©1986 The Institute of Mind and Behavior, Inc.
The Journal of Mind and Behavior
Winter 1986, Volume 7, Number 1
Pages 1–10
ISSN 0271-0137

Formalism and Psychological Explanation

John Heil

Virginia Commonwealth University

The prospects of a scientific psychology, that is, a discipline (1) in which representational content figures essentially and (2) more or less continuous with biology and physiology, are assessed. It is suggested that the determinants of content may be at odds with psychology's distinctive scientific pretensions. Scientific standing appears to require that contentful psychological states be determined by (supervene on) underlying biological states. Two biologically identical creatures, thus, ought to be psychologically indistinguishable. Content, however, appears not to be so determined. If this is so, we are faced with a choice: either we abandon the possiblity of a scientific *psychology*, or we broaden our conception of what is to count as properly *scientific* psychology.

Consider what may be called the *standard picture* of the science of psychology. Psychology aims at an understanding of what makes us tick. It is to be distinguished from biology and physiology, on the one hand, by the fact that its explanations appeal not to biological or physiological mechanisms, but to psychological states and processes embodying *representational content*. On the other hand, scientific psychology is to be distinguished from everyday explanations of behavior that advert to content by providing a systematic, causal story that promises eventually to mesh with physiology and biology. Psychology affords explanations at a *higher level* than those focusing exclusively on biological hardware, and at a lower, *more basic* level than those of "folk psychology." We assume, of course, that intermediate level states and processes are "realized in" the underlying hardware. Thus, two creatures, biologically identical, must be psychologically identical. (Though the converse need not hold: psychological characteristics need not, on the standard view, be "reducible to" biological characteristics.)

I shall refer to the two components of this picture of psychological science as the *content* component and the *autonomy* component: content figures essentially in psychological explanation, and content is determined exclusively by hardware. My thesis is that a truly scientific psychology (by which I mean

Versions of this paper were presented at meetings of the American Psychological Association (Los Angeles, August 25, 1985) and the Virginia Philosophical Association (Williamsburg, October 17, 1985). Requests for reprints should be addressed to John Heil, Department of Philosophy, Box 2025, Virginia Commonwealth University, Richmond, Virginia 23284.

a psychology that is, in straightforward ways, continuous with biology and physiology) requires both components, but that each is, as it happens, at odds with the other. I shall endeavor to defend this somewhat pessimistic appraisal by appealing to certain characteristics of mental contents. The characteristics in question are commonly taken for granted in the construction of psychological explanations, so that there is, or certainly seems to be, an evident and very deep problem at the heart of the discipline.

One may feel—justifiably— that philosophers have no business telling psychologists (or, for that matter, anyone outside their immediate families) what they can or cannot do. The trouble is that philosophers are not the only philosophers. Psychological theorizing seems often to embody, though unselfconsciously, substantive philosophical commitments. On the rare occasion when these bubble to the surface, they may be (like repressed materials generally) ridiculed and disowned, only to return to the depths, their efficacy undiminished. Philosophers, perhaps, can at least play the role of therapists whose contribution, if any, lies in assembling reminders that serve to focus attention on what ordinarily eludes scrutiny. From time to time, of course, philosophers may succumb to the temptation to offer conceptual replacements or retreads. When this happens, however, psychologists have a right to be suspicious: *caveat emptor* (which, in the present context may be translated: *beware of philosophers bearing gifts*).

The Importance of Content

Why exactly is the notion of representational content (intentionality, semantic content, mental representation—I shall use these expressions interchangeably) important to psychology? In the first place, the capacity to construct and employ representations is surely a significant human capacity, one on a par, for example, with perceiving and cogitating. Considered in this light, representing is on all fours with a host of psychological goings-on.

There is, however, another, rather different point of view on representation. It is the representational aspect of intelligent agency that renders it distinctively *psychological*. Psychological explanation differs from physiological or biological explanation in its appeal to states and processes that exhibit representational (intentional, semantic) *content*. To turn one's back on content is, it seems, to turn one's back on psychology.

I do not mean to suggest that representational goings-on exhaust the subject matter of psychology, only that they occupy a central place in the conceptual network that distinguishes psychology from other attempts to understand agency systematically. The possibility of a genuinely scientific psychology evidently rests on the possibility of a naturalistic account of mental representation, that is, an account of representation that fits smoothly with biology

and physiology. Such an account is commonly taken to require a characterization of the *mechanism* of representation framed in the vocabulary of some more basic idiom (see, e.g., Sayre, in press). This, at any rate, is one pervasive motivation behind computational approaches to psychological explanation. The idea, very roughly, is that if representational processes can be given a computational specification, then the way is open to the eventual (and in-principle) discovery of particular biological mechanisms in which these computations are *realized*. I shall return to this point presently.

The Character of Content

Let us imagine, then, that the possiblity of a viable science of psychology rests on the possibility of our producing a coherent account of the phenomenon of mental representation. Not just *any* account will do, of course: the scientific pretentions of psychology severely constrain prospective candidates. That these constraints in effect rule out any plausible candidate is a central contention of this paper.

Here, in any case, is where the trouble begins. There are, it seems, compelling reasons to suppose that the determinants of mental content undermine (in perhaps surprising ways) the scientific promise of psychology. On the one hand, psychology evidently requires reference to mental contents. It is this that makes psychology *psychology*. On the other hand, a specification of the mechanisms of representational content apparently rules out the possibility that such things could figure in anything approximating a *science* of behavior. We seem faced with a choice: either we abandon our scientific pretentions, or we change the subject.

Consider, first, what I shall describe as the *nonautonomous* character of representation generally. The point is easily grasped in the case of ordinary, nonpsychological representational devices—road signs, maps, diagrams, pictures, and gestures, for example. The content conveyed by such items, their *meaning*, is not determined just by their internal constitution, their structure or architecture, not even by their relations to other, similar items. A single inscription or system of inscriptions can be made to represent now one thing, now another (see Heil, 1981). I may use a system of squares and triangles to represent the disposition of troops at the battle of Borodino, or the layout of a Japanese garden, or an imaginary gathering of unicorns and virgins.

This simple, uncontroversial point about familiar *signs* can, however, be extended to any sort of representational entity at all—including, it would seem, *mental* signs. Suppose, for instance, that such things are in fact nothing more than complicated states and processes in the brains of intelligent creatures. (Nothing whatever hinges on this supposition, incidentally; the very same point would apply to representations conceived of as modifications

of special mental substances or as features of nonneural physical occurrences.)

Imagine now a pair of brains, internally indistinguishable but differently connected to the outside world. One brain is linked in the usual way via an array of sensory-motor systems animating a human body, the other floats in vitro attached to a computing machine programmed to simulate normal inputs and outputs. Although the goings-on in each brain may perfectly mirror goings-on in its counterpart, we should not thereby be entitled to regard the brains as identical with respect to their mental contents. One brain, perhaps, is aware—I shall pretend that this way of talking about brains is unobjectionable—of Guy Lombardo driving past in a DeSoto. The second brain, given that it is differently situated in the world, is aware of something utterly different—the innards of a computing machine, perhaps.

This will be so even if the phenomenal character of the awareness of one brain were indistinguishable from that of the other, even though, from the point of view of each brain, there is no difference whatever. Hilary Putnam has elaborated on this theme in a colorful way (see Putnam, 1981; see also, Heil, 1983; other, perhaps less contrived, arguments on the same point may be found in Baker, 1985a, 1985b). What a given brain—or human being—mentally represents, depends not merely on internal occurrences, but also, and crucially, on connections between those occurrences and external goings-on. If one thinks of mental representations as constituting a language-like system, then it seems best to regard the meaning of the signs in this language as depending at least in part on their relations to happenings *outside* the system.

These observations on representational content are captured in the following priniciple:

> *The Principle of Architectural Inadequacy.* Intentionality (representation, semantic content) cannot be accounted for solely by reference to architectural (i.e., structural, formal, syntactic, nonrelational) properties of intelligent creatures.

If something—an inscription, an utterance, a brain state—possesses representational content, it must do so at least partly in virtue of connections it bears to objects and events external to it. Purely formal or architectural properties of representations determine content only *given* these external connections. Formal differences among symbols on a map, for instance, determine distinct representations only against a background of connections between the *system* of symbols and objects these are taken to represent.

Why should the irreducibly relational character of intentional content threaten to undermine psychological science? Because psychological truths are, as we suppose, *autonomous* truths about intelligent creatures. A creature's psychological response to a given situation is a function of that creature's internal constitution at the time of the response. We may wish to explain the origin or development of that internal constitution by looking at events

in the creature's surroundings or at its history. But a *psychological* explanation of behavior appeals only to *autonomous* states of behaviors (see Stich, 1978). The point may be expressed in the form of a second principle.

> *The Principle of Autonomy.* Psychological states and processes are determined by (supervene on) the internal (presumably physical) states and processes of creatures possessing them.

Two creatures, identical with respect to their internal properties, must, given the Principle of Autonomy, be psychologically identical. Autonomy, however, although evidently required by psychology, excludes mental content, another seemingly essential ingredient of psychological explanation. The irreducibly relational—*nonautonomous*—character of representational states clashes with the requirement of autonomy. A pair of essential components is at odds, the one undermining the other.

If we are to continue along the road mapped for us by traditional psychology, we must, it appears, be prepared to sacrifice one or the other of these elements. We must, that is, give up appeals to mental contents and run the risk of losing psychology to biology and physiology, or abandon hope that psychology can ever be in a position to provide a viable *scientific* account of what makes intelligent creatures tick. Neither prospect is an attractive one.

The Formalist Alternative

As such things go, the conceptual considerations I have been discussing seem relatively uncontroversial. This is not to say that they are unchallengeable. One might, for instance, wish to insist that representational content—intentionality—is, after all, a purely formal property (see, e.g., Palmer, 1978), or that it is a perfectly natural secretion of a certain sort of biological system on a par, perhaps, with chlorophyll (Searle, 1980). Neither of these alternatives, however, has much to recommend it. I shall, for this reason, simply assume the prima facie plausibility of both the Principle of Architectural Inadequacy and the Principle of Autonomy, and proceed to ask what room these leave for a science of psychology.

One answer, recently defended by Stephen Stich, is that psychology must transform itself into a science that explains mental goings-on *syntactically* rather than *semantically* (Stich, 1984). Stich's recommendation is that psychologists might come to regard human beings (and intelligent creatures generally) as syntactic—i.e., formal—systems, rather than, as in the past, semantic systems whose operations depend upon representational content. Formalists do not (or need not) deny that architectural features of intelligent creatures *have* semantic properties, only that these properties play a role in the production of behavior—hence that reference to them has a place in psychological explanations.

The familiar analogy with computing machines may prove useful here. A computing machine operates exclusively on the architectural properties of its inputs. That a given input has a specifiable sense, that it represents something or other, is, for such a device, a matter of indifference. A particular syntactic configuration may represent, at one time, shoes and sealing wax and, at another time, cabbages and kings. So long as the representations remain *formally* identical, their contribution to the machine's operation remains unaffected. Two computing machines that were formally identical though different with respect to their representational properties would, in one important sense, behave identically. The similarities are to be explained by reference to autonomous, syntactic features of the devices. Appeals to representational (semantic, intentional) similarities and differences, though important for certain purposes, would play no role in explanations of this sort.

The formalist recommendation is that psychology be in this way transmogrified into a purely *syntactic* explanatory science. Mental states, as distinct from physiological states, would survive the transition, but only as structures and operations drained of content. Their role in the production of behavior would be pegged exclusively to their formal aspects. Semantic–representational, intentional–properties would drop out of the causal picture altogether, hence reference to mental contents would find no place in psychological explanations. Again, formalists need not contend that particular syntactic items *lack* representational content, only that content plays no causal role in the execution of intelligent behavior.

There are, then, on the formalist view, three distinct perspectives on intelligent creatures. First, such creatures may be characterized as *biological systems* and their behavior explained accordingly. Explanations of this sort belong to the domain of the biologist, physiologist, and anatomist. Second, an intelligent creature may be thought of as a system the behavior of which is determined by familiar intentional states and processes. We all adopt this perspective in our everyday dealings with one another–and, though perhaps to a lesser extent, in or dealings with dogs, cats, and computing machines. Explanations in this vulgar idiom belong to the domain of folk psychology. Formalists despair over attempts to refine the folk psychological intentional framework into a viable scientific enterprise and opt instead for a third perspective on intelligent behavior, one in which explanations are framed exclusively in content-neutral, syntactic terms. What is to be said for this third perspective?

Doubts About Formalism

Much of formalism's plausibility stems from our willingness to embrace an analogy between computing machines and intelligent creatures. The opera-

tion of machines can be described and explained at each of the three aforementioned levels. We can grasp the operation of a given machine at the hardware level (as an engineer might), at the semantic level (as when we explain what a chess-playing device does by saying that it wants to protect its queen), and at the programming level. Which perspective we take on a particular machine may depend, mostly, on our own pragmatic interests (see Dennett, 1978). Ordinarily we regard the first of these—the hardware perspective—as too fine-grained to be of much use. More to the point, so long as we confine ourselves to a consideration of hardware, we are apt to miss important generalizations across machines. Different machines may "realize" the same program in utterly different ways. By the same token, we are likely to overlook another important class of generalizations when we remain at the semantic level. Different machines may be identically programmed (or embody identical sub-programs), yet, owing to matters extraneous to their operation, differ in many of their most interesting semantic properties. A machine that takes inventory on shoes and another that keeps track of sealing wax may run identical programs.

We may grant, then, that there is an important syntactic level of explanation for the operation of computing machines. Might not the same be true of intelligent creatures? Might not there be a syntactic descriptive and explanatory level *between*, as it were, the biological and intentional levels? We are sometimes encouraged to believe that there *must* be such a level on the grounds that representation—allegedly—*requires* a formal syntax of some sort. If this were so, and if we are willing to ascribe representational states to human beings, then we should have to suppose as well that it is possible, at least in principle, to characterize those states and operations purely syntactically, purely in terms of their formal architecture.

An argument of this sort inherits much of its force from the analogy with computing machines. That analogy is innocuous so long as it is employed to illuminate a distinction between *levels* of description and explanation. It reminds us that we are apt to miss important generalizations if we focus exclusively, in the case of computing machines, on their hardware and, in the case of intelligent creatures, on their biology. But of course we knew this without having to appeal to features of computing machines.

Difficulties arise when the computational analogy is extended in a particular direction. One is invited to accept the notion that intelligent creatures resemble computing machines not simply in respect to their susceptibility to distinct levels of description and explanation, but also in respect to their *operation*. Computing machines are—essentially—devices designed to perform operations on uninterpreted formal strings. We understand such operations and, in addition, have a grip on how they can be realized in a variety of physical systems. It is tempting, now, to slide into the view that there *must* be a pure-

ly syntactic way of describing the operations of intelligent creatures, that is, a level of description that appeals exclusively to the *formal* features of inputs and internal processes. This is manifestly so in the case of computing machines, why not, then, in the case of human beings and their near relations?

The question, I suggest, must be turned around. Is there any reason, other than a prior infatuation with computational models, to suppose that the capacities of intelligent creatures can be given a purely formal specification? Even if such a specification were available, of course, additional argument would be required to support the much stronger claim that intelligent behavior in humans and brutes is actually determined by purely formal manipulations. From the fact that one might concoct a formal account of the operation of a given device—a mercury thermometer, for instance—it does not follow that such an account sheds light on the actual operation of the device. It is useful here to bear in mind the Kantian distinction between acting *in accord with* a given rule and acting *on* that rule (see, e.g., Heil, in press).

Sometimes we are told that there is no other reasonable way of accounting for intelligent performances: one or another version of formalism is the only game in town. It is scarcely surprising that alternative accounts have been disappointingly sparce, however. We seem to have thoroughly inculcated the computational model and, with it, the methodological stricture that, above the hardware level, only formal explanations are explanations. But if formalism *is* the only game in town, then this is due in part to formalists' insistence that nothing else could count as a game.

The attitude is a convenient one. It relieves those who harbor it of having actually to confront and assimilate wayward data. Promissary notes may be issued ad lib. When we step back from the notion that descriptions and explanations of intelligent goings-on must be couched in a formal idiom in order to be scientifically admissible, however, we can begin to appreciate the extent to which this stricture, like the metaphysical view from which it stems, is importantly at odds with what we know about ordinary mental operations. We seem not, for example, to employ well-defined formal categories, but structures loosely organized around salient prototypical exemplars (Rosch, 1978). Tasks like pattern recognition that seem trivial for intelligent creatures have thus far resisted formal specification (Dreyfus, 1979). In contrast, human beings boggle at computational activities that even the lowliest computing machines perform in a twinkling (Hirst, 1977). The burden of proof, it would appear, lies with those who wish to espouse formalism. What grounds are there for thinking that psychological occurrences are amenable to formal specification? A certain ingrained metaphysic tells us they must be; experience tells us otherwise.

A second, though perhaps not unrelated, worry about formalism concerns a point that is so obvious it tends to escape attention. Human beings—and

perhaps other creatures as well—evidently react to and operate on *semantic stimuli*. You respond to my utterances, to traffic lights, and to newspaper editorials in virtue of what these things *mean*. We are situated in the world awash in semantic information, and our lives are conducted largely in terms of the *significant* aspects of our surroundings. Formalism seems committed to the view that all of this can somehow be recast as syntax. But surely there is no reason to feel optimistic about the prospects for reductionist schemes of this sort. They merely present, in a new and putatively scientific guise, a venerable positivist dream, a dream evidently embodying perennial philosophical sex appeal. It is one thing for philosophers to allow themselves to be seduced by metaphysical sirens, however, and quite another matter for the affair to be turned into a methodological point of honor to be foisted on psychologists toiling in the vineyards.

Concluding Remarks

One may (and, I have been suggesting, ought to) reject formalism without rejecting the formalist critique of conventional psychological theorizing. There remains an inevitable tension at the core of the subject. The need for scientific respectability pushes in the direction of autonomous explanation, while a desire for disciplinary integrity inclines oppositely. Perhaps what we now regard, for no very good reason, as a single discipline, will diverge, one component evolving away from mental content and toward physiology and biology, another component, that concerned with content, declaring its independence from narrow conceptions of scientific acceptability foisted by those with philosophical axes to grind. The second of these, at any rate, is an interesting possibility, one that, depending on one's ideological preferences, might be described either as a final abandonment on the part of psychology of its scientific pretentions, or as a long overdue broadening of our conception of science.

References

Baker, L. (1985a, April). *Cognitive suicide?* Paper presented at the Oberlin Philosophy Colloquium, Oberlin, Ohio.

Baker, L. (1985b). Just what do we have in mind? In P. French, T. Uehling, and W. Wettstein (Eds.), *Midwest studies in philosophy: Volume 10* (pp. 25–48). Minneapolis: University of Minnesota Press.

Dennett, D. (1978). *Brainstorms*. Montgomery, Vermont: Bradford Books.

Dreyfus, H. (1979). *What computers can't do* (2nd ed.). New York: Harper and Row.

Heil, J. (1981). Does cognitive psychology rest on a mistake? *Mind, 92*, 321–342.

Heil, J. (1983). *Perception and cognition*. Berkeley: University of California Press.

Heil, J. (in press). Does psychology presuppose rationality? *Journal for the Theory of Social Behaviour*.

Hirst, W. (1977, December). *Machine simulation of human cognitive functions*. Paper presented at the American Association for Symbolic Logic, Washington, D.C.

Palmer, S. (1978). Fundamental aspects of cognitive representation. In E. Rosch and B. Lloyd (Eds.), *Cognition and categorization* (pp. 259–303). Hillsdale, New Jersey: L.J. Erlbaum Associates.
Putnam, H. (1981). *Reason, truth, and history.* Cambridge, England: Cambridge University Press.
Rosch, E. (1978). Principles of categorization. In E. Rosch and B. Lloyd (Eds.), *Cognition and categorization* (pp. 27–48). Hillsdale, New Jersey: L.J. Erlbaum Associates.
Sayre, K. (in press). Intentionality and information processing. *Behavioral and Brain Sciences.*
Searle, J. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3,* 417–424.
Stich, S. (1978). Autonomous psychology and the belief-desire thesis. *Monist, 61,* 573–591.
Stich, S. (1984). *From folk psychology to cognitive science.* Cambridge, Massachusetts: M.I.T. Press.