations external to the agent, but all of these cases call for careful examination.

Before considering the answers that have been proposed for these questions, we introduce a new bit of terminology. Functional specifications of psychologies have thoughts playing certain roles in the agent. We can turn this relation around and for convenience say that the agent bears a certain relation to its thoughts. Indeed, this latter phrasing suggests some of the motivation for RUSSELL'S "propositional attitude" terminology, since the agent-thought relation can be viewed as a certain stance or attitude taken by the agent toward a possible thought. In this phraseology, the requirement of methodological solipsism becomes the requirement that thoughts are graspable, that is, some how manipulable and determinable by the agent itself. The agent presumably bears some relation to the real world, but if the agent is only a small part of the world and the weakest forms of skepticism are justified, the agent cannot fully grasp its relation to the external world.

§**15.** The first suggestion we consider about the nature of thoughts holds that the content of thoughts are abstract entities called *propositions.* The term "propositional attitude" stems from this orientation. Propositions as the content of thoughts comes as a suggestion from two quarters: BRETANO'S theory of intentionality and FREGE'S theory of language.

BRETANO distinguished mental states from physical states of an agent with the idea of intentionality.[21] He claimed that all acts of consciousness, such as beliefs and desires, are *directed towards* or *about* some object or objects. For example, a hatred of paperclips is an act of consciousness whose intentional objects are all paperclips. In the case of the belief or desire that some condition obtains, BRETANO maintained that the object of the belief or desire is a proposition.

From another orientation, FREGE distinguished the sense of names or terms from their reference.[22] To use the celebrated example, the terms "author of *Waverley*" and "Scott" have different senses, but refer to the same person. FREGE claimed that thoughts express propositions as their sense, and may be either true or false as their reference. (Actually, this

[21] [Bretano 1874]
[22] [Frege 1892]

takes liberties with FREGE'S conception, but the details are not crucial here. Moreover, he used a term translated as both "thought" and "proposition", calling what we call thoughts "ideas" or "concepts".)

These suggestions have attracted many critics and defenders, but the main point we note about these suggestions is that such propositions cannot be part of a suitably narrow psychological specification for agent's structure, i.e., propositions are not graspable. PUTNAM observed that one can distinguish between real and apparent propositional content in beliefs, just as earlier skeptical arguments distinguish between the real and apparent truth of beliefs.[23] PUTNAM'S examples are rather involved, but KAPLAN gives the simpler examples of thoughts involving indexicals, i.e., implicit self or temporal reference.[24] For example, on Monday I think the thought "Today is beautiful." In the propositional theory, I can think exactly the same thought on Tuesday by thinking "Yesterday was beautiful." But this identity of actual content between these thoughts is indiscernible to me, since I might lose track of the days and think "Yesterday was beautiful" on Wednesday, thus actually expressing a proposition about Tuesday while thinking I am merely reminiscing on Monday's glories. Thus what proposition I actually express with my thought is not determined by my narrow psychological state.

§**16.** One can back off from propositions with the suggestion that thoughts are the relation of the agent to sentences. This suggestion has been motivated both in terms of an aversion for introducing abstract objects like propositions, and in terms of seeking graspable thoughts. Rather than introduce propositions as the meaning of sentences, and having propositions be true or false depending on circumstances, QUINE recommends simply having the sentences themselves be true or false in circumstances directly, rather than indirectly through the medium of propositions.[25] And rather than have an agent's thoughts involve grasping ungraspable abstractions, FODOR recommends having thoughts involve realizable representations, such as electrical patterns in the brain, markings on paper, or data structures in computer storage.[26] While this suggestion avoids the pitfall of a non-narrow psychological specification for the structure of mind, it raises other serious questions.

The first problem for the sentential view is that of the language of thought (sometimes called "brain-writing" or "Mentalese"). Since the sentences involved in thoughts are not just abstract entities but members of some language or representational system, they have a concrete syntax, and this places the sentential theory on the horns of a dilemma. Either the language of thought is common to all agents at all times, in which case diachronic and inter-agent comparisons of thoughts can be made within the confines of solipsistic psychology, or else the language of thought can depend on the temporal development of individual agents, in which case functionalism's claims to general applicability are sabotaged.

[23] [Putnam 1975a]
[24] [Kaplan 1977]
[25] [Quine 1970]
[26] [Fodor 1975]

The hypothesis of a universal language of thought is difficult to accept for many reasons. The first objection is that if there are several (actual or possible) species of agents, the hypotheses claims a universality that can only be defended on grounds of cognitive necessity, on grounds that the very nature of successful or rational thought and action entails, by means of physical or computational necessity, the features of Universal Mentalese. While it seems plausible (perhaps weakly so) that some general features of language must be forced by the necessities of the task, it requires considerable demonstration that all features of the languages must be the same. One might claim that all agents of a certain species begin with the same innate language, or grow to accept the same language through learning of cultural conventions, but these suggestions are suspect on the grounds of simple genetic and educational variability.

But if the hypothesis of a universal language of thought is implausible, the acceptance of individual evolving languages is not without theoretical difficulties or unpleasant consequences. Theoretically, one may have to retreat from the view that a language has a definite set of sentences as members, for actual questions of membership may be infeasible or may change the language, so challenging the sensibility of determining the language by combining hypothetical judgments. Practically, acceptance of this hypothesis means that the accuracy of memory becomes much more problematic, since sentences brain-written in the past may play a different role in the agent's language in the future, if they still are in the language at all. This entails either continuous rephrasing of memories when the language changes, or loss of the power to interpret memories. This last possibility is not without attraction, since it fits well with Piagetian-style theories of psychological development, in which the conceptualization of the world changes radically as a child grows. I can hardly recall a thing of my childhood: perhaps I still have all my brain-records, yet cannot recall or make sense of them any more in terms of my present mental language.

§**17.** Psychological solipsism seems to force us to accept individual, evolving languages of thought with their attendant temporal and contentual ambiguities for the agent. This conclusion seems in some ways to defeat the motivation for functional specifications of psychologies. For example, recall specification (1): "If agent A desires p, and believes M a method for achieving p, then A desires to do M." Since narrowly interpreted psychology cannot refer to actual content, we cannot include specifications like "A desires p", since that refers to our interpretation of A's attitude. We cannot even say instead "A thinks that A desires p", since that merely shifts the problem to the accuracy of A's introspective beliefs. Instead, we are driven to write instead things like ($1''$): "If agent A incorporates A-now-Mentalese sentence S_1 and A-now-Mentalese sentence S_2, then A incorporates (or adopts) A-now-Mentalese sentence S_3," where we must write the concrete A-now-Mentalese sentences S_1, S_2, and S_3 in our theory, since we cannot refer to their external interpretation. But does such a theory fit our intentions for a psychological theory? We started by thinking we could sharpen up our naive beliefs about psychology by expressing them as a functional theory, but we find that if we want a narrowly interpreted theory, we get a purely formal set of specifications like ($1''$) for each individual agent at each atomic interval of time. Such a theory may leave open some questions about the details by which the agent actually realizes all these sentences and computes with them: but it certainly seems to have abandoned the generality of realization motivating functionalism in the first place. It may be possible to salvage these ideas by formulating general psychological theories which describe both universals of a species's psychology and which postulate the existence, for each agent at each time, a personal language instantiating or approximating the general one, but we have none to propose at this point.

§**18.** Another suggestion about the nature of thoughts is less easy to place within a non-behavioristic theory of mind, and that is the idea of possible-world interpretations of propositional attitude ascriptions.[27] In this approach one says A believes p if p is true in every possible world compatible with what A believes. This may seem circular, but one can take as primitive A's acceptance or grasping of a set of possible worlds, among which A's actual world is supposed to lie. With such a primitive conception, talk about beliefs is reduced to talk that does not involve beliefs in an attractive way.

However, this approach suffers from several difficulties. First, it entails that an agent's beliefs are closed with respect to logical consequence. It is certainly not obvious that (or if) the functional specifications for beliefs of any sufficiently rigorous intuitive psychology entail deductive closure, yet such closure seems to be an inescapable consequence of the possible-world approach. A second difficulty is that the earlier arguments about the ungraspability of meaning would seem to have similar force for the graspability of a set of possible worlds. If these problems are not enough, the third difficulty with the possible world approach arises when we try to extend the idea to other attitudes. If we take exact analogies for interpreting ascriptions of desires, hopes, fears, angers, etc., we get the same closure problem. But it is a strange psychology which does not allow for conflicting yet limited set of desires (or beliefs, for that matter). I can desire to have my cake and eat it too and still not desire nuclear war. In addition, a psychology should be sensitive not just to logical incompatibility between desires but also to compatibility with respect to the agent's beliefs, yet the straightforward possible-world approach cannot incorporate such a notion. Instead, the only way open for making sense of possible-world compatibility seems to be by using the psychological theories to determine possible worlds as compatible sets of mental attitudes. And this is simply an acknowledgement of the anti-reductionist arguments mentioned earlier in support of functionalism.

[27] See, for instance, [Hintikka 1971].

Views on the nature of machines

Models of computation

§**19.** For most of history, man has had only simple machines. Machines had at most a few parts, often fixed rather than moving, and the invention of a new useful machine was a significant event for civilization. Part of the reason for this contrast with current times, in which vast numbers of new machines are constructed each day, must be that machines were less reliable in the past, and so if a complex machine was constructed from simpler ones, its chance of working was small. The big advances were not simply the construction of new machines, but of ones simple enough to be reliable, rather than Rube Goldberg contraptions. Would any king today pine for a mechanical nightingale?

The obstacle of unreliability is serious. One of our most powerful intellectual tools for understanding the world and for taking action is *problem decomposition*. Problem decomposition is the technique of breaking one problem up into several hopefully simpler subproblems together with a way of combining the subproblem solutions into a composite solution to the original problem. When I make a cake, I do not throw all the ingredients into the oven and expect a cake to appear. Instead, I break the task into first making the batter, then baking it, then making the frosting, and finally assembling the finished cake. Moreover, this trick works for more serious problems than cooking. But for problem decomposition to work, the composition methods must be effective. When most machines were not terribly reliable or precise, new machines could not be designed by straightforward problem decomposition because the composite machine would be hopelessly unreliable, if it worked at all. Until science and technology had progressed to the point where most machines could be built to be reliable, invention of useful new machines remained largely a matter of inspiration, luck, or natural analogy.

§**20.** The modern theory of machines developed with the advent of reliable machines, and has focussed mostly on computational devices, since modern computers and the specialized machines realized in them via programs are the most complex machines ever known. Yet though modern machines are more reliable and can be combined in more complex constructions, theoretical models of machines do not always reflect this. Instead, the earliest developed models of machines offered little sense of machines decomposable into parts, and concentrated solely on a notion of the machine as a whole.[28]

Whatever their inadequacies regarding decomposability, machine models fulfill the subsequently formulated philosophical demands about narrowness. Normal uses of machine specifications involve narrow interpretations of the machine states. The idea of machine is

[28]For presentations of many machine models, consult [Minsky 1967] and [Aho, Hopcroft, and Ullman 1974].

closely connected with the idea of effective calculability, and non-narrow interpretations of machine specifications make the machine operations non-effective. In fact, effectiveness is a stronger requirement than simple narrowness, and later we see what things might fill this gap.

§**21.** One of the first models of machines was that of the *finite state machine*, introduced at least as early as 1936 by TURING.[29] A finite state machine is simply a transducer of input strings into output strings, with a finite amount of memory. Each finite state machine M is completely described by two functions F_{M} and G_{M}, such that if M is in state $S_j \in \{S_1, \ldots, S_k\}$ and receives input $I_j \in \{I_1, \ldots, I_l\}$, it emits the output $F_{\mathrm{M}}(S_j, I_j) \in \{O_1, \ldots, O_m\}$ and moves to state $G_{\mathrm{M}}(S_j, I_j) \in \{S_1, \ldots, S_k\}$. The finite set of states of the machine limits its memory capacity, so that there are severe restrictions on what can be computed with a finite state machine. Nevertheless, finite state machines are theoretically interesting and frequently useful in practice. Unfortunately, this model of machines provides a relatively poor foundation for understanding machines by means of problem decomposition. Each of the states of a finite state machine is atomic, so there is no overt sense of a finite state machine having parts or being constructed out of submachines. Indeed, one of the heights of the theory of finite state machines is a characterization of when a finite state machine is equivalent to the sum, product, or concatenation of smaller finite state machines. We must look further to find models of machines congenial to problem decomposition.

§**22.** Another early machine model is the *Turing machine*.[30] Turing machines are simply finite state machines which can read and write symbols from a finite alphabet on an infinite tape. The tape symbols form the input alphabet of the finite-state controller, and combinations of symbols to write and tape motion signals form the output alphabet. Although Turing machines put some of their state onto the tape, where decompositions can be observed by using separate areas of the tape for separate sub-computations, Turing machines do not facilitate problem decomposition much more than do finite state machines, since the tape controllers are just finite state machines with the difficulties observed previously.

§**23.** Machine models closer to the structure of modern computers were formalized in the *random access machine* (RAM) and *random access stored program machine* (RASP).[31] A RAM has a read-only input tape, a write only output tape, an infinite array of memory cells, and a program. Each tape square or memory cell is either blank or may contain an integer. The machine's program is a finite list of numbered instructions from a certain fixed repertoire which can make arithmetic computations and comparisons on the memory cells as well as specifying the number of the next instruction to execute. RASP's are just like RAM's except that the program is stored in the memory cells, so that the machine can modify its own instructions by altering the contents of the memory cells storing the program.

[29] [Turing 1936]
[30] [Turing 1936]
[31] See [Shepherdson and Sturgis 1963] and [Elgot and Robinson 1964].

RAM's and RASP's are much better than finite state machines or Turing machines at facilitating problem decomposition. Because both the program and memory are broken into discrete components, RAM's and RASP's computing a combination function can be built more or less by concatenating and renumbering the programs and by relocating the memory segments used by each sub-machine to disjoint (possibly interleaved) components of the combined memory. These are essentially the ideas of subroutines and linking loaders so important in modern programming systems. Such combination of machines is possible because each of the operations of the machine changes only a bounded component of the machine's state, i.e., a couple of cells and the program counter. This means that all operations ignore almost all of the machine state, so that separate sub-machines ignore each other when combined, except in their desired communication channels.

§**24.** Models of computation moved most recently to a position of complete decomposability in the functional programming languages.[32] We must live with an unfortunate coincidence of terminology between "functional" specifications of roles and "functional" programming languages. The notion of roles played by some object need not be the same as the mathematical notion of function. We try to minimize confusion by always referring to the latter notion as "functional programming." In the functional programming model, sub-machines realize self-contained, arbitrary type functions, and composite machines are constructed by functional composition and application from sub-machines. Since no machines share any structure, there can be no interference when they are combined. This greatly facilitates problem decomposition. In fact, the principal theoretical difficulties with functional programming languages involve how to re-introduce shared structures and side-effects in a useful way. Shared structures are often important for economy of storage usage, and for economy of effort in updating a database common to a number of separate processes. Unfortunately, functional programming languages have gone too far in seeking decomposability, since structure sharing is outside the domain of a pure functional programming language, so that a less extreme position must be found.

Functional specifications in programming

§**25.** Most actual computers in use today are organized something like RASP's in the so-called *Von Neumann architecture*. Reflecting this common architecture, most programming languages are based on the idea of combining procedures with local instruction and data sets communicating through global or shared data sets, although there are a few languages like LISP and APL which come close to the functional language conception. One important consequence of the typical structure of computing machines and programming languages, if it is merely a consequence and not more deeply intertwined, is the phenomenon of procedural thinking among programmers. Most programmers find it easiest to compose programs by

[32] See, for example, [Backus 1978].

conceiving or imagining sequences of operations that wind up with the intended result. Only after they have composed the program in this way do they, if ever, reflect on what the program is computing to explain the functional relationships between the pieces of information it manipulates. This do first, reflect later phenomenon may be simply a consequence of the intellectual culture in which the programmers were raised, or it might stem from the human mind being a decomposable system in which most of one's mental state is automatically conserved from one moment to the next, facilitating envisioning of individual actions. In any event, programming methodologists have had to develop ways by which conscious functional decomposition can be facilitated in programming. The popular methodologies ("structured programming," "stepwise refinement," etc.) are suggestions for how to break problems up into explicit subproblems, how to combine the separate sub-solutions, and how to ensure or check the correctness of any non-interference assumptions made in the process of combination. Of course, these methodologies are still more hints about how to think than recipes for programming. The topic of how to decompose problems into subproblems is still more the domain of artists than of technicians, as only incomplete heuristics have been articulated, and the mechanization of these is still part of artificial intelligence. In contrast, much more is known about ways of checking the correctness of non-interference assumptions. There are two issues here: one of how to state what a machine is intended to compute, as opposed to how it is supposed to compute, and one of how to relate the intentions for sub-machines to intentions for the whole machine.

§**26.** Although each computer program is a precise set of instructions written in an interpreted formal language, most programming languages provide no formal means for restating what the program's instructions are intended to accomplish. Most programming languages provide only a commenting facility, with which the programmer can attach to lines of program text comments in English (or some other natural language) to indicate what the individual program instructions mean in the larger scheme of the program. Unfortunately, these comments do not have equal force of specification as the program text, but instead are completely non-operative. In part this is due to the informal language of comments, to the lack of a formal language within the programming language for stating comments. The result is that the program forms a uniquely privileged specification of the machine realizing it, the only description that really matters to the operation of the machine. Some programming methodologists abhor the discretion this privileged status gives programmers in deciding whether or not to document their programs, but the number of machine descriptions does not seem to be the crucial issue. While undocumented programs are often odious, the underlying problem seems one of lack of force of specification (either to the computer or to programmers) rather than one of discretionary use.

To remedy these deficiencies of programming languages, computer scientists have developed numerous formal specification languages, logical metalanguages of programming languages with which one can state the intended effects of a program (its ecological specifications) and the intended roles of program components in program operation (its internal functional specifications). The ecological specifications for the program as a whole connect to the internal functional specifications by means of ecological descriptions of elementary program instructions. For example, one common type of instruction operates on certain data-structures so as to mimic arithmetic operations on numbers. Since almost everyone thinks of these instructions as arithmetic operations, or uses them for other purposes by means of arithmetic encodings, the axiomatizations of these instructions usually state the effects of the instructions' execution in terms of arithmetic operations on numerically interpreted data-structures. Similarly if computers had instructions mimicking the operations of sewing machines on cloths, one might axiomatize these instruction in terms of sewing operations on cloth-interpreted data-structures. Other instructions call for further fabrication of interpreting axioms.

§**27.** Complementing the notion of formal specification of machines is the notion of verification of machine specifications, in which one checks that the specification of a composite machine follows from the specifications of the sub-machines and properties of the combination method. That is, one asks whether the ecological specifications of program operation follow from its functional specifications plus the ecological axiomatizations of instruction execution effects. Verifications can be approached syntactically or semantically, either by giving a logic of programs for formally deriving relations between specifications, or by giving a model which simultaneously satisfies all the specifications. While the syntactic approach is more immediately amenable to mechanization (and many attempts at mechanization populate the literature), the semantic approach is more fundamental, since a theory of models must underlie any logic of programs. Models of external domains like arithmetic, symbol strings, cloths, etc., present no difficulties peculiar to computer science, but many instructions operate on the internal components of machine states, affecting by their actions the meaning of the instructions themselves.[33] Models for these instructions are much more complex and much less familiar than everyday models involving ships and sealing wax. This unfamiliarity may be part of why programming seems so hard to teach; in any case, this complexity presented serious obstacles to development of satisfactory models for program specifications. However, much progress has been made, the most striking advance being the models for type-free functional programming languages developed by Scott, Plotkin, and others.[34] These developments make doubly appropriate the term "functional specification," since they allow functions to be elements of the domains of models as well as parts of the relational structures of the models.

Functional specifications in artificial intelligence

§**28.** While formal specifications fill the literature and textbooks of computer science, on first glance they seem almost totally absent from the artificial intelligence literature, their appearance there being restricted to attempts at mechanizing the problem decomposition

[33] See [Elgot and Robinson 1964] for instructive models of RASP's in these terms.
[34] See [Scott 1973], [Plotkin 1972], [Barendregt 1981].

and specification verification processes for "automatic" (i.e., machine performed) programming. In actuality, however, formal specifications do play a significant role in artificial intelligence, but a very different role from that played in computer science. This difference in role stems partly from the nature of artificial intelligence research, where problems of formulation play so great a role. If the primary purpose of writing a program is to increase one's understanding of the psychology realized by the program, verifying the specifications of the program becomes a pointless activity compared to reformulating the psychology and the implementation to accord with the insights gained in the experiment. Most programs written in artificial intelligence are not meant to be solutions, but are meant to be rungs on the ladder of understanding, rungs which allow one to progress, but rungs to be discarded after use. This is quite a contrast with the usual situation in computer science, where the focus is on better algorithms for tasks with stable formulations. But however important this difference in the purposes of programming in the two fields, it is dwarfed by the difference between the audiences of specifications in the two fields.

§**29.** In computer science, the study of machine specifications focuses on how a human programmer can think about programs and their use. This means that the principal contribution of formal specifications and semantics is a way of interpreting the structure of machines in terms of their real-world meaning. A hand calculator or slide rule is interesting only because we interpret the data-structures it manipulates as numbers, and because we interpret its operations on these data-structures as arithmetic operations. We do not care about changes in machine states, flip-flops, or bit patterns. We care about numbers and arithmetic. Similarly, we do not specify a bank's accounting machine in terms of patterns of bits and their manipulation, but in terms of customers, their deposits, their withdrawals, and their balances. The interpretations of interest are those relating the machine's state to the external world, and any discussion of the relations between machine states is merely part of a proof that the specifications of external interest are reflected in the structure and behavior of the machine. In our earlier language for describing psychological theories, the specifications of interest in most of computer programming are non-narrow, ecologically interpreted specifications which refer to circumstances external to the machine.

§**30.** In artificial intelligence, however, the study of machine specifications focuses on how the machine can think about itself. One key component of intelligence seems to be adaptiveness, the ability of the agent to change itself or its surroundings when it so desires. Mundane adaptiveness involves, for example, the agent updating its beliefs to reflect the effects of its actions, to accommodate new information, or to adopt a stand on some question. Similarly, the agent might adapt by changing other mental structures, such as its desires and intentions (resolving inconsistencies, adopting stands, etc.) or its skills. In all these examples, the agent is acting as the designer of its new state, as its own programmer.

As its own programmer, the agent needs some way of guiding its adaptations, some way of stating its intentions about the relations of parts of its mental structure at single instants and over time so that it can modify itself to satisfy these intentions. But this is just the problem described above facing any programmer, one addressed by functional specifications relating parts of mental states to others at the same time, to others at other times, or to the mental state as a whole. The agent modifies itself so that its new state still satisfies (is a model of) the set of self-specifications. But since the machine is doing this revision itself with only its current state to work with, the machine's interpretation of these self-specifications cannot refer to external affairs, but must be narrowly interpreted specifications referring only to parts of the agent itself. In classical terminology, artificial intelligence programs are not merely rule-obeying machines, but are rule-following machines, interpreters of the rules, and as such must be able to grasp the import of their rules.

Examples of such self-specifications are abundant in artificial intelligence programs. Perhaps the clearest example is that of reason maintenance systems, also called truth maintenance, belief revision, or dependency network systems.[35] In a machine based on reason maintenance techniques, the fundamental sort of self-specification states that if the current state of mind contains certain components and lacks certain other components, then the current state of mind should also contain some further component. These specifications are termed *reasons* or *justifications*. The mental components related by reasons may be mental attitudes like beliefs, desire, etc.; descriptions, procedures, etc.; or whatever the psychology of the agent employs as building blocks for mental states. The agent follows these self-specifications by deriving its current mental state from the current set of reasons, using groundedness principles to construct a set of mental components satisfying all the reasons (i.e., by falsifying an antecedent or by including the conclusion). The agent's thinking and acting may change the state of mind by adding new reasons to or subtracting reasons from the current set, or by switching to another model of the current set. In each of these cases, the machine revises its current mental state by using the reasons as guides to what mental components should be adopted or abandoned.

Reasons are not the only sort of narrowly interpreted self-specification employed in artificial intelligence. The least complex self-specifications are those simply declaring the existence of some component of mental structure. Most of the declarations of the so-called "knowledge representation languages" take this form, asserting the existence of a belief (a statement in a database), of a desire (a goal statement), or of an intention (an item on an agenda). More complex self-specifications relate two or more components of states of mind, including reference or coreference relationships (procedural attachments and equalities).[36] The structure sharing, inheritance, or "virtual copy" relationships so common in representational systems are simply self-specifications stating that one description should be considered as having certain components if other descriptions have those components as well.[37] Likewise, MINSKY'S "K-lines" can be viewed as self-specifications stating that the state of mind should contain one subset of (K-node) components if it also contains the corresponding enabling (K-node) component.[38] In fact, much of what goes by the name of "self-description" in

[35] See [Stallman and Sussman 1977], [Doyle 1979], [London 1978], [McAllester 1980], [Doyle 1983a], and [de Kleer 1986].

[36] See, for example, [Weyhrauch 1980] and [Sussman and Steele 1980].

[37] See, for example, [Fahlman 1979] and [Doyle 1983c].

[38] [Minsky 1980]

artificial intelligence is not merely descriptive but instead normative, and so properly viewed as self-specification rather than self-description.[39] Self-specifications have also begun to make appearances in computer science more generally, the best examples being the "integrity constraints" of database theory[40] and N. MINSKY'S approach to program design based on formally expressed and automatically maintained "laws" of the system.[41]

The concept of mind

Convergence of the theories

§**31.** We can now not help but see a convergence of ideas between philosophy of mind and artificial intelligence. Philosophy began with the idea that psychological theories describe the mind in ecological terms, but abandoned that view in favor of the idea of narrow psychological theories since ecological facts offer little predictive power about behavior, even when they are accurate descriptions of mental states. On the other hand, the theory of machines arrived somewhat earlier at the notion of effectiveness, a stronger notion entailing narrowness. Early on, descriptions of machines (either abstract or programmed) were effective, hence narrow, theories of machines states and behavior. But effectiveness was so strong as to obscure matters. Effectiveness of machine description implies rather direct physical realizability: this fact lies at the heart of the practice of programming. Because of their privileged status in these direct realizations, programs became identified with the machines realizing them. Considerable effort was required to regain the perspective of abstract, general, and multiple theoretical specifications of machines, as opposed to unique or privileged machine descriptions. In computer science, the recovered perspective proved crucial in ecological specifications of machines in their environment of application. In artificial intelligence, the recovered perspective proved crucial in the design of adaptive machines reconstructing themselves by means of narrow self-specifications. We attribute narrowness rather than effectiveness to the self-specifications of artificial intelligence because often these specifications were introduced as ideal (but narrow) specifications only approximated in their interpretation by an accompanying effective algorithm. While philosophy came to narrowness seeking predictive power, artificial intelligence came to narrowness seeking adaptive power.

[39] See [Doyle 1983d] for a semantical approach to self-specifications, and [Doyle 1988] for a treatment of self-specifications as "mental constitutions."

[40] [Reiter 1988]

[41] [Minsky 1988]

Possible minds defined

§**32.** The two paths to narrowness of psychological theories suggest the importance of the idea for rational psychology. It might seem that the motivation for adaptiveness subsumes the motivation of predictive power, since the rationality of a particular adaptation involves the agent's expectations about the effects of the changes. But in fact the motivations are separate, since many of the adaptation applications leading to the idea of narrow self-specification need not be deliberate or considered adaptations, but may be automatic reorganizations choosing some possible adaptation without regard to comparative advantages or disadvantages. These converging motivations mean that while any psychological theory, narrow or not, may be of interest in ecological studies of minds in their environments, only narrowly realized psychological theories need matter to rational psychology. We draw on the apparent significance of this idea to turn it around and say that *all* narrowly realized psychological theories matter to rational psychology, that the set of such realizations forms the set of possible minds.

§**33.** Defining the set of possible minds as the set of narrowly realized theories has several advantages for rational psychology, advantages of neutrality on several important questions. This neutrality permits the use of these questions as dimensions for classification rather than as presuppositions of the science. Specifically, the definition is neutral on questions of psychological ontology, complexity, effectiveness, and determinateness.

The definition is neutral on psychological ontology because as long as the theory has some realization, we may posit any mental entities we desire. For example, one can express narrow stimulus-response psychologies in terms of relations between sensors and effectors; attitudinal psychologies in terms of relationships between beliefs and desires (or whatever attitudes one chooses); Freudian psychologies in terms of ego, superego, id, energy, and flows; and even the monolithic states of finite state machines. Thus the first dimension of classification of minds is by the mental components, by psychological ontology.

The definition is neutral on psychological complexity for reasons similar to its neutrality on psychological ontology. As long as the theory is narrowly realizable, it is a possible psychology no matter how trivial or complex it is. Possible minds may be as simple as a soda machine, or as complex as Lev Tolstoy. Thus the second dimension of classification of minds is by their structural complexity, by the variety of ways their components may be combined to form mental states.

The definition's neutrality on effectiveness follows since effectiveness is a stronger notion than narrowness. The range of possible minds includes both mechanically realized psychologies as well as physically realized psychologies which might not admit mechanized realizations. The definition allows for universes in which the notions of narrowness and effectiveness coincide, and for universes like our own in which narrowness subsumes effectiveness.

Moreover, it allows for universes in which effectiveness is even more restricted than in our own. This neutrality opens a whole range of new questions for physical theorists. One might conjecture from the amazing coincidence of effective calculability and recursive computability in our universe that other universes might offer analogous coincidences of effectiveness with the degrees of recursive unsolvability or with the degrees of computational complexity.[42] What form would the physics of such universes take? Need they have different laws, or might they differ from ours only in the values of the fundamental constants?

Finally, the definition of possible minds is neutral on the question of determinism, admitting both theories in which a mental state may have at most one possible successor state and theories in which a mental state might be followed by any of several others. Deterministic and non-deterministic Turing machines are examples of these. Of course, even non-deterministic psychologies can have deterministic realizations, and the existence of non-deterministic realizations is a question of physics.

§**34.** Defining possible minds to be narrowly realized theories has a seemingly unavoidable and perhaps unwelcome consequence in addition to the previously discussed advantages. The neutralities on psychological ontology, complexity, effectiveness, and determinateness add up to a neutrality on realizations. According to the definition, any satisfiable theory is a possible psychology, and by taking the entire domain of the model to be an agent, any realized theory is a possible mind. This makes the idea of narrowness of theoretical realizations analogous to the idea of closed systems in physics, and with similar importance, the laws of physics being taken to describe all and only closed systems. The generality of the definition means that any object and an accurate internal description of it constitute a possible mind. For example, supposing the known laws of physics correct, the universe is a possible mind, with the laws of physics as its psychology. Similarly, the U.S. economy and a correct economic theory constitute a possible mind, as does *Hamlet* together with a correct descriptive analysis, as do the natural numbers together with PEANO'S axioms (assuming their correctness). If the aim of rational psychology is to classify all possible minds, then it includes as subdisciplines not just psychology and artificial intelligence, but also physics, theology, model theory, sociology, etc., etc., etc. If rational psychology is so all embracing, what endeavor is *not* rational psychology?

My own feelings are mixed about this problem. On the one hand, I find the seeming intellectual imperialism of this view distasteful, but on the other hand I can offer some perspectives from which it seems less so, even natural. First, the generality of the definition does no real harm, for the point of the science is to introduce distinctions, and the first action of anyone interested in thinking would be to introduce distinctions of domains so as to reinstate the disciplines in their traditional fiefs. Indeed, most disciplines have ways of viewing the whole world from their perspective, but modesty and common sense keep them from overstepping their most fruitful bounds.

Second, the generality seems unavoidable if one wants the field to encompass both the

[42] [Rogers 1967], [Aho, Hopcroft, and Ullman 1974], [Garey and Johnson 1979]

trivial mind (e.g., the soda machine) and physically realized but non-effectively realizable minds. Bizarre and pointless minds need not be very interesting to anyone, but the scope of the science must include them precisely so that terms like "bizarre" and "pointless" can be substantively applied. Group theory has its "monster" groups: rational psychology needs its "psychopaths." It is worth emphasizing that rational psychology *requires* this generality of domain, lest restrictions on its freedom to consider possible psychological structures prevent it from discovering the best mathematical forms of psychological concepts. This intended neutrality on questions of what constitutes human thinking sharply distinguishes the notions of possible mind appropriate for rational psychology from the notions of human understanding and mental states involved in SEARLE'S critique of the "strong AI" view that minds are programs.[43] From the perspective presented here, artificial intelligence is not about programs, but about theories of machines, about narrowly realized psychological theories of machines, whether human or otherwise, and about whether humans are indeed machines.

Third, and finally, there is a hidden, underlying rightness to the definition. All possible minds must incorporate all possible mental structures, of that there can be no doubt. Yet what are the mathematical structures of physics, the logical theories of model theory, the programs of computer science, the constitutions of governments, and the articles of the faiths if not ideas in the minds of men?

Conclusion

§**35.** Cognitive psychology and artificial intelligence currently dominate cognitive science, and with them come the dominant methodologies of experimental data collection and speculative construction. If the arguments of this paper are understood, practitioners of these fields will at least have been introduced and perhaps converted to a methodological viewpoint that begins with the set of all possible minds and proceeds by formulating distinctions and classifying minds. The new viewpoint does not reject, but subsumes, the previous viewpoints since experiment and speculation remain useful in the problems of formulation and classification.

Even if one remains attached to one's familiar methodology and hesitant before the general conception of mind introduced above, I think important practical benefits follow from approaching the study of mind in the proposed way. I am not familiar enough with the literature of cognitive psychology to promise cognitive psychologists these benefits, but I am familiar with artificial intelligence, and feel the intelligibility of its literature might be increased by adoption of the proposed viewpoint: not just intelligibility to outsiders, but to insiders as well. One of the commonly acknowledged problems in artificial intelligence is the difficulty of telling what someone else has done. Almost every researcher has his own vocabulary and viewpoint, and while the gist of papers is usually intelligible, the details present

[43][Searle 1980]

more difficulties because one person's omitted explanatory trivialities may be someone else's stumbling blocks. While the classificatory viewpoint cannot by itself reconcile vocabularies or world-views, consciousness of the task of classification might strengthen incentives for scholarly analysis of comparable works. Moreover, the classificatory viewpoint also makes it possible to state and infer what results are scientific results and what results are engineering advances. The scientific aim of artificial intelligence is a question of formulation and existence: whether or not there are interesting psychologies realizable in Turing-equivalent machines. The answer to this question will likely involve an actual realization. The details of this realization are not scientifically interesting, but instead are matters of engineering. This is not meant to belittle engineering problems, merely to distinguish two endeavors. The situation is similar to that in mathematics, where in the context of a particular theory, an individual proof is without mathematical significance except for its conclusion. Of course in a wider context the proof will have importance as an example of a method for discovering analogous proofs. Likewise, the details of an intelligent machine's construction will be answers to important engineering questions, perhaps useful in constructing realizations of other psychologies, but scientifically insignificant as far the answered existential question is concerned. This distinction between scientific and engineering questions allows clearer explanations of the problems and results reported in artificial intelligence. Is the purpose of a paper to show how to realize feature f of the psychology ψ currently being worked toward (a scientific advance)? Or is it to show ways in which feature f of the machine M might be utilized independently of any particular psychology (an engineering advance)? Is the paper's purpose to modify the target psychology in light of changing estimates of feasibility of realizations? Or is it to modify the underlying machine so as to enhance feasibility of realization? The first two of these questions serve to mitigate the apparent scientific irrelevance of papers on programming techniques and programming systems by viewing them as engineering advances rather than psychological theories. Likewise, the second two questions serve to clarify the revolutionary claims often made in the literature. Papers changing the target psychology or underlying machine change particular scientific aims. Papers inventing a new programming technique do not, but advance the progress of the scientific or engineering investigation of a particular set of scientific aims. Artificial intelligence may well be yet a field most concerned with problems of formulation, but let us at least make clear what is being formulated and studied, namely pairs of psychologies and machines.

§**36.** The proposed perspective on the aims of artificial intelligence can benefit the clarity and ease of construction of programs as well as the clarity and construction of their exegeses. The advantages of initial specification and subsequent implementation are well-known from programming methodology in computer science, even if intermediate specifications and verifications of implementations rarely matter in artificial intelligence's problems of formulation. But the proposed perspective can have a far more substantial impact on the construction of artificial agents than simply as an explanatory aid. Earlier, we introduced the idea that intelligent machines can interpret narrow self-specifications to guide their adaptations. If we take this suggestion seriously, then designing machines to continually construct and interpret sets of self-specifications becomes a powerful aid to programming. In this methodology, intelligent machines are always designed to construct and update their mental states by interpreting their self-specifications. Reason maintenance systems were mentioned earlier as an example of this idea, but so are the constraint-based programming systems currently enjoying interest. Constraints are nothing more than self-specifications, and in programming systems based on them one simply describes the psychology of the program as a set of narrowly interpreted logical statements, which the programming system examines to maintain the current mental state. Of course, in artificial intelligence there must be not just narrow, but effective and feasible ways of interpreting the specifications, so that one repeatedly reformulates the specifications of the psychology as one's store of feasible interpretation techniques changes. For example, the extant constraint-based programming systems fantastically restrict the complexity of psychological specifications because they incorporate means for interpreting only a few of the very simplest specifications. But these restrictions are not essential, and future systems might allow incremental addition of interpretation techniques for specifications of more complex logical form.

The proposed methodology can be expressed as the thesis of self-interpretation: *Thinking is a process of narrow self-specification and self-interpretation.*[44] That is, mental actions are all described and realized by means of self-interpretive psychological theories which augment themselves with new self-specifications, purge themselves of unwanted self-specifications, or reinterpret their existing self-specifications. In this way mental actions are understood in terms of the psychology itself (e.g., reasons, beliefs, desires, etc.), rather than in terms of the realization of the psychology (e.g., CONS's, RPLACA's, etc.).

Acknowledgements

I owe much to JOSEPH SCHATZ for advice on this paper. I also benefited from discussions with NED BLOCK, JOHAN DE KLEER, MERRICK FURST, ALLEN NEWELL, DANA SCOTT, and RICHMOND THOMASON. Preparation of this paper was supported by National Institutes of Health Grant No. R01 LM04493 from the National Library of Medicine.

[44] This thesis is a descendant of the comparable thesis about reasoning proposed in [Doyle 1979].

References

1. Aho, A. V., J. E. Hopcroft, and J. D. Ullman, 1974. *The Design and Analysis of Computer Algorithms*, Reading: Addison-Wesley.

2. Backus, J., 1978. Can programming be liberated from the von Neumann style? A functional style and its algebra of programs, *Communications of the Association for Computing Machinery* **21**, 613-641.

3. Barendregt, H. P., 1981. *The Lambda Calculus: its syntax and semantics*, Amsterdam: North-Holland.

4. Boring, E. G., 1950. *A History of Experimental Psychology* (2nd ed.), New York: Appleton-Century-Crofts.

5. Bretano, F., 1874. *Psychologie vom empirischen Standpunkte*, Leipzig.

6. Chomsky, N. A., 1965. *Aspects of Syntax*, Cambridge: MIT Press.

7. de Kleer, J., 1986. An assumption-based TMS, *Artificial Intelligence* **28**, 127-162.

8. Descartes, R., 1637. Discourse on the method of rightly conducting the reason and seeking for truth in the sciences, in *The Philosophical Works of Descartes* (E. S. Haldane and G. R. T. Ross, trans.), V. I, Cambridge: Cambridge University Press, 1931.

9. Dennett, D. C., 1978. Intentional systems, *Brainstorms*, Montgomery, Vermont: Bradford Books, 3-22.

10. Doyle, J., 1979. A truth maintenance system, *Artificial Intelligence* **12**, 231-272.

11. Doyle, J., 1983a. Some theories of reasoned assumptions: an essay in rational psychology, Pittsburgh: Department of Computer Science, Carnegie-Mellon University, report 83-125.

12. Doyle, J., 1983b. What is Rational Psychology? Toward a modern mental philosophy, *AI Magazine*, V. 4, No. 3, 50-53.

13. Doyle, J., 1983c. A society of mind: multiple perspectives, reasoned assumptions, and virtual copies, *Eighth International Joint Conference on Artificial Intelligence*, 309-314.

14. Doyle, J., 1983d. Admissible state semantics for representational systems, *IEEE Computer*, V. 16, No. 10, 119-123.

15. Doyle, J., 1988. Artificial intelligence and rational self-government, Pittsburgh: Department of Computer Science, Carnegie Mellon University, TR CMU-CS-88-124.

16. Elgot, C. C., and A. Robinson, 1964. Random access stored program machines, *Journal of the Association for Computing Machinery* **11**, 365-399.

17. Fahlman, S. E., 1979. *NETL: A system for representing and using real world knowledge*, Cambridge: MIT Press.

18. Feigl, H., 1958. The 'mental' and the 'physical,' in *Minnesota Studies in the Philosophy of Science* (H. Feigl, M. Scriven, and H. Grover, eds.) V. II, Minneapolis: University of Minnesota Press, 370-497.

19. Fodor, J. A., 1968. *Psychological Explanation: An introduction to the philosophy of psychology*, New York: Random House.

20. Fodor, J. A., 1975. *The Language of Thought*, New York: Crowell.

21. Fodor, J. A., 1981. Methodological solipsism considered as a research strategy in cognitive psychology, *Representations: Philosophical essays on the foundations of cognitive science*, Cambridge: MIT Press, 225-253.

22. Frege, G., 1884. *Die Grundlagen der Arithmetic*, Breslau: Verlag von Wilhelm Koebner, translated by J. L. Austin as *The Foundations of Arithmetic* (2nd revised ed.), Evanston: Northwestern University Press, 1968.

23. Frege, G., 1892. Über Sinn und Bedeutung, *Zeitschrift für Philosophie und philosophische Kritik* **100**, 25-50. Also, On sense and reference (M. Black, trans.), in *Translations from the Philosophical Writings of Gottlob Frege* (P. Geach and M. Black, eds.), Oxford: Basil Blackwell, 1952.

24. Freud, S., 1895. Project for a scientific psychology, *Standard Edition of the Complete Psychological Works of Sigmund Freud* (J. Strachey, tr. and ed.), V. I, London: Hogarth Press, 1966, 281-397.

25. Gandy, R., 1980. Church's thesis and principles for mechanisms, in *The Kleene Symposium* (J. Barwise, H. J. Keisler, and K. Kunen, eds.), Amsterdam: North-Holland, 123-148.

26. Garey, M. R., and D. S. Johnson, 1979. *Computers and Intractability: a guide to the theory of NP-completeness*, San Francisco: W. H. Freeman.

27. Goldstein, H. H., 1972. *The Computer from Pascal to von Neumann*, Princeton: Princeton University Press.

28. Hilbert, D., 1900. Mathematical Problems, lecture delivered before the International Congress of Mathematicians at Paris (M. W. Newson, trans.), *Bulletin of the American Mathematical Society* **8** (1902), 437-479.

29. Hintikka, J., 1971. Semantics for propositional attitudes, in *Reference and Modality* (L. Linsky, ed.), 145-167.

30. Kaplan, D., 1977. *Demonstratives*, manuscript.

31. London, P. E., 1978. Dependency networks as a representation for modelling in general problem solvers, Department of Computer Science, University of Maryland, TR-698.

32. McAllester, D. A., 1980. An outlook on truth maintenance, MIT Artificial Intelligence Laboratory, Memo 551.

33. Minsky, M., 1967. *Computation: Finite and infinite machines*, Englewood Cliffs: Prentice-Hall.

34. Minsky, M., 1980. K-lines: a theory of memory, *Cognitive Science* **4**, 117-133.

35. Minsky, N. H., 1988. Law-governed systems, unpublished manuscript, Rutgers University, Computer Science Department.

36. Peters, R. S., and C. A. Mace, 1967. Psychology, in *Encyclopedia of Philosophy* (P. Edwards, ed.), New York: Macmillan and Free Press, V. 7, 1-27.

37. Place, U. T., 1956. Is consciousness a brain process?, *British J. of Psychology* **47**, 44-50.

38. Plotkin, G., 1972. A set-theoretical definition of application, School of Artificial Intelligence, Memo MIP-R-95, University of Edinburgh.

39. Putnam, H., 1975a. The meaning of 'meaning,' *Mind, Language, and Reality*, Cambridge: Cambridge University Press, 215-271.

40. Putnam, H., 1975b. Minds and machines, *Mind, Language, and Reality*, Cambridge: Cambridge University Press, 362-385.

41. Putnam, H., 1975c. Philosophy and our mental life, *Mind, Language, and Reality*, Cambridge: Cambridge University Press, 291-303.

42. Quine, W. V., 1970. *Philosophy of Logic*, Englewood Cliffs: Prentice-Hall.

43. Randell, B., (ed.), 1975. *The Origins of Digital Computers: Selected papers*, Berlin: Springer-Verlag.

44. Rogers, H. Jr., 1967. *Theory of Recursive Functions and Effective Computability*, New York: McGraw-Hill.

45. Ryle, G., 1949. *The Concept of Mind*, London: Hutchinson.

46. Scott, D., 1973. Models for various type-free calculi, *Logic, Methodology and Philosophy of Science IV* (P. Suppes, L. Henkin, A. Joja, Gr. C. Moisil, eds.), Amsterdam: North-Holland.

47. Searle, J. R., 1980. Minds, brains, and programs, *Behavioral and Brain Sciences*, **3**, 417-424.

48. Shepherdson, J. C., and H. E. Sturgis, 1963. Computability of recursive functions, *Journal of the Association for Computing Machinery* **10**, 217-255.

49. Skinner, B. F., 1957. *Verbal Behavior*, New York: Appleton-Century-Crofts.

50. Smart, J. C. C., 1959. Sensations and brain processes, *Philosophical Review* **68**, 141-156.

51. Stallman, R. M., and Sussman, G. J., 1977. Forward reasoning and dependency-directed backtracking in a system for computer-aided circuit analysis, *Artificial Intelligence* **9**, 135-196.

52. Sussman, G. J., and G. L. Steele Jr., 1980. CONSTRAINTS—A language for expressing almost-hierarchical descriptions, *Artificial Intelligence* **14**, 1-39.

53. Thomason, R. H., 1979. Some limitations to the psychological orientation in semantic theory, unpublished manuscript, University of Pittsburgh.

54. Truesdell, C., 1966. *Six Lectures on Modern Natural Philosophy*, Berlin: Springer-Verlag.

55. Turing, A. M., 1936. On computable numbers with an application to the entscheidungsproblem, *Proc. London Math. Soc. Ser. 2* **42**, 230-265.

56. Watson, J. B., 1914. *Behavior: An introduction to comparative psychology*, New York: Holt.

57. Weyhrauch, R. W., 1980. Prolegomena to a theory of mechanized formal reasoning, *Artificial Intelligence* **13**, 133-170.

58. Whitehead, A. N., and B. Russell, 1910. *Principia Mathematica*, Cambridge: Cambridge University Press.

59. W. Wundt, 1874. *Grundzüge der physiologischen Psychologie*, Leipzig: W. Engelemann. Also *Principles of Physiological Psychology* (E. B. Titchener, trans.), New York: Macmillan, 1904.