Reset reproduction of CMU Computer Science report CMU-CS-83-106. Abbreviated version published in *AI Magazine*, V. 4, No. 3 (1983), pp. 50–53. Reprinted July 1994. Current address: MIT Laboratory for Computer Science, Cambridge, Massachusetts.

What is Rational Psychology?

Toward a modern mental philosophy

by Jon Doyle
at Carnegie-Mellon University

November 1982–January 1983
Pittsburgh, Pennsylvania

January 12, 1983

Abstract: We introduce rational psychology as the conceptual investigation of psychology by means of the most fit mathematical concepts. Several practical scientific benefits should accrue from its recognition.

This research was supported by the Defense Advanced Research Projects Agency (DOD), ARPA Order No. 3597, monitored by the Air Force Avionics Laboratory under Contract F33615-81-K-1539. The views and conclusions contained in this document are those of the author, and should not be interpreted as representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the Government of the United States of America.

Some years ago — never mind how long precisely — having little or no money in my purse, and nothing particular to interest me on shore, I thought I would sail about a little and see the watery part of the world. It is a way I have of driving off the spleen and regulating the circulation.

H. Melville, *Moby Dick, or The Whale*

Peregrinus expectavi pedes meos in cymbalis.

S. Prokofiev, *Alexander Nevsky*

Apology

§**1.** This paper explains "rational psychology," the field in which I place my current work. This is not a scientific paper, but an apology: a necessary one, I think, however distasteful. My current work begins in artificial intelligence, a recognized, if infant, field. However, I cannot merely address the work to artificial intelligence because my aims and conceptions are somewhat different than those ordinarily held in artificial intelligence, and so are easily misunderstood without some sort of explanation. I would instead address the work to the field whose aims I share, except that I find no such field at present, and face similar obstacles to communication in the alternatives to artificial intelligence. In this paper, I hope both to explain the idea of rational psychology, and to place it among the associated fields of artificial intelligence, psychology, cognitive science, etc. The reader is free to reject my introduction of this enterprise, but I hope at least to communicate my aims, whatever fields lay claim to them.

§**2.** I was trained in the field of artificial intelligence, initially working with issues of problem solving and planning, but many workers in those areas took my work to focus on representational issues. If I turned to workers in representation, I found myself thought a problem solver, since what I did was plainly not just representation (to the extent that anyone in artificial intelligence is willing to admit that anything is not representation). But neither did I study language, vision, or cognitive simulation. I began to seek congenial scholars outside artificial intelligence, but found no clear home there either. Philosophers knew that whatever my philosophical interests, the questions I pursued were ones of engineering. Turning back to artificial intelligence, the "knowledge engineers" and system builders saw I was no engineer, but a logician instead. Logicians thought it quaint one of the artificial intelligentsia should appreciate something substantial like logic, but knew my concerns were extra-logical, instead psychological. Psychologists concluded from my disinterests in the human mind that I was a computer scientist. Those computer scientists who had not already written off artificial intelligence could not understand my disinterest in writing programs. After all, is that not the point of artificial intelligence?

As Truesdell remarks, it is pleasant to afford the luxury of being an eccentric, but no one likes to be nothing. To combat my apparent isolation, I worked to more clearly formulate the aims of my researches, to see if they were rightly "something else" to everyone. I wish to be clear that the people implicitly referred to previously are all generous and friendly, and are all correct in their judgements of the dissimilarity of our respective aims. But due to my inarticulate explanations of my work, they each concluded my work to fit where others would not have it. I benefited greatly from this string of rejections in refining my views on my aims, and I thank all those involved. While my prior training as a mathematician and the counsel of Joseph Schatz led me on the first steps of this formulation, happily I eventually read works of Truesdell, whose aims for mechanics coincided remarkably with my aims for mental concerns. I have borrowed heavily from the lessons of Schatz and the writings of Truesdell in the following, to the point of adopting the term "rational psychology" in correspondence with the name "rational mechanics." I explain this choice momentarily, together with the enterprise it labels.

§**3.** In the following I first describe the concerns, aims, and methods of rational psychology, together with its connections to associated fields. This precedes discussion of some practical benefits I hope will follow from its recognition. Better exposition perhaps would begin with the problems I perceived in the practice of artificial intelligence that motivated my pursuit of rational psychology, for these might serve as a better bridge to the new from the familiar, but rational psychology is an enterprise in its own right, not merely a temporary therapeutic project. However, as I elaborate later, whatever the proper stature of rational psychology, the points to grasp are its aims and methods. These are important regardless of how one labels the fields they serve.

§**4.** I owe much to many who must remain nameless for unwitting help in formulating these views. However, I am pleased to thank several for special help: Joseph Schatz, Johan de Kleer, Gerald Sussman, Jaime Carbonell, Allen Newell, Merrick Furst, Raymond Reiter, and John MacNerney for several discussions, and Marvin Minsky, Clifford Truesdell, Juris Hartmanis, Dana Scott, and Michael Rabin for their many writings. I am also indebted to the institutions which have supported my pursuits: the Massachusetts Institute of Technology, Stanford University, Carnegie-Mellon University, the Fannie and John Hertz Foundation, and the Government of the United States of America.

Rational Psychology

§**5.** Rational psychology is a part of mathematics, the conceptual investigation of psychology. "Rational" here indicates psychological investigations based on reason alone, rather than on experiment, engineering, or computation, the rational analysis of the concepts and theories whose applicability and feasibility are studied in experimental, engineering, and computational projects. Rational psychology is not the study of rational agents, but instead the mathematical approach to the problems of agents and their actions, whether these agents and actions are themselves thought rational or irrational. The name stems from the rational mechanics of Newton, and is merely adaptation to the realm of mental philosophy of the principles, aims, and methods found in his natural philosophy. Although I contrast rational psychology with other disciplines, the term is not meant to exclude others so much as to name something excluded by everyone else, and to highlight the common project occurring in specialized and isolated manifestations. It is not meant merely to agglomerate numerous disciplines, as unfortunately seems to be the situation in cognitive science, nor to prevent specialization. The aim is instead to reset the common foundations of mental fields to make the unity apparent mathematically while aiding the prosecution and communication of specialized inquiries.

The reader is warned that this enterprise involves a different conception of what is meant by "mind," "mental," and "psychology" than that common in the existing mental sciences. For me, a psychology is merely a specification of the structure and behavior of some agent, and a mind is the realization of a psychology in an agent. I decouple these terms from any connotation of human minds or actual physical realizability, admitting as "possible minds" agents including vending machines and logically omniscient intelligences. These conceptions are developed at length in my paper *The Foundations of Psychology*.

§**6.** The aim of rational psychology is understanding, just as in any other branch of mathematics. Where most of what is labelled "mathematical psychology" consists of microscopic mathematical problems arising in the non-mathematical prosecution of human psychology, or in the exposition of informal theories with invented symbols substituting for equally precise words, rational psychology seeks to understand the structure of psychological concepts and theories by means of the most fit mathematical concepts

and strict proofs, by suspiciously analyzing the informally developed notions to reveal their essence and structure, to allow debate on their interpretation to be phrased precisely, with consequences of choices seen mathematically. The aim is not simply to further informal psychology, but to understand it instead.[1]

This aim entails classifying sorts of agents and actions, classifying all possible minds, so that the detailed properties of an agent may be predicted from its fundamental classifications. Just as group theory seeks to classify the set of all groups in terms of their isomorphism classes and their relations to other mathematical structures, rational psychology seeks to classify the set of all possible minds and their relations to possible environments. In either endeavor, a complete classification allows selection of standard representatives from each isomorphism class, representatives chosen to maximally facilitate their presentation and discussion. Put another way, rational psychology is one of the "sciences of the artificial," aiming to classify possibilities rather than to identify actualities. Classification can procede without metaphysical doctrine, and as COURANT and ROBBINS observe, some of the greatest achievements in physics have come as rewards for courageous adherence to the principle of eliminating superfluous metaphysics. One must have a metaphysics, but it can be chosen, as well as inherited.

§**7.** The method of rational psychology is to describe and study mental organizations and phenomena by the most fit mathematical concepts. This does not mean pursuit of the mathematical tools for their own sake, nor forced application of pet mathematical abstractions, but simply the use of a precise language instead of vague formulations, and the borrowing of whatever analyses the current mathematics provides. The standards directing the investigation are those of mental importance rather than difficulty of proof or abstruseness or mathematical importance of the mathematical tools employed. If a result is not psychologically crucial, the difficulty of its proof does not lend it importance, and neither does the use of mathematical esoterica. But if analytic function theory captures the properties of some agent more clearly than simple number theory, then it should not be shunned simply because of its relatively advanced position in mathematics curricula.

The method of rational psychology follows that of the mathematical study of mathematical concepts. One phrases subjects of investigation and specialized theories as sets of axioms about the constitution of agents. These are called "constitutive assumptions" in modern rational mechanics. Rational psychology takes psychologies as givens for analysis, classification, prediction, and reformulation, rather than as mysterious qualities of agents to be discovered by experiment, computation, or philosophical speculation. These sets of constitutive assumptions can be formulated and studied for many external purposes: as ideals against which actual or constructed agents may be compared; as theories of actual or desired agents in special circumstances; as special aspects of actual or desired agents; and as approximations to the properties of actual or desired agents. Clean theories of special cases may "leave things out," but they so trade restricted range of applicability for enhanced accuracy within their domain of interest.

[1] I cannot explain the mathematical aim of understanding, even though every mathematician grasps it. For hints, see HARDY or COURANT and ROBBINS.

Comparision

§**8.** I wish I could give a detailed sampling of important contributions to rational psychology, but that would make this a long textbook rather than a brief apology. Instead, I list some of the areas I would include as contributions. Only a tiny fraction of this work has occurred within artificial intelligence, and rightly so, for artificial intelligence is only one of the newest of the fields of mental philosophy. Perhaps artificial intelligence will someday attempt to equal its more accomplished sister-fields. At any rate, prominent among the areas with which the (ideal) student of rational psychology should be acquainted are (1) the sciences of rationality and rational agents, namely mathematical logic, metamathematics, and parts of mathematical economics (especially decision theory, game theory, utility theory, equilibrium theory, and social choice theory); (2) the sciences of mental representation and realizability, namely information theory, mathematical linguistics (both syntatical investigations and semantical studies), and the mathematical theory of computation; and (3) the sciences of mental ecology, for instance cybernetics and the new mathematical theories of perception. To these substantial theories, artificial intelligence contributes only a few smaller topics at present, such as the theory of perceptrons, search theory, and theories of reasoned assumptions. These topics are still at the beginnings of their development and integration with other areas. As a non-example of rational psychology I offer the theory of measurement. This theory appears prominently in texts on mathematical psychology, but is really no more relevant to psychology than to physics or demography. It supplies analysis of methodological questions and experimental procedure, but has little bearing on the nature of mental or physical entities. This does not reflect badly on the theory of measurement, any more than the irrelevance of ceramics to psychology reflects badly on ceramics.

§**9.** I build on this non-example of rational psychology to make the principal aims and methods of rational psychology clearer by contrasting them with the principal aims and methods of related fields. These brief characterizations are all somewhat unjust, for fields are populated by people with mixed interests; but they serve nevertheless to illustrate different emphases. To begin: the modern discipline of Psychology is the experimental investigation of human psychologies, with studies of other animals as paths to humans. Humans and experiment form the focus of Psychology, rather than all possible minds and mathematical analysis. The philosophy of mind, while employing conceptual (but typically not mathematical) analysis, also focusses on humans almost exclusively. In economics, where mathematical analysis has become standard, the focus is on rational agents, individual and collective, rather than on agents in general. Similarly, logic and metamathematics look to rationality, not general psychologies. Chomskyan linguistics is explicitly oriented toward the human mind, via the mechanism of language. The neurosciences are similarly both human- and mechanism-oriented. Cognitive science, to the extent that it admits a consensus, is an amalgamation of the human-oriented fields and artificial intelligence. Artificial intelligence itself, which from its name might seem the natural companion to the aims of rational psychology, is quite fragmented in aims, but almost universally oriented toward recursive realizability in agents using modern digital computers. Its subfield of cognitive simulation is explicitly human-oriented, and its subfields of formal reasoning, automated deduction, and "theorem proving" are all oriented towards issues of rationality rather than psychologies in general. "Reasoning" means deduction to almost all involved. The focus of the field on gaining insight from computational experience is valuable, for exact analysis always has current limits, but few pursue any exact analysis at all. Nevertheless, the enterprise of cognitive simulation is suspect: its practitioners think to study humans, but what do simulations of speculations tell about anything except the speculations? And given the difficulty of programming large systems, computational experience without much attempt at exact analysis certainly seems the slow boat to China.

Benefits

§**10.** By this point I have likely offended every reader beyond redemption, so comments seem in order on the motivations for and benefits of recognizing rational psychology. This paper is *not* a call for others to drop their own aims and methods and take up those of rational psychology. Instead, it simply describes a point of view that may be necessary to properly understand my substantive works. Differing fields have different aims and emphases, as is proper, and all are important in the pursuit of understanding. But even if their interests and methods overlap, recognizing their differences helps avoid misunderstandings and enmity. When as many issues and interests are at stake as in the fields of mental philosophy, being clear about the form and substance of one's claims is important. (Simply labelling works with disciplinary names is not only sometimes pointless, but occasionally even an impediment to communication.) For example, in artificial intelligence it is important to distinguish scientific from engineering claims. Claims about the properties and recursive realizability of psychologies are scientific claims, while the details of specific realizations and their relative "efficiencies" are claims of engineering. The former claims can remain significant to understanding no matter what engineering improvements are discovered, while yesterday's engineering compromises are sometimes best forgotten. Scientific advances are not necessarily "better" than engineering advances, for engineering advances often influence the areas of primary scientific interest as well as the immediate quality of life. But many arguments have appeared in artificial intelligence due to claiming scientific status for engineering techniques, which annoys people oriented toward science, and due to claiming immediate practical importance for theoretical developments, which irritates people primarily responsible for engineering achievements. It is not rare to find the programmer who claims to have solved the problem of mind (just use his program!), or the mathematician who claims to have discovered the structure of knowledge (modulo the programming details!). Both are equally guilty of arrogant claims no matter how outstanding their contributions appear when properly directed. This problem is larger than usual in artificial intelligence because unlike older fields, it has not yet divided into scientific and engineering branches. This means that conferences and journals must serve the entire audience, with completeness proofs boring the programmers and reinventions of Modus Ponens boring the theoreticians.

If nothing else, common consciousness about clarification of aims would relieve some personal aggravations. My work (for example on non-monotonic logic) has often been received with questions like "Have you implemented it yet?," "But what is your thesis?," "You don't really think humans do that, do you?" My work often consists of mathematical analyses of ideas. Asking about its implementation mistakes analysis for psychological proposal. One might as well ask if I have implemented the laws of physics recently. Asking for my thesis mistakes classificatory analysis for philosophical proposal. What is the thesis of VAN DER WAERDEN'S *Moderne Algebra*? And requiring that thinking be studied only in human terms is groundless chauvinism. These wrong-headed questions are not the fault of my questioners. Instead, they are the fault of inadequate cultural consciousness of and sensitivity toward the variety of aims with which one can approach psychological questions.

§**11.** In addition to these perhaps dubious hopes for intellectual amity, there are a number of far more practical benefits facilitated by rational psychology. The first of these is that of formal, precise statements of artificial intelligence problems, theories, and techniques. Formal specifications of program intent and proofs of program correctness are well-known in computer science. These concepts, though hardly a panacea, now allow concise and correct description of systems whose understanding previously required apprenticeship and experience. These exact formulations permit variations in problem and solution to be studied as technical questions rather than as banners in battles between methodologies and world-views. Mathematical formulation of concepts has hardly been prominent in artificial intelligence, with good reason. For the most part, complete ignorance prevails about the appropriate mathematical structures to employ in formulating psychological notions, and there is every reason to suspect that many new

mathematical notions must yet be invented in order to develop current informal psychological theories in precise terms. To draw a parallel, no matter how much one hoped to assign meanings to computer programs and their components, all early attempts to do so foundered on the reflexive nature of the domain of all computable functions, so that every proposal prior to SCOTT'S discovery of appropriate models was either obviously inadequate or of such complexity as to be of doubtful correctness. Unfortunately, for most of artificial intelligence, suitable mathematical tools are similarly undiscovered, so no matter what their standards when discussing computer science, many researchers find that doing artificial intelligence requires abandoning the usual crutches of confidence for wild and woolly adventures in intellectual hinterlands. Some never return to tell their tales, and some return speaking in tongues to the rue and mutterings of the stick-at-homes. Thus, I cannot recommend formal specifications to artificial intelligence as an immediate path to benefits. The way doubtless requires much toil to discover the appropriate concepts. But someday, it must be done.

§**12.** The second benefit rational psychology offers, even to the hard-core hacker, is savings in time and resources. Mathematics can be viewed as the science of avoiding unnecessary calculation, and rational psychology can be used as a way of avoiding some labors of programming and computation. It is commonplace in artificial intelligence research that systems are developed at costs of man-years and CPU-months, and when finished, their authors discover trivial examples of fundamental inadequacies and seemingly unmotivated limitations of abilities that to remedy would require the effort all over again. One cannot hope to discover all difficulties with a pet idea through thought alone, nor hope to avoid all unconscious intellectual blinders, but cultural practice in artificial intelligence calls for implementing ideas as sufficient means to "understanding" them. Often some inadequacies and tacit limitations come to light in this process, but diluted by months or years of wondering where the next CONS is coming from. Consider instead a cultural imperative which called for three weeks of pure critical thought and strict abstinence from computers prior to beginning any important implementation effort. The problems of artificial intelligence would not become any easier, but progress might be faster, since one might trade a week of analysis for a year of wasted programming. SOCRATES might well have said "The unexamined idea is not worth programming," and had the Athenians personal computers with LISP-controlled graphics they might well have sentenced him anyway. There is great contrast between the pleasures of programming and the tedium of analysis, between the challenge of the mysterious bug and the beautiful theory killed by ugly facts.

§**13.** The third benefit, and to some extent my motivation for writing this splenetic, is improved communications. The frequency of reinvention of ideas in artificial intelligence is legendary. While it is unreasonable to expect (and undesirable to attempt) to make reinventions rare occurrences, artificial intelligence clearly seems extravagant. It is not alone in this. There is the old joke in computer science about the result that was lost because it was only published four times. But even the magnitude of the problem is unclear. Not only do researchers lack deep understanding of their own proposals, but they usually cannot understand those of others either. This incomprehension is not due to stupidity, but to the vague, metaphorical terms on which the field relies in the absence of precise, formal vocabularies for presenting theories. In mathematics, physics, and many other sciences, papers, if properly written, define concepts in terms of the accepted vocabulary, state claims or discoveries, and then leave comprehension up to the intelligence and motivation of the reader. In artificial intelligence, even conscientiously written papers can be unintelligible no matter how capable and motivated the reader, for the accepted vocabulary is about as precise as that of poetry, and about as substantive as that of advertising copy. If we had adequate mathematical concepts, if we had conventions for clear, exact statements of problems — two large ifs — then we could hope for reduced reinvention, more rapid communication, comparison, and reproduction of ideas, and a true chance to build on the work of others: things all taken for granted in other fields.

Conclusion

§**14.** I have stated the aims, the methods, some examples, the distinction, and some benefits of rational psychology. The only remaining question is that of taste. Questions of taste are unusually prominent in all sciences of the artificial. Ask any mathematician, and his judgements of his own and of others tastes almost overshadow his mathematics. But I refrain from proposing any standards here since, as TRUESDELL observes, it is tasteless to recommend one's own taste, but scarcely honest to recommend any other. At most, I urge those struck by the above to examine their own aims, methods, and tastes, and then accept them consciously. There are many temptations today to sanction illegitimate arbiters of intellectual taste: witness the stampede of artificial intelligence toward immediate industrial applicability and respectability. It may be hard to call for milk in a crowded tavern, and the noise may make it difficult to bolster one's resolve, but the decision, whether to depart or to follow, should be conscious. At least I hope it will be so in science.

> There are certain queer times and occasions in this strange mixed affair we call life when a man takes this whole universe for a vast practical joke, though the wit thereof he but dimly discerns, and more than suspects that the joke is at nobody's expense but his own. However, nothing dispirits, and nothing seems worth while disputing. He bolts down all events, all creeds, and beliefs, and persuasions, all hard things visible and invisible, never mind how knobby; as an ostrich of portent digestion gobbles down bullets and gun flints. And as for small difficulties and worryings, prospects of sudden disaster, peril of life and limb; all these, and death itself, seem to him only sly, good-natured hits, and jolly punches in the side bestowed by the unseen and unaccountable old joker. That odd sort of wayward mood I am speaking of, comes over a man only in some time of extreme tribulation; it comes in the very midst of his earnestness, so that what just before might have seemed to him a thing most momentous, now seems but a part of the general joke. There is nothing like the perils of whaling to breed this free and easy sort of genial, desperado philosophy; and with it I now regarded this whole voyage of the Pequod, and the great White Whale its object.
>
> H. Melville, *Moby Dick, or The Whale*

Some readings for comparision and illustration

Abraham, R., and Marsden, J. E., 1978. *Foundations of Mechanics* (second edition), Reading: Benjamin.

Courant, R., and Robbins, H., 1941. *What is Mathematics? An elementary approach to ideas and methods*, London: Oxford University Press.

Debreu, G., 1959. *Theory of Value: an axiomatic analysis of economic equilibrium*, New York: Wiley.

Doyle, J., 1982. Some theories of reasoned assumptions: an essay in rational psychology, Pittsburgh: Department of Computer Science, Carnegie-Mellon University.

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

Hardy, G. H. 1940. *A Mathematician's Apology*, Cambridge: Cambridge University Press.

Hartmanis, J., 1981. Remarks in "Quo Vadimus: computer science in a decade," (J. F. Traub, ed.), *Communications of the ACM* **24**, 351-369.

Hartmanis, J., 1981. Observations about the development of theoretical computer science, *Annals of the History of Computing* **3**, 42-51.

Jauch, J. M., 1968. *Foundations of Quantum Mechanics*, Reading: Addison-Wesley.

Marr, D., 1982. *Vision*, San Francisco: W. H. Freeman.

Minsky, M., 1962. Problems of formulation for artificial intelligence, *Proc. Symp. on Mathematical Problems in Biology*, Providence: American Mathematical Society, 35-46.

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

Minsky, M., and Papert, S., 1969. *Perceptrons: an introduction to computational geometry*, Cambridge: MIT Press.

Newton, I., 1728. *Philosophiae Naturalis Principia Mathematica* (A. Koyre and I. B. Cohen, eds.), Cambridge: Cambridge University Press, 1972.

Nilsson, N. J., 1980. The interplay between experimental and theoretical methods in artificial intelligence, Menlo Park: SRI International, TN 229.

Rabin, M. O., and Scott, D., 1959. Finite automata and their decision problems, *IBM Journal of Research and Development* **3**, 114-125.

Rohrlich, F., 1965. *Classical Charged Particles: foundations of their theory*, Reading: Addison-Wesley.

Scott, D., 1976. Data types as lattices, *SIAM Journal of Computing* **5**, 522-587.

Smale, S., 1981. The fundamental theorem of algebra and complexity theory, *Bulletin of the AMS* (New Series) **4**, 1-36.

Thom, R., 1975. *Structural Stability and Morphogenesis: An outline of a general theory of models* (D. H. Fowler, tr.), Reading: Benjamin.

Truesdell, C., 1958. Recent advances in rational mechanics, *Science* **127**, 729-739.

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

Truesdell, C., 1977. *A First Course in Rational Continuum Mechanics*, Vol. 1, New York: Academic Press.

Truesdell, C., and Noll, W., 1965. *The Non-Linear Field Theories of Mechanics, Encyclopedia of Physics* **V. III/3** (S. Flügge, ed.), Berlin: Springer-Verlag.

Truesdell, C., and Toupin, R., 1960. The classical field theories, *Principles of Classical Mechanics and Field Theory, Encyclopedia of Physics* **V. III/1** (S. Flügge, ed.), Berlin: Springer-Verlag.

van der Waerden, B. L., 1931. *Moderne Algebra*, Berlin: Julius Springer.

von Neumann, J., and Morgenstern, O., 1944. *Theory of Games and Economic Behavior*, Princeton: Princeton University Press.