Philosophy of Science

June, 1995

FOLK PSYCHOLOGY AND COGNITIVE ARCHITECTURE*

FRANCES EGAN†‡

Department of Philosophy
Rutgers University

It has recently been argued that the success of the connectionist program in cognitive science would threaten folk psychology. I articulate and defend a "minimalist" construal of folk psychology that comports well with empirical evidence on the folk understanding of belief and is compatible with even the most radical developments in cognitive science.

1. Introduction. The last decade has seen the development in cognitive science of a research program that challenges the "classical" model of the mind prevalent since the 1960s. There is continuing discussion within cognitive science about whether connectionist models might provide adequate explanations of human cognitive phenomena that significantly differ from and compete with explanations provided by classical models.

The cognitive architecture dispute appears to have implications outside of cognitive science itself. It has recently been argued (Ramsey, Stich, and Garon 1991, Davies 1991, Rey 1991) that the connectionist program in cognitive science threatens our commonsense conception of ourselves as thinkers and invites the elimination of the folk psychological notions of belief and desire. In this paper, I articulate and defend a construal of commonsense psychology that is immune to the latest eliminative challenge. I argue that folk psychology imposes no substantive constraints on accounts of our underlying cognitive architecture; consequently, it is not

*Received February 1994; revised July 1994.

†Thanks to William Bechtel, David Davies, Paul Lodge, Robert Matthews, Robert McCauley, and William Ramsey for helpful comments and criticism.

‡Send reprint requests to the author, Department of Philosophy, Rutgers University, New Brunswick, NJ 08903, USA.

Philosophy of Science, 62 (1995) pp. 179–196

threatened by the success of connectionism. Unlike recent defenses of folk psychology (see, for example, Dennett 1987, 1991, Van Gelder (forthcoming), Wilkes 1986, 1991), my argument does *not* turn on denying a central premise of the eliminativist argument—that folk psychology is an explanatory theory that purports to describe the inner causes of behavior. I take this to be an advantage of my account, since the causal–explanatory construal of folk psychology is itself well-supported.

2. Two Families of Cognitive Models. Classical computational architectures treat cognitive processes as rule-governed manipulations of internal symbols or data structures. In connectionist models, by contrast, there are no fixed representations over which the device's operations are defined. Rather, there are activated units (nodes) which increase or decrease the level of activation of other units to which they are connected until the ensemble settles into a stable configuration. Semantic interpretations, if assigned at all, are assigned either to individual units (in localist networks) or to patterns of activation over an ensemble of units (in distributed networks).

Proponents of the classically-inspired *language of thought thesis* (hereafter, LOT; see Fodor 1987) argue that propositional attitudes can be construed as computational relations to symbols in an internal code. On this view, to believe that Clinton is a Democrat is to bear the computational relation characteristic of belief to a internal sentence that means *Clinton is a Democrat.* The LOT, if true, would provide a scientific vindication of folk psychology, as it posits states which are not only type-correlated with propositional attitudes, but also individuated, like propositional attitudes, by reference to a relation-type and a content sentence. Indeed, the LOT promises a *reduction* of commonsense psychology, as LOT-based explanations of mental processes will be isomorphic to the explanations offered by common sense.

Current connectionist models typically do not bear a transparent relationship to folk psychology. This is due in part to the fact that connectionism eschews the process/data structure distinction central to the classical program and does not construe mental processes as operations on (relations to) data structures. Thus, connectionist states are not individuated along the two dimensions by which we taxonomize propositional attitudes. Activated nodes in a network (or patterns of activation across an ensemble of nodes) have no natural interpretation as relations to contentful structures. (I am not suggesting that connectionist states could not be so interpreted, simply that they do not wear such interpretations on their sleeves.) Consequently, a connectionist-based vindication of folk psychology is more difficult to envision.

However, scepticism about connectionism's prospects of providing a much hoped for vindication of commonsense psychology does not justify the claim that connectionism implies the *falsity* of folk psychology, just as quantum physics' failure to vindicate our commonsense ontology of middle-sized objects does not imply that there are no tables or chairs. Let us turn, then, to the most widely-debated argument from connectionism to eliminativism.

3. Ramsey, Stich, and Garon's Argument. Arguments for eliminativism presuppose that folk psychology is a *theory*, and hence a candidate for elimination or replacement. Beliefs, desires, and the other propositional attitudes, on this view, are plausibly regarded as posits of a commonsense theory. Central to Ramsey, Stich, and Garon's (hereafter, RSG) argument is the claim that folk psychology is committed to a cluster of claims called *propositional modularity*, according to which propositional attitudes are

> . . . *functionally discrete, semantically interpretable*, states that play a *causal role* in the production of other propositional attitudes, and ultimately in the production of behavior. (1991, 204)

The claim that beliefs and desires, as characterized by folk psychology, are causally efficacious is supported by the fact that the folk psychological explanation of a behavior typically cites a particular desire (in conjunction with appropriate beliefs) as the cause of the behavior. Thus, for example, the explanation of Alice's going to her office might plausibly cite her desire to send e-mail messages as the cause of her going to the office, even though Alice also wants to speak to her research assistant, and had she not wanted to send e-mail messages, the desire to speak to her research assistant might have caused the behavior. The thesis that beliefs and desires are functionally discrete states amounts to the claim that it makes sense to talk of acquiring or losing them one at a time. Thus it makes sense, for example, to say that after awakening from a nap Henry forgot that he had unplugged the phone. He may have forgotten nothing else.

RSG's claim that propositional attitudes are functionally discrete states needs qualification in light of the oft-noted holism (or, at least, anatomism) of belief. If Henry has forgotten that he has unplugged the phone, he has also forgotten (hence, currently does not believe) that he has unplugged something, that he cannot currently be contacted by telephone, etc. The individuation of belief in commonsense practice is fine-grained enough that these count as distinct beliefs. It is therefore doubtful that individual beliefs *can* be acquired or lost one at a time. However, this qualification aside, I shall grant RSG's claim that folk psychology is committed to something like propositional modularity. The question, then, is whether connectionism poses a threat to the propositional modularity of beliefs and desires.

The LOT is clearly compatible with propositional modularity, since it posits states which are themselves functionally discrete, semantically evaluable, and causally efficacious in the production of behavior, with which propositional attitudes are type-correlated. Indeed, according to Fodor and Pylyshyn (1988, 57), "conventional [computational] architecture requires that there be distinct symbolic expressions for each state of affairs that it can represent." Thus, the LOT, if true, would explain the propositional modularity of beliefs and desires.

Some connectionist networks are similarly compatible with propositional modularity. In "localist" connectionist models, individual units or small clusters of units are assigned a semantic interpretation. Thus, if a unit representing *fur* is always activated when a unit representing *dog* is activated, then the ensemble consisting of the two units and the connection between them might be construed as the system's representation of the proposition *all dogs have fur*.

However, not all connectionist networks exhibit this sort of functional localization. In all but the simplest networks, the connections between the input units of the network (whose activation values represent an encoding of the input to the system) and the output units (whose activation values encode the output the system has computed from the input) are mediated by hidden units, which represent neither the input nor the output. RSG describe a class of connectionist cognitive models with the following additional properties: (1) individual hidden units (and weights and biases) in the network have no plausible symbolic interpretation; and (2) the encoding of information in these networks is not *local* but rather *distributed* over many nodes and connection strengths. These networks may plausibly be regarded as holistically encoding a set of propositions, although the representational strategy employed is what Smolensky (1988) has termed 'subsymbolic,' since none of the hidden units, weights, or biases can comfortably be construed as symbols.

In the connectionist models under consideration, no distinct state or part of the network serves to represent any particular proposition. Large chunks of the network (that is, many units, many connection strengths, and many biases) play a role in each computation, with individual units, weights, and biases encoding information relevant to many propositions. The representation of any given proposition is widely scattered throughout the network. It appears that such models do not posit distinct states with which the functionally discrete, semantically evaluable, causally efficacious states characterized by folk psychology might plausibly be identified, from which RSG conclude:

> If these models turn out to offer the best accounts of human belief and memory, we will be confronting an *ontologically radical* theory change—the sort of theory change that will sustain the conclusion that propositional attitudes, like caloric fluid and phlogiston, do not exist. (1991, 218)

RSG's argument can be explicitly formulated as follows:

(1) the network lacks functionally discrete, identifiable substructures that are semantically interpretable as representations of individual propositions;
(2) therefore, the representation of a particular proposition cannot plausibly be said to play a causal role in the network's computation;
(3) however, folk psychology is committed to the propositional modularity thesis, which implies that particular beliefs do play causal roles in specific cognitive episodes; and
(4) therefore, if the best models of human cognitive processes are distributed connectionist networks of the sort described, then folk psychology is false.

Compelling though this argument may appear, it fails to establish its conclusion. (1) is true for a wide range of connectionist models—so-called "distributed" networks. However, (2) follows from (1) only if the representation of a particular proposition must be realized as a discrete, identifiable substructure to be causally efficacious. RSG do not offer an argument for this claim. To establish its truth, and hence to establish claim (2) of their argument, RSG need to argue that distributed representations are *epiphenomenal*—that they play no causal roles in the network's behavior—something that they do not attempt. The growing literature on distributed representations does not construe them as epiphenomenal. (See, for example, Hinton, McClelland, and Rumelhart 1986, and Van Gelder 1991.) If distributed connectionist models are taken at their face, and in the absence of an argument to the contrary surely they should be, then claim (2) of RSG's argument appears to be false.

The argument from connectionism to eliminativism collapses with the apparent falsity of step (2). But suppose that (2) were true. It might turn out that distributed representations are epiphenomenal—that the complex states that are assigned semantic interpretations in the best connectionist models play no causal roles in the networks' computations. The causal generalizations that describe the networks' behavior, let us suppose, do not advert to these particular complex states. The semantic interpretation of these states in the envisioned models would play a purely heuristic role, allowing us to keep track of what the network is doing. Would the eliminativist conclusion follow? It follows only if propositional attitudes, to be causally efficacious, must be realized as structures which figure explicitly in the causal generalizations of a lower level cognitive theory. However, to assume that they must is to presuppose an unsubstantiated, and very strong, constraint on inter-theoretic compatibility. It is not generally true that the causal generalizations of a lower level theory will advert to the complex of structures that realize a causally efficacious state posited at a higher level of theory. Very often, the complex will be arbitrary from the perspective of the lower level theory. (This will be true even if the higher level states are not multiply-realized by lower level structures.) For example, there is at present no biochemical characterization of the gene responsible for sickle cell anemia, but it is very unlikely that the complex biochemical structure that realizes the gene is theoretically significant from the perspective of biochemistry. Nonetheless, the likelihood that the causal generalizations of biochemistry do not advert to this particular structure does not impugn the molecular geneticist's claim that the gene causes the sickle cell condition. (This example was suggested to me by Robert McCauley.) Analogously, the possibility that those complex structures that precisely realize beliefs and desires do not figure in the causal laws of the cognitive-level science does not threaten the causal efficacy of the propositional attitudes.

Nor is there any reason to assume that beliefs and desires must be realized as functionally discrete cognitive structures to satisfy the functional discreteness component of propositional modularity. It is not generally true that functionally discrete items posited at one level of theory must be realized by structures which are treated as functionally discrete at lower levels. The sickle cell gene, as characterized by molecular genetics, is functionally discrete—it plays a distinct role in the development of the phenotype—yet the complex of chemical structures realizing it may have no discrete biochemical role.[1]

The point here is that beliefs and desires need not be realized in structures that are causally efficacious, functionally discrete, and semantically evaluable as characterized by a lower level theory to satisfy the demands of propositional modularity. This cluster of commitments describes how folk psychology itself characterizes such states. Thus, even if the representations of particular propositions were to turn out to be epiphenomenal, from the standpoint of connectionist theory (because they play no

[1] Similarly, beliefs and desires are presumably not realized by functionally discrete neurological structures; nor is it assumed that the structures posited by cognitive level theories (both connectionist and classical) are realized by functionally discrete neurological hardware. Given that connectionist networks typically do not bear a transparent relation to the neurological structures that realize them, the description of connectionist networks as "neural nets" is somewhat misleading.

characterizable causal roles in connectionist models), the eliminativist conclusion that RSG want would not follow.

However, while RSG have failed to discharge the eliminativist burden, the following worry may remain: if one's *entire* psychological state underlies a particular belief, as seems to be the case in the connectionist models under consideration, does it not follow that propositional attitudes are *emergent* out of underlying psychological processes, that there is no explanation of how beliefs are realized psychologically? Not necessarily. It remains true that the device has a particular belief because of the information represented in the system. Standing beliefs may well be explainable in terms of the learning history and connection strengths of the system, and occurent beliefs (those causally efficacious in a particular behavioral episode) may be explainable in terms of the current pattern of activation of hidden nodes of the system. It is no mere accident that the network behaves as it does.[2]

Consider, once again, the functional discreteness component of propositional modularity. Recall that RSG (1991, 205) gloss it as the claim that "it typically makes perfectly good sense to claim that a person has acquired (or lost) a single memory or belief" as, for example, it makes sense to say that after awakening from a nap Henry forgot that he had unplugged the phone. By adjusting the input to a network, the activation levels of individual units, and the connection weights and biases of the ensemble, one can change the representations in a distributed connectionist network. It may not be transparent how this is to be accomplished for a particular proposition, as it is in classical models or localist connectionist networks, where one simply changes the local state (in classical models, by adding or deleting the appropriate data structure); however, inter-theoretic realization relations in science are rarely so tidy. (Forster and Saidel (forthcoming) describe a very simple distributed network where a single representation, realized in a distributed fashion, can be added to or deleted from the network's representational repertoire, indicating that functional discreteness is not incompatible with non-discrete realization.)

[2] There is no reason to insist, as RSG seem to (see their responses to replies, pp. 216–217), that standing and occurrent beliefs be given a uniform account in connectionist models. Classical models of belief are unlikely to give a uniform account either, since there are not enough data structures available to account for all of an individual's standing beliefs. (Recall Fodor and Pylyshyn's (1988, 57) claim that "conventional architecture requires that there be distinct symbolic expressions for each state of affairs that it can represent.") Classicists have attempted to solve this problem by appealing to a distinction between *core* and *derivative* cases, with only core cases—those corresponding to "episodes in mental processes" (Fodor 1987, p. 25)—explicitly represented as data structures. Derivative cases, corresponding to the system's standing beliefs, are given a dispositional analysis. Since classicists have so far failed to explain how core and derivative cases are related, in particular, how a standing belief could come to play a role in a mental process, there is at present no classical account of standing belief.

I have argued that propositional attitudes need not be realized by discrete computational-level structures to be causally efficacious and functionally discrete. RSG's conclusion could be salvaged if the claim that propositional attitudes are realized by discrete computational structures is a fundamental commitment of folk psychology itself, entirely independent of its commitment to propositional modularity. If folk psychology did make such a claim, then it would be rightly viewed as in part a theory of psychological *processes*, and consequently a competitor to some forms of connectionism. In the next section, I shall argue that folk psychology makes no commitment regarding how propositional attitudes are realized and imposes no substantive constraints on cognitive architecture. If I am correct, then folk psychology is not incompatible with any form of connectionism.

4. A Minimalist Construal of Folk Psychology. My argument against RSG does not claim that distributed connectionist networks really do posit structured states that might plausibly be identified with propositional attitude tokenings. Nor does it appeal to what might be called a neo-Rylean construal of folk psychology according to which beliefs and desires are construed not as causally efficacious internal states of agents but as *abstracta* or logical constructs out of behavioral patterns (see Dennett 1991, Van Gelder (forthcoming)), and thus immune, in principle, to an eliminativist challenge.

I shall call the construal of folk psychology defended here, according to which our commonsense theory involves no substantive commitments about how the internal causes of behavior are realized, either computationally or physically, the *minimalist* construal, or simply, *minimalism*. The minimalist construal of folk psychology might be seen as a special case of Mark Johnston's more general Minimalist thesis, according to which "metaphysical pictures of the justificatory undergirdings of our practices do not represent the crucial conditions of justification of those practices." (Johnston 1992, 590) Jackson and Pettit (1990) and Horgan and Graham (1991) have also defended versions of minimalism.

Minimalism construes folk psychology as committed to something like the propositional modularity thesis: propositional attitudes are semantically-evaluable internal states of agents and are characterized by their roles in the production of behavior and other mental states. These roles are specified by generalizations of which the following are typical examples:

(S) (p) (A) [If S wants p, and S believes that doing A is the only way to bring about p, then (*ceteris paribus*) S will do A]

(S) (p) (q) [If S believes p, and S comes to believe that if p then q, then (*ceteris paribus*) S will come to believe q]

(S) (p) [If S fears p, then (*ceteris paribus*) S does not want p].

Generalizations of this sort, which taken together constitute a theory tacitly known by the folk and deployed by them in the explanation of mental phenomena and behavior, characterize propositional attitudes by reference to their typical causes and effects, intentionally described. They say nothing about the physical or computational realization of these states.

As a construal of folk psychology, minimalism occupies an intermediate position between neo-Rylean and behaviorist positions which deny that propositional attitudes are to be construed as causally efficacious internal states and architecturally committed, or *extravagant*, interpretations of folk psychology. A defense of the minimalist position, therefore, requires argument on both flanks.

Neo-Ryleans typically construe folk psychological explanations of behavior as *rationalizing*, rather than *causal*. Their purpose, it is claimed, is to "situate a piece of behavior in the space of reasons," to show that it is rational (and hence, predictable) in the circumstances, rather than to cite actual causes. A non-causal construal of folk psychology insulates it from seemingly hostile developments in empirical science. If the argument in the previous section is correct, then insulation is not necessary. Moreover, a non-causal construal is not supported by actual practice.

Neo-Ryleanism is revisionary about our shared explanatory practices. It certainly *seems* as if we take beliefs to be effects of perception and inference, and causes (in conjunction with desires) of action. We often say things such as "He believed he was about to be fired *because* he saw a confidential memo that criticized his job performance" and "She quit smoking *because* she believed it was affecting her health." There is no reason to suppose that "because" here functions any differently than in locutions which are clearly causal, such as "The fire started *because* the electrical system was overloaded." In claiming that commonsense explanations of belief fixation and action are not causal, despite appearances, the neo-Rylean assumes a rather heavy burden of proof.

The case for a causal construal of belief–desire attributions does not rest solely on linguistic intuitions. *Attribution theory* is the branch of social psychology that studies the perceived causes of behavior. (See Heider (1958) for the classic statement of attribution theory; see Kelley and Michela (1980) and Weiner (1990) for more recent surveys of the attribution literature.) Attribution theory's applications take us well beyond the domain of folk psychology. For example, attribution theorists have argued that an agent's self-esteem and susceptibility to depression are a function of whether the prime determining factor ("locus of causality") of the success or failure of a behavioral event is thought to be an external condition beyond the agent's control (for example, *luck*) or an internal condition of the agent, such as an enduring character trait (see Peterson and Seligman 1984). Of present relevance is the fact that a large and influential body of research in empirical psychology is predicated on the claim that beliefs, desires, and more permanent conditions such as character traits are implicated in causal explanations of behavior. For example, an agent's failure to expend the amount of effort required to secure a goal may be attributed to the fear that he will fail, or a world-class athlete's Herculean efforts in the face of adversity may be attributed, in part, to her belief that she is the best at her sport. According to attribution theory, the propositional attitudes ascribed in such explanations are construed (by the folk) as *causes*.

While it seems clear that folk psychology does construe beliefs and desires as internal causes, the available evidence supports nothing stronger than minimalism. Minimalism contrasts sharply with extravagant construals of folk psychology, some of which claim that propositional attitudes, as explicated by folk psychology, have a language-like structure (e.g., LOT), must be realized by discrete computational states, etc. Extravagant construals typically underlie eliminativist arguments,[3] although advocates of folk psychology, especially those who anticipate that computational psychology will provide a vindication of the folk categories, have typically assumed extravagant construals as well. (Lycan (1991) and Rey (1991) assume that folk psychology construes propositional attitudes as relations to mental representations; i.e., both assume a version of the LOT.) Folk psychology is better off without such friends. Tying folk psychology to the fate of particular processing-level proposals that purport to explain why the folk theory is true (or how it works) is an invitation to eliminativism. More to the present point, extravagant construals are simply unsupported by folk psychological practice.

It has been noted that some form of intentional explanation is universal among adult humans (Forguson and Gopnik 1988, Fodor 1987). Significantly, there is a striking degree of interpersonal agreement on belief and desire ascription, despite widely divergent opinion (or more accurately, ignorance) about what beliefs and desires really are (that is, what they are made of and how they do their causal work). For the purposes of predicting and explaining the behavior of others—folk psychology's spe-

[3] Churchland (1981) assumes that folk psychology is committed to what he calls "sentential kinematics". He does not explicitly define this thesis, but it appears to involve the idea that mental processes are manipulations of sentence-like structures. Stich (1983) assumes that folk psychology is committed to a *modularity* condition, which requires local realization. These assumptions are not defended in any detail.

cial forte—it does not matter to folk psychology's practitioners whether propositional attitudes are realized by discrete computational states, large-scale neural structures, or some sort of mysterious soulstuff, where the global *vs.* local/discrete distinction has no clear application.

The point is incontrovertible if we bring young children within our purview. There is compelling evidence from developmental psychology that children have acquired the concepts *belief, desire*, and *intention* by the age of six or seven (Astington, Harris, and Olsen 1988, Wellman 1990). The basic elements of our folk psychological understanding of ourselves are in place by this time. Of course, young children often have trouble predicting (and explaining) the behavior of others, and there is considerable disagreement among developmental psychologists regarding the correct explanation for their deficiencies. (See Wimmer and Perner (1983) and Forguson and Gopnik (1988) for competing accounts.) There is, however, no evidence to support the view that very young children's problems with belief ascription are attributable to ignorance of our cognitive architecture; nor is there any evidence to support the idea that six year olds have a more sophisticated understanding of our cognitive architecture than younger children. Furthermore, developmental psychologists agree that by six years of age children are proficient folk psychologists, although there is no evidence that by this age they have acquired a general belief that causally efficacious states must be realized by structures that are themselves functionally discrete. (It is sometimes suggested that children have a crude "billiard ball" model of causation, but a mechanistic model of causation provides no support for a claim about how beliefs and desires are internally realized.) What the evidence does support is a minimalist construal of folk psychology: young children are fully capable of attributing beliefs and desires to their fellows; what they are attributing are simply semantically evaluable internal states that are causally efficacious in the production of behavior (and other propositional attitudes). There is no evidence suggesting that they have any beliefs about how these causally efficacious states are realized.

What about adults? Attribution theory is once again of relevance, and it indicates that the adult's conception of propositional attitudes is continuous with the child's. While the folk posit enduring, stable states that play causal roles in producing behavior, there is no evidence that they share any particular views on underlying psychological or neural processes or mechanisms.

It appears, then, that the available empirical evidence supports the minimalist construal of folk psychology over the alternatives. So construed, folk psychology is not susceptible to eliminativist arguments of the sort that have been offered, although it is *conceivable* that folk psychology could be false. Since propositional attitudes are internal causes, being behaviorally indistinguishable from a believer is not sufficient for being a believer. The behavior must be caused by internal states that play the causal roles characterized by folk psychological generalizations. The fact that a system's behavior is predictable using folk psychological generalizations is compelling evidence that the system has such states, and is a true believer. But this evidence is defeasible. I have argued that it would *not* be defeated by the failure of scientific psychology or neuroscience to find independently characterizable states or structures with which propositional attitudes tokenings can be identified.

However, it is not hard to imagine a case where the behavioral evidence would be defeated. Imagine a robot, behaviorally indistinguishable from a typical human (right down to behavioral dispositions), whose "actions" are produced by Martian scientists manipulating its motor and speech organs by remote control (cf. Peacocke's (1983) story). Although folk psychological generalizations would be useful for predicting the robot's behavior, the robot is *not* a model of folk psychology, because its behavior is not caused by internal states of the sort characterized, in terms of their causal roles, by folk psychological generalizations. The robot's behavior is not caused by its own beliefs and desires. It may have no internal states with the appropriate causal roles (i.e., no beliefs and desires). It seems to be a mere conduit for the intentions of others.

It is possible, therefore, to imagine circumstances where folk psychology would be false. If most of the "human" population were remote-controlled robots, then all "output-side" folk psychological generalizations—those purporting to explain behavior—would be false. Behavior would not be caused by beliefs and desires of the behaving subject.

Let us call the extrabehavioral condition on being a folk psychological subject the *autonomy condition*. I shall not attempt to give the autonomy condition a positive characterization—suffice it to say that it requires simply that the subject's behavior is not the result of manipulation of its motor and speech organs by an external agent. There is indirect evidence to support the claim that the autonomy condition is an essential commitment of the folk psychological understanding of ourselves. Courts of law are often asked to rule on issues of intentional agency. In doing so they consider not only the subject's behavior, but also whether the behavior was caused by intentional states of the subject. Imagine a case where a person kills someone while under the control of a hypnotist. Assuming that it could be determined that the killing really was the result of hypnotic suggestion, it would not be considered an intentional action of the hypnotized subject, and (provided that he had not paid the hypnotist to make the fatal suggestion) the person would not be convicted of murder. (The hypnotist is more likely to be regarded as the murderer, inasmuch as the intentional action resulting in the death is attributable to him.) The hypnotized subject's behavior was not caused by the subject's own beliefs and desires. It does not fall under a folk psychological generalization, precisely because the subject's bodily movements were under the direct control of an external agent. We can generalize to get the autonomy condition—if all of a subject's behavior were under the direct control of an external agent (e.g., a hypnotist), then since none of folk psychology's output-side generalizations would be true of the subject, it would not be a model of folk psychology. Indeed, if the etiology of its behavior were understood, it is unlikely that such a subject would be considered a person.

To summarize the argument in this section: empirical evidence on the folk understanding of belief supports what I have called minimalism, the thesis which holds that propositional attitudes are construed as causally efficacious internal states, but denies that there are widespread views about how such states are realized. Since folk psychology imposes a minimal extrabehavioral condition on being a folk psychological subject (viz. *autonomy*), it is not compatible with every conceivable cognitive architecture. If most "humans" had the architecture of remote-controlled robots, then folk psychology would be false. But nothing short of fantastical scenarios of this sort would clearly falsify it. It is not threatened by discoveries about our cognitive architecture of the sort that connectionists are hoping for.

5. Lingering Doubts: Another Eliminativist Argument. The worry may remain that distributed connectionism and folk psychology, while not incompatible, nonetheless do not make congenial bedfellows. In the hope of assuaging lingering doubts about their compatibility, I shall conclude by considering a somewhat different argument from connectionism to eliminativism. Martin Davies (1991), drawing inspiration from the work of Gareth Evans, claims to have found a tension between connectionism and our commonsense conception of ourselves as thinkers. According to Davies,

> A thinker who has the thought that a is F appreciates that it follows that a is H, say; and he also appreciates that from the thought that b is F it follows that b is H. But that is not all. It is not just that there is an input-output pattern in the inferences that the thinker is disposed to make. The two inferences are manifestations of a common underlying capacity; namely, mastery of the concept of being F. (1991, 243)

So, for example, someone who knows that Oscar is a bachelor and also knows that Elmer is a bachelor appreciates that both Oscar and Elmer are unmarried, in virtue of a mastery of the concept of being a bachelor. Davies calls this the *neo-Fregean conception of thought*, and he claims that it is part of our commonsense conception of what it is to be a thinker. He goes on to say,

> . . . the idea of a common capacity being manifested in the two inferences should be unpacked in terms of a common explanation, adverting to a common state. In short, there is a causal systematicity relative to the input-output pattern in a thinker's inferential practice. (1991, 243–44)

Davies gives an informal characterization of causal systematicity: a process is *causally systematic* relative to a discernible pattern in its input-output behavior. Suppose that a generalization G describes such an input-output pattern. Then,

> . . . the requirement for causal systematicity relative to a pattern described by G is that there should be a mechanism whose presence in the system explains all the input-output transitions that conform to the pattern described by G. It is not sufficient that this common mechanism should merely figure as a component somewhere along the way in the several transitions. Rather, the common mechanism should actually mediate between inputs and outputs in accordance with G. . . . Causal systematicity requires real commonality of process. (ibid., pp. 235–36)

Davies' unpacking of the neo-Fregean conception of thought in terms of causal systematicity imposes a constraint on internal architecture that can be formulated as follows:

> A has a concept p only if (for whatever computational mechanisms A has) there is a computational state (structure) of type m such that (i) m realizes in A the concept p, and (ii) m is uniformly deployed in all cognitive processes involving the concept p.

I shall call this the *uniform realization constraint*. In Davies' view, our commonsense conception of ourselves as thinkers commits us to the uniform realization constraint on cognitive architecture.

He further points out that distributed connectionist models typically do not have syntactically structured representations of the sort seemingly required by causal systematicity. In particular, they fail to satisfy the uniform realization constraint. In distributed connectionist networks, as we have seen, a proposition is represented as a pattern of activation over many units. As Smolensky (1988) admits, the constituent subpatterns of activation that represent coffee in various contexts—coffee in a cup, coffee in a jar, coffee with sugar—"are activity vectors that are not identical, but possess a rich structure of commonalities and differences (a family resemblance, one might say)" (p. 17). Strictly speaking, in such networks, there is no common subpattern of activity that can be identified as a realization of the concept *coffee*. Hence, there is no component or state of the network uniformly deployed in all coffee transitions. So these networks fail to satisfy the uniform realization constraint on concepts.

The similarity between Davies' argument and Fodor and Pylyshyn's (1988) systematicity argument should be apparent. Fodor and Pylyshyn conclude, from the fact that no common constituent is uniformly deployed in all transitions involving the exercise of a concept, that connectionist networks are unable to explain a pervasive feature of thought readily explainable by classical models and are clearly inadequate as theories of cognition. But Davies draws an eliminativist conclusion: since our commonsense understanding of ourselves as thinkers is committed to causal systematicity (via the neo-Fregean conception of thought), and hence to the uniform realization constraint, connectionism is incompatible with the commonsense conception. According to Davies, a being whose cognitive architecture is correctly described as a distributed connectionist network will fail to meet a necessary condition on being a believer. If distributed connectionist models turn out to provide the best accounts of *our* internal architecture, *we* are not believers.

In response to someone who claims that the fact that we are believers is not open to serious question, Davies argues that the only refuge from the eliminativist threat is behaviorism:

> If it is to be non-negotiably true that we who produce interpretable behavior are thinkers, then the concept of a thinker must impose no necessary conditions that go beyond behavior. In particular, it must impose no necessary conditions at all upon internal cognitive architecture. But this means that what the critic wants is a form of behaviorism. . . .
>
> This form of behaviorism is itself arguably incompatible with the commonsense scheme. . . . In any case, if the choice lies between behaviorism and facing up to eliminativism, then there are many of us who know which way we are voting. (1991, 255)

Davies, presumably, intends to cast his vote for eliminativism. But if I am right that folk psychology is properly construed along minimalist lines, then he has posed a false dilemma. Our commonsense scheme does impose (minimal) conditions that go beyond behavior, yet it is not vulnerable to eliminativist arguments of the sort that Davies offers.

Folk psychology, on the minimalist construal, is compatible with the neo-Fregean conception of thought. According to this conception, the inferences a thinker is disposed to make are manifestations of a common underlying capacity, namely, the mastery of a particular concept. The minimalist can happily endorse this claim, although it is far from clear what, if anything, follows from it. Folk psychology has very little to say about concepts, beyond, perhaps, that they are the constituents of thoughts, in the sense of *propositions* or *senses* (i.e., in Frege's sense of "thought"). Mastery of a concept, for all folk psychology has to say about the matter, might involve mentally grasping objects in Platonic heaven. On the other hand, a behaviorist analysis of concepts, which treats them as unanalyzed dispositions to draw certain inferences, is quite compatible with a realist construal of propositional attitudes (according to which beliefs and desires are causally efficacious internal states of agents). Folk psychology makes no commitments concerning what mastery of a concept involves (i.e., how concepts are realized computationally, or how they are deployed in psychological processes). Specifying the psychological mechanisms underlying concept mastery is of no interest to the folk—it is a job left for cognitive scientists. In "unpacking" the neo-Fregean conception of thought in terms of causal systematicity (with its commitment to the uniform realization constraint), Davies builds in substantive commitments about cognitive architecture about which folk psychology is silent.

Let us say, then, that folk psychology is committed, at most, to a *stripped-down* version of the neo-Fregean conception of thought, which claims that the inferences a thinker is disposed to make are manifestations of a common underlying capacity, namely, mastery of a concept, but makes no claims about how the capacity is exercised nor what mastery of a concept involves. (Frege himself would be happier with the stripped-down version, given his famous opposition to the psychologizing of thought.) The question, then, is whether distributed connectionist models of psychological processing are compatible with the stripped-down neo-Fregean conception of thought. There is no reason to think they are not. Nothing prevents us from describing the network's F-involving inferences as manifestations of a mastery of the concept of being F, *provided that mastery of a concept does not require that a common computational state or structure mediates all F-involving inferences*, which, according to the stripped-down neo-Fregean conception of thought, it does not.

The problem with Davies' argument is that the account of concept mastery at play (in the "overdressed" version of the neo-Fregean conception) is itself an integral part of the classical computational picture of thought. It is not surprising, therefore, that distributed connectionist models appear to be incompatible with it. The account of concept mastery implicit in the causal systematicity assumption and the uniform realization constraint does not by itself entail the LOT thesis (because the computational states or structures involved in the relevant inferences do not necessarily constitute a language). Nonetheless, it does involve substantive commitments about psychological processing that go well beyond the folk psychological conception of thought. The failure of connectionist cognitive models to comport with such an account, therefore, has no eliminativist implications.

The lingering worry that distributed forms of connectionism do not sit well with folk psychology may be attributable to a similar source. In the twenty years since the publication of Fodor's seminal book (1975) the classical computational model of the mind has become the received view in the philosophy of mind. It is not surprising that integral components of this conception, such as the account of concept mastery underlying Davies' argument, have seeped into the collective philosophical consciousness to the point where it may be difficult for philosophers to separate these components from the body of theory that they share with ordinary folk. But they should be separated. The perceived tension indicates the extent to which distributed connectionism departs from its more entrenched computational rival; it does not reflect any incompatibility with the folk conception of the mind.

REFERENCES

Astington, J., Harris, P., and Olsen, D. (1988), *Developing Theories of Mind*. New York: Cambridge University Press.

Churchland, P. M. (1981), "Eliminative Materialism and the Propositional Attitudes", *Journal of Philosophy 78*: 67–90.

Davies, M. (1991), "Concepts, Connectionism, and the Language of Thought", in Ramsey, Stich, and Rumelhart (1991), pp. 229–257.

Dennett, D. C. (1987), *The Intentional Stance*. Cambridge, MA: MIT Press.

———. (1991), "Real Patterns", *Journal of Philosophy 88*: 27–51.

Fodor, J. A. (1975), *The Language of Thought*. New York: Thomas Y. Crowell.

———. (1987), *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Cambridge, MA: MIT Press.

Fodor, J. A. and Pylyshyn, Z. (1988), "Connectionism and Cognitive Architecture", *Cognition 28*: 3–71.

Forguson, L. and Gopnik, A. (1988), "The Ontogeny of Common Sense", in Astington, Harris, and Olsen (1988), pp. 226–243.

Forster, M. and Saidel, E. (forthcoming), "Connectionism and the Fate of Folk Psychology", *Philosophical Psychology*.

Heider, F. (1958), *The Psychology of Interpersonal Relations*. New York: Wiley.

Hinton, G. E., McClelland, J. L., and Rumelhart, D. E. (1986), "Distributed Representations", in D. E. Rumelhart, J. L. McClelland, and the PDP Research Group (eds.), *Parallel Distributed Processing, Vol. 1: Foundations*. Cambridge, MA: MIT Press, pp. 77–109.

Horgan, T. and Graham, G. (1991), "In Defense of Southern Fundamentalism", *Philosophical Studies 62*: 107–134.

Horgan, T. and Tienson, J. (eds.) (1991), *Connectionism and the Philosophy of Mind*. Dordrecht: Kluwer.

Jackson, F. and Pettit, P. (1990), "In Defense of Folk Psychology", *Philosophical Studies 59*: 31–54.

Johnston, M. (1992), "Reasons and Reductionism", *The Philosophical Review 101*: 589–618.

Kelley, H. H. and Michela, J. (1980), "Attribution Theory and Research", *Annual Review of Psychology 31*: 457–501.

Lycan, W. (1991), "PDP meets Humuncular Functionalism", in Ramsey, Stich, and Rumelhart (1991), pp. 259–286.

Peacocke, C. (1983), *Sense and Content.* Oxford: Oxford University Press.

Peterson, C. and Seligman, M. E. P. (1984), "Causal Explanations as a Risk Factor for Depression: Theory and Evidence", *Psychological Review 91*: 347–374.

Ramsey, W., Stich, S., and Garon, J. (1991), "Connectionism, Eliminativism, and the Future of Folk Psychology", in Ramsey, Stich, and Rumelhart (1991), pp. 199–228.

Ramsey, W., Stich, S., and Rumelhart, D. (eds.) (1991), *Philosophy and Connectionist Theory.* Hillsdale: Lawrence Erlbaum Associates.

Rey, G. (1991), "An Explanatory Budget for Connectionism and Eliminativism", in Horgan and Tienson (1991), pp. 219–237.

Smolensky, P. (1988), "On the Proper Treatment of Connectionism", *Behavioral and Brain Sciences 11*: 1–74.

Stich, S. P. (1983), *From Folk Psychology to Cognitive Science.* Cambridge, MA: MIT Press.

Van Gelder, T. (1991), "What is the 'D' in 'PDP'? A Survey of the Concept of Distribution", in Ramsey, Stich, and Rumelhart (1991), pp. 33–59.

———. (forthcoming), "Connectionism and the Mind-Body Problem."

Weiner, B. (1990), "Attribution in Personality Psychology", in L. A. Pervin (ed.), *Handbook of Personality: Theory and Research.* New York: Guilford Press, pp. 465–485.

Wellman, H. M. (1990), *The Child's Theory of Mind.* Cambridge, MA: MIT Press.

Wilkes, K. (1986), "Nemo Psychologus nisi Physiologus", *Inquiry 29*: 169–185.

———. (1991), "The Long Past and Short History", in R. Bogdan (ed.), *Mind and Common Sense.* New York: Cambridge University Press, pp. 144–160.

Wimmer, H. and Perner, J. (1983), "Beliefs about Beliefs: Representation and Constraining Function of Wrong Beliefs in Young Children's Understanding of Deception", *Cognition 13*: 103–128.