Published by Blackwell Publishers, 108 Cowley Road, Oxford OX4 1JF, UK and 350 Main Street, Malden, MA 02148, USA
METAPHILOSOPHY
Vol. 33, Nos. 1/2, January 2002
0026–1068

COMPUTER MODELING AND THE FATE OF FOLK PSYCHOLOGY

JOHN A. BARKER

ABSTRACT: Although Paul Churchland and Jerry Fodor both subscribe to the so-called *theory-theory* – the theory that folk psychology (FP) is an empirical theory of behavior – they disagree strongly about FP's fate. Churchland contends that FP is a fundamentally flawed view analogous to folk biology, and he argues that recent advances in computational neuroscience and connectionist AI point toward development of a scientifically respectable replacement theory that will give rise to a new common-sense psychology. Fodor, however, wagers that FP will be largely preserved and vindicated by scientific investigations of behavior. Recent findings by developmental psychologists, I argue, will push both Churchlandians and Fodorians toward the pessimistic view that FP is a misguided theory that will never be displaced, because it is, so to speak, built into our cognitive systems. I explore the possibility of preserving optimism by rejecting the theory-theory and adopting the *simulation theory*, a competing view developed by Robert Gordon, Alvin Goldman, and Jane Heal. According to simulationists, common-sense interpretation of behavior is accomplished by means of pretense-like operations that deploy the cognitive system's own reasoning capabilities in a disengaged manner. Since on this view no theory-like set of principles would be needed, the simulation theory seems to enjoy a simplicity advantage over the theory-theory. Steven Stich and Shawn Nichols, however, contend that as the cognitive system would require special mechanisms for disengaged operation, the simplicity question cannot be resolved until suitable computational models are developed. I describe a set of models I have constructed to meet this need, and I discuss the contribution such models can make to determining FP's fate.

Keywords: AI, autism, computer modeling, folk psychology, mental simulation, mindreading, ProtoThinker, simulation theory, theory-theory, ToMM.

Folk psychology, the common-sense view that human behavior is guided by beliefs, desires, and other mental states, is an enigma. Despite its humble origins, this age-old conception of behavior undergirds much of the theorizing of the cognoscenti in such fields as psychology, sociology, anthropology, economics, and philosophy. But what if folk psychology turned out to be a fundamentally flawed empirical theory destined to be displaced by a neurophysiological theory? Mental entities would join phlogiston, caloric, and other assorted debris on the trash heap of discarded

posits, taking along myriad theories that presuppose the reality of the mental. The social sciences would undergo drastic revisions, and philosophy as we know it might fade away. Paul Churchland is one philosopher who contemplates such an outcome with enthusiasm. He denigrates folk psychology as an untrustworthy prescientific theory analogous to folk physics and folk biology, and celebrates recent advances in computational neuroscience and connectionist AI as initial steps toward development of a scientifically respectable replacement theory (Churchland 1981, 1988, 1989, 1995, 1998). This theory, he predicts, will eventually give rise to a new common-sense psychology: "The new framework, like any other, will gradually work its way into the general population. In time, it will become the common property of folks generally. It will contribute to, or even constitute, a new folk psychology – one firmly rooted, this time, in an adequate theory of the brain" (Churchland 1995, 323).

In opposition to Churchland's Optimistic Eliminativism, Jerry Fodor advocates Optimistic Preservationism: folk psychology is a fundamentally sound empirical theory that will be vindicated rather than be dislodged by scientific investigations of behavior (Fodor 1983, 1987, 1992a, 1992b, 1993, 1998, 2000). Challenging Churchland's presumption that the common-sense conception is a cultural product, Fodor calls attention to folk psychology's cultural universality and emergence in early childhood, and he hypothesizes that it is an innate intellectual endowment: "Homo sapiens is . . . I suspect, uniquely the species that is born knowing its own mind" (Fodor 1987, 133).

In this essay I discuss recent findings by developmental psychologists that at first blush appear to support Fodor's position. I argue, however, that these appearances are misleading, and that the findings may push both Fodorians and Churchlandians toward Pessimistic Preservationism, the disheartening view that folk psychology is, so to speak, radically wrong but built into our brains. I explore the possibility of rescuing Optimistic Preservationism by dispensing with Fodor's presupposition (shared by Churchland) that folk-psychological interpretation of behavior is based on deployment of a theory. This so-called *theory-theory* has a formidable competitor: the *simulation theory*, which has been advanced by Robert Gordon, Alvin Goldman, Jane Heal, and others (Gordon 1986, 1996; Gordon and Barker 1994; Goldman 1989, 1993; Heal 1986, 1995). According to the simulationist, folk-psychological interpretation of an individual's behavior is accomplished by means of "mental simulations" of the individual's decision-making situation. These pretense-like activities, in which the cognitive system operates in a disengaged, or off-line, manner, allegedly generate reliable predictions and explanations without need of a theory. I discuss the role computational models of rational agents can play in adjudicating the dispute between the simulationist and the theory-theorist, and I describe several models I have constructed for this purpose.

Evidence for Innateness

Simon Baron-Cohen, Uta Frith, Alan Leslie, Josef Perner, Heintz Wimmer, and other developmental psychologists have amassed impressive evidence that children possess innate proclivities toward mentalistic interpretations of behavior (Baron-Cohen 1995; Baron-Cohen, Leslie, and Frith 1985; Wimmer and Perner 1983). By the age of four most developmentally normal children can successfully deploy concepts of belief, desire, and other mental states in explaining and predicting behavior. This "mindreading" capability, as Baron-Cohen calls it, emerges spontaneously, with no need of instruction or superior intellectual endowments. Children with Down syndrome, for example, typically perform in accord with their mental ages. Children afflicted with autism, however, exhibit a strange "mindblindness" that selectively impedes their grasp of the mental. Experimental studies by Baron-Cohen and others reveal that even intelligent autistic adults usually perform poorly on "false belief" tests that are easy for most five-year-olds. Suppose, for example, that Sally puts a marble in a basket and departs, whereupon someone else transfers it to a box. When asked where Sally will look for the marble upon returning, autistic individuals of all ages usually indicate the box, a response characteristic of normal three-year-olds. Numerous other tasks requiring understanding of false belief prove to be beyond the ken of most people with autism, irrespective of intelligence or training.

A Nativistic Theory-Theory

In attempting to explain mindreading capability, Baron-Cohen and Leslie, like Fodor and Churchland, adopt the theory-theory approach (Baron-Cohen 1995; Leslie 1987, 1994). Theory-theorists hypothesize that mindreading abilities are grounded in inferential deployment of theory-like sets of principles. If asked how a child manages to pass the above-mentioned false-belief test, a theory-theorist might reply along the following lines: The child derives the conclusion that Sally will look in the basket by invoking premises about Sally's desires, beliefs, and so on, together with such principles as

> (P1) Anyone who wants to find something and believes it to be in a certain place will tend to look for it there.

The premises and principles may be inaccessible to awareness and may be represented in the cognitive system in implicit, nonlinguistic form; they may even be false – as eliminativists are fond of observing, many false theories work fairly well.

Baron-Cohen and Leslie posit the existence of an innate information-processing module, the Theory of Mind Mechanism (ToMM), which

interrelates information about volitional and perceptual states and draws conclusions about the full range of mental states: thinking, believing, knowing, deceiving, imagining, pretending, and so forth. On their view, ToMM grounds mindreading ability, and malfunctions in ToMM account for the core impairments associated with autism.[1]

It may appear that the ToMM theory supports Optimistic Preservationism – scientific research already seems to be making strides toward vindicating folk psychology. But we must not take it for granted that evolutionary processes would have tended to provide ToMM with *true* rather than merely *workable* principles. We must ask ourselves a crucial question: How good are Mother Nature's credentials as a theorist? Unfortunately, her theory-construction record is unimpressive. Indeed, as I shall argue, if folk psychology is Mother Nature's own theory of behavior, then there is good reason to fear that it may be radically mistaken.

Empirical theories that match folk psychology with respect to cultural universality and emergence in early childhood are not easy to find.[2] Perhaps the best candidate is the part of folk physics – call it intuitive physics – that posits the existence of relatively stable objects moving about in various ways under the influence of a variety of forces (pushes, pulls, and so on). Intuitive physics has a conservative, observation-oriented ontology, in that the unobservable objects, forces, trajectories, and so forth that it posits closely resemble observables. Despite this fact, however, key components of the theory are seriously flawed. Consider, for example, the intuitive principle that motion requires force. Notoriously, acceptance of this principle generates numerous difficulties that (thanks to Galileo and Newton) gave rise to a replacement theory containing the unintuitive principle that only change of motion – that is, acceleration or deceleration – requires force. Although this replacement theory has been taught in schools for centuries, it has had little impact on most people's conception of motion. Studies have documented the existence of a deeply ingrained proclivity toward the intuitive principle, a proclivity that not only generates numerous errors in

[1] Some developmental psychologists who adopt the theory-theory approach contend that the normal child's innate endowments are insufficiently determinate to account for acquisition of mindreading capability, and that the child learns folk psychology in ways analogous to those in which scientists develop their theories (see, e.g., Gopnik 1996). This view, however, makes it extremely difficult to explain why normal humans in all cultures acquire and retain the same psychology. (To her credit, Gopnik, unlike many proponents of nonnativistic theory-theories, explicitly acknowledges the existence of a culturally universal conception of human behavior and makes a concerted effort to account for its universality, emergence in early childhood, persistence throughout adult life, and continuing influence on the sciences of behavior.)

[2] Some Fodorians may be tempted to cite universal grammar as an example of an innate theory that is closely akin to folk psychology. But even if universal grammar can be construed as a theory of some kind, it does not seem to be of the right kind. Fodorians claim that folk psychology is the kind of theory that scientific research could in principle show to be fundamentally false. Such a claim about universal grammar would be highly implausible, to say the least.

prediction, explanation, and behavior but also significantly impedes the process of learning the replacement theory (McCloskey 1983).

Although intuitive physics may turn out not to be innate, it appears to exemplify what Mother Nature could have been expected to produce were she to have undertaken theory-construction. She would have tended to posit unobservables that resembled observables and would have been apt to make significant errors when venturing much beyond the observable realm. This point can be illustrated by imagining what would have happened had she decided to equip humans with a built-in explanation of the daily movements of celestial bodies. It seems likely that she would have inculcated a theory that these bodies orbit the earth rather than a theory about the earth's rotation. Now it is plausible that she would have found it quite useful for humans to be born knowing their own minds. If folk psychology is indeed her own theory of human behavior, is it apt to be basically true or, like the celestial-orbit theory, fairly workable but fundamentally misguided?

Folk psychology is an intricately structured complex of principles involving a large number of unanalyzable notions that are related to observables in an indirect and collective fashion. The failed projects of the logical behaviorists inadvertently established that these notions cannot be analyzed in terms of relationships to observables. The history of philosophers' unsuccessful attempts to analyze such concepts as knowledge, belief, desire, and purpose, even in terms of other folk-psychological notions, provides strong evidence that these concepts acquire their natures at least partly from their roles within the complex of principles. These failures did not stem from vagueness or instability of the notions themselves – both proponents and opponents of proffered analyses could typically reach consensus regarding the effectiveness of numerous counterexamples, something that attests to the existence of a shared fund of intuitively grasped principles.

Clearly, then, folk psychology is a high-level interpretative framework containing a theory-oriented, and hence misconception-prone, ontology. And since many key folk-psychological principles are descriptive only of language-endowed creatures, our distant ancestors would not have constituted fitting subjects for the protracted experimentation that would have been required for discovery and inculcation of principles that correctly explain the behavior of language-using humans. The conclusion seems inescapable that in constructing folk psychology Mother Nature cannot be expected to have hit upon a fundamentally accurate representation of the deeply hidden springs of human behavior. If Fodor is right that folk psychology is an *innate* theory, then there is good reason to fear that Churchland may be right that it is a *false* theory. The bleak prospect of Pessimistic Preservationism looms – folk psychology may turn out to be a misguided theory that will never be displaced, because it is built into our cognitive systems.

Drawing inspiration from Daniel Dennett (1984) and Donald Davidson (1984), we could forestall this disconcerting development by construing folk psychology not as an empirical theory but as a collection of principles that specify what it takes to qualify as a genuine rational agent. We could posit the existence of an innate Rational Agency Mechanism (RAM), which contains representations of the principles and enables the mindreading child to deploy them in a pretense-like fashion. In dealing with the false-belief situation mentioned above, for example, the normal four-year-old could produce the correct prediction by taking "the intentional stance" toward Sally, pretending that she qualifies as a rational agent, as someone who by definition would tend to look for things where they were thought to be. Autistic children (who are known to suffer from pretense-behavior deficits) would fail to produce the correct answer because they could not make proper use of RAM. To help explain how mindless evolutionary processes could have managed to supply RAM with such a remarkably descriptive set of prescriptive principles, we could hypothesize that, via maturation and socialization, the principles themselves tend to shape human behavior, thereby becoming more descriptive of it.[3]

The RAM theory of mindreading capability would have difficulty providing folk psychology with genuine explanatory capability, and it would most likely enjoy no significant advantages over the ToMM theory with respect to simplicity. Furthermore, even though the RAM theory would help dispel Pessimistic Preservationism's air of despair, it would retain so much of its substance that rescuing Optimistic Preservationism would be out of the question. Let us therefore consider a seemingly minor modification in the approach: let us hypothesize that the typical four-year-old is a (reasonably) rational agent who can pretend *to be* Sally and, simply by reasoning as she would, generate the correct prediction. This maneuver, which dispenses with the need to posit representations of folk-psychological principles, exemplifies the guiding strategy of the simulationist.

A Nativistic Simulation Theory

According to the simulation theory, mindreaders do not need a theory about mental states or a conception of rationality to predict and explain behavior. Instead, their cognitive systems employ a pretense-like process to construct "mental simulations" of a target individual's decision-making situation. Their own reasoning capabilities are deployed in a disengaged, or off-line, manner, operating on "pretend" beliefs, desires, feelings, intentions, and so on and yielding generally reliable information about the individual's mental

[3] This hypothesis might generate legitimate concern about the desirability – and even the rationality – of rationality as defined by the principles in RAM, unless it could be shown that any deleterious or arbitrary principles would tend to be eliminated over time, and that human lives are not now being shaped by a specification of rationality that was appropriate only for our primitive ancestors.

states and associated actions. The culturally universal proclivity toward mentalistic interpretation of behavior is grounded in an innate capacity for mental simulation rather than in an innate theory of mind.[4]

Utilization of the mindreader's own reasoning and decision-making system can generate reliable predictions and explanations of conspecifics because human minds generally function in similar ways. There is no need for explicit or implicit representations of psychological principles – the requisite principles are embodied as operating regularities in the cognitive systems of both the simulating persons and the simulated persons. In the false-belief situation described above, for example, the mindreader's system operates in an off-line manner, forming a desire to find the marble, a belief that it is in the basket, and a rational decision to look there; the simulation culminates in the formation of the on-line belief that Sally will look in the basket. The mindreader need not possess introspective ability or understanding of self; indeed, engaging in primitive forms of simulation grounds acquisition of such ability and understanding. Nor need the mindreader employ any assumptions about what is transpiring – the entire simulation process can take place automatically, with no conscious awareness on the part of the simulator.[5]

To counter the threat of Pessimistic Preservationism, the simulationist can begin by acknowledging that the mindreader's cognitive system operates as if it contained representations of folk-psychological principles, such as

> (P1) Anyone who wants to find something and believes it to be in a certain place will tend to look for it there.

Hence, the system can be described as containing *operative representations* of the principles. The simulationist and the theory-theorist can agree that the mindreader possesses a large fund of such operative representa-

[4] The version of simulationism presented here should not be taken to represent all of the many extant versions. For example, this "externalist" simulationism, unlike Heal's "internalist" version, explicates mindreading from a third-person rather than a first-person perspective; unlike Goldman's "introspectivist" version, it entails mindreading not requiring introspective capability; and unlike Gordon's "radical" version, it entails mental state ascriptions not being intrinsically linked to mental simulation, so mindreading could in principle be accomplished without it.

[5] While mental simulations can be usefully viewed as inference-like processes that yield "conclusions," the mindreader need not harbor any assumptions to the effect that the two cognitive systems work in similar ways – no analogical inferences are involved. It is the simulation theory itself that contains the assumption that human cognitive systems generally work alike. Hence the theory, if correct, entails and explains simulation-based mindreading as reliably able to generate true beliefs and predictions about behavior, and correct explanations thereof. (Whether or not such mindreading usually provides the mindreader with justified beliefs or with genuine knowledge is a large question that is not being addressed here.)

tions, a fund constituting what amounts to an *operative theory* of behavior. (To the simulationist, of course, this operative theory is a "virtual" rather than an actual theory, consisting as it does of "virtual" rather than actual representations.) When it comes to dealing with Pessimistic Preservationism, the simulationist enjoys a distinct advantage over the theory-theorist: if simulationism is true, then the mindreader possesses *an operative theory of behavior that cannot be radically misguided*, for its contents reflect the operating regularities of the human cognitive system. There would have been no need for Mother Nature to undertake theory construction. Having produced a basic version of the human cognitive system, she could initiate primitive mindreading activity simply by getting the system to function in a pretense-like manner. And given her penchant for inculcating capacities for mimicry, deception, and pretense in so many of her creatures, she would be apt to adopt this simple strategy to equip humans with an innate device for discovering their minds.

In place of ToMM, the simulationist can posit the existence of what may be called a Disengaged Operation Mechanism (DOM), an innate control mechanism that enables the cognitive system to disengage from normal input-output channels, form and process appropriate mental states, and ascribe them to the relevant individual when engaged operation resumes. Because the system functions in accord with regularities characteristic of mental states, DOM provides the normal child with a "head start" toward successful mindreading of conspecifics before any actual representations of principles governing mental goings-on have been acquired. Primitive forms of mindreading lead to development of the standard concepts of belief, desire, and so on – since the concepts of mental states the normal child acquires will be characterized by relationships that mirror the nomic regularities governing the states themselves, there is no need to posit either innate mental-state concepts or concept-acquisition processes that rely on detection of culture-transcending patterns within human behavior. Mindreading practice generates a continuously expanding fund of actual representations of folk-psychological principles, culminating in the accomplished mindreader's consciously accessible conception of human behavior.[6] Concerning autism, abnormalities in DOM's functioning could account for mindreading impairments and the absence of even a basic understanding of human behavior. Furthermore, the fact that DOM operates in a pretense-like fashion could explain the puzzling correlation between such impairments and pretense-behavior deficits linked to autism (Gordon and Barker 1994).

[6] Although simulational activity (according to this account) initiates the mindreading process in young children, such activity is rapidly and continuously supplemented with formation and use of representations of psychological principles acquired and refined via experience and communication, as well as by additional simulational practice. As many simulationists have emphasized, simulationism need not claim that mental simulation is the only mindreading heuristic, or even the one that most mindreaders predominantly employ.

Computational Models of Rational Agents

Stephen Stich and Shawn Nichols (1992) have pointed out that, despite initial appearances, the simulation theory may not enjoy any advantages over the theory-theory with regard to simplicity. According to the simulationist, mindreaders do not need a special database of psychological principles, for off-line reasoning deploys the principles embodied in the cognitive system. According to the theory-theorist, however, mindreaders do not need a special off-line control mechanism, for the principles in the database are deployed via ordinary reasoning processes. Stich and Nichols contend that this dispute cannot be settled until suitable computational models are developed: "While the simulation theorist gets the data base for free, it looks like the theory-theorist gets the 'control mechanism' for free. . . . But we don't think either side of this argument can get much more precise until we are presented with up-and-running models to compare. Until then, neither side can gain much advantage by appealing to simplicity" (Stich and Nichols 1992, 53).[7]

In an effort to meet this need, I have developed a software program containing three simple computational models of a rational agent, ProtoThinker (Barker 1998, 1999, 2001). Menu options enable the user to activate and experiment at will with each of the three models. The user interacts with ProtoThinker (PT) through conversation in natural language and can opt to have PT's mental processing displayed in varying degrees of detail to facilitate analysis and evaluation of the modeling strategies.

The three models are the *Core-model*, the *DOM-model*, and the *ToMM-model*, the latter two consisting of the Core-model supplemented with mechanisms that provide PT with mindreading and related capabilities. In all three models PT can be usefully viewed as understanding statements, questions, and requests, forming thoughts and memories, reasoning deductively and inductively, and making rational decisions. In the Core-model, however, PT cannot engage in mindreading, pretending, deceiving, identifying empathetically with others, or understanding metaphorical discourse. As absence of these capabilities in children has been linked to autism, PT can be construed as exhibiting some of the characteristic deficiencies of this affliction. These deficiencies are "remedied," at least to a modest extent, by the mechanisms added to the Core-model to form the DOM-model and the ToMM-model.

[7] It may seem that the simulationist can easily win this simplicity competition – even the theory-theorist must acknowledge that normal children already possess a suitable off-line control mechanism, i.e., the mechanism that enables them to engage in ordinary pretense. But some theory-theorists (e.g., Leslie and Baron-Cohen) claim that in order to engage in pretense the child must possess beliefs about pretending, and hence must deploy a theory of mind in the process. Even if this claim could be shown to be predicated on a failure to distinguish carefully between engaging in pretense and understanding pretense, the simulationist would still need to establish that the mechanism that grounds ordinary pretense is also used for mindreading.

The DOM-model, inspired by the simulation theory, contains several disengaged operation mechanisms that give PT limited abilities to mindread, pretend, deceive, empathize, and understand metaphorical statements. They also provide simple forms of additional capabilities that appear to utilize disengaged operations – constructing conditional and indirect proofs, engaging in hypothetico-deductive and analogical reasoning, and predicting grammaticality judgments. The ToMM-model, inspired by the theory-theory, differs from the DOM-model in only one significant respect: PT's mindreading capability is implemented via a separate module called ToMM.[8] As I explain below, the functional distinctness of ToMM from PT's own cognitive operations ensures that the principles employed in mindreading could in principle differ radically from the principles embodied in these operations. In the DOM-model, in contrast, PT's own cognitive operations are themselves deployed in mindreading (albeit in a disengaged manner), and hence the relevant sets of principles are essentially identical.

Currently PT's mindreading abilities are limited to belief prediction, and they do not yet extend to desire prediction, behavior prediction, and so on. To facilitate comparison of the DOM-model and the ToMM-model with respect to simplicity, PT has been given equivalent belief-prediction abilities in each. In both models PT operates as if she possessed actual representations of thousands of folk-psychological principles, such as

> (P2) If S believes that X is an A and that every A is a B, then S tends to believe that X is a B.

In other words, PT possesses operative representations of these principles, which taken together constitute the content of her operative theory of belief. By design, PT possesses the same operative theory in both models; but the way in which the theory is implemented differs significantly. In the DOM-model, the content of the theory is directly linked to, and in effect incorporates, numerous principles embodied in PT's own cognitive system. In the ToMM-model, however, the content is wholly contained in the ToMM module, and the principles constituting this content could have differed radically from PT's own principles – ToMM could have been given a theory of belief that was not even roughly descriptive of PT's own

[8] The ToMM-model, which was absent from the earliest versions of ProtoThinker, is contained in DOS Version 3.1, Windows Version 4.1, and in all subsequent versions. Windows Version 5.1, which can be downloaded from www.mind.ilstu.edu/research/ai/pt, contains the best implementation of all three models and is the version recommended for experimentation with PT's mindreading capabilities. In the default setting of this version, the DOM-model is active. To activate the ToMM-model, click the "Abilities" button on the main screen and select "Theory-based mindreading." To activate the Core-model, click the "Abilities" button and select "Autism." (For a discussion of computational models of simulational processes from an AI perspective, and arguments for the efficiency of such models as compared with theory-theory models, see Barnden 1995.)

cognitive system. Because at present ToMM's theory is highly descriptive of PT's system, the existing ToMM-model accords with Fodorian intuitions. In a suitable Churchlandian model, ToMM's theory would be radically misguided.

Modeling Mindreading Capability

The DOM-model and the ToMM-model utilize the same belief-prediction strategies. It turned out to be necessary to employ a disengaged operation mechanism in the ToMM-model as well as in the DOM-model because reliable belief prediction requires inference replication. To achieve reliability, the mindreader must attend to the propositions constituting the contents of the target individual's known beliefs and utilize these propositions in off-line inferences that roughly accord with the individual's own inferences.[9] To illustrate this, suppose we give a math test to three students, Alice, Betty, and Carol, and try to predict their answers. Alice, let us suppose, is good at math, Betty always divides when she should multiply, and Carol thinks that the word *five* designates the number 9. Those of us who are mathematically challenged will be poor predictors, no matter how expert we are at psychology – we shall have trouble replicating the correct and the incorrect inferences that the students are apt to make.

The following printout of a conversation with PT illustrates the steps in the process by which she attempts to predict beliefs. (For the sake of brevity, only a small portion of PT's mental processing has been reproduced.) Given information about an individual's beliefs, PT simulates possession of the beliefs, tries to replicate the individual's inferences, and, if all goes well, ascribes belief in the conclusions to the individual. To increase the chances that her own off-line inferences really will replicate the individual's inferences, PT always tries to make appropriate adjustments in the level of reasoning competence she uses within the scope of her simulations.

As the conversation begins, PT is told that a certain individual, Brown, believes that Smith is a banker and that every banker is rich. PT is also told that Brown is intelligent. When asked whether Brown believes that Smith is not poor, PT simulates possession of Brown's beliefs and, in view of Brown's high level of reasoning ability, makes a concerted effort to derive the proposition within the context of the simulation.

[9] See Heal 1995 for detailed arguments for the indispensability of inference replication. It may appear that eliminativists could not accept the thesis that reliable belief prediction requires inference replication, as it seems to presuppose realism about beliefs and inferences. But examination of the manner in which inference replication is accomplished in the ToMM-model will indicate that eliminativists need not reject the thesis, even though they would probably choose to word it somewhat differently.

Conversation 1

Entry: Brown believes that Smith is a banker.
PT's response: I can accept that statement.
Entry: Brown believes that every banker is rich.
PT's response: That's intriguing.
Entry: Brown is intelligent.
PT's response: I did not know that.
Entry: Does Brown believe that Smith is not poor?
(*Simulating subject's beliefs*)
Simulated belief: Smith is a banker.
Simulated belief: Every banker is rich.
(*Assessing subject's reasoning ability; simulating* HIGH *level of reasoning ability*)
Premise: Every banker is rich.
Conclusion: Every banker is wealthy.
Premise: Every banker is rich.
Conclusion: Every banker is not poor.
Premise: Every banker is rich.
Premise: Smith is a banker.
Conclusion: Smith is rich.
Premise: Every banker is wealthy.
Premise: Smith is a banker.
Conclusion: Smith is wealthy.
Premise: Every banker is not poor.
Premise: Smith is a banker.
Conclusion: Smith is not poor.
(*Simulation completed; ascribing relevant belief to subject*)
PT's response: Yes.

By mentally simulating Brown's cognitive situation and attempting to replicate his inferences, PT succeeded in identifying additional beliefs that Brown most likely holds. If we erased PT's thoughts and repeated the conversation without supplying the information that Brown is intelligent, PT would attenuate her reasoning activity within the scope of the simulation; as a result, she would not reach the conclusion that Smith is not poor, and would report that she cannot tell whether Brown believes the proposition. If we asked her whether Brown believes that Smith is wealthy, she would say "Yes," as her off-line reasoning would generate this conclusion with relatively little effort.

Mindreaders must take account of the fact that normally reliable inferences about the beliefs of others, just like normally reliable inferences about world situations, are capable of being defeated by special circumstances. For example, when reasoning leads to tension-generating consequences (for example, conclusions that obviously conflict with existing

beliefs), the reasoner usually tries to make adjustments, such as rejecting certain premises or refraining from accepting certain conclusions. PT models this process not only in her own reasoning but also, as the following conversation printout shows, in attempting to replicate the reasoning of others.

As the conversation begins, PT is told that Green believes the following propositions: that Smith is a banker, that every banker is rich, and that Smith is not wealthy. PT is also told that Green is intelligent. When asked whether Green believes that Smith is not poor, PT simulates possession of Green's beliefs, discovers the inconsistency within Green's set of beliefs, and refrains from attributing to Green the belief she would normally ascribe.

Conversation 2

Entry: Green believes that Smith is a banker.
PT's response: I will bear that in mind.
Entry: Green believes that every banker is rich.
PT's response: I can accept that assertion.
Entry: Green believes that Smith is not wealthy.
PT's response: That's nice to know.
Entry: Green is intelligent.
PT's response: I can believe that.
Entry: Does Green believe that Smith is not poor?
(*Simulating subject's beliefs*)
Simulated belief: Smith is a banker.
Simulated belief: Every banker is rich.
Simulated belief: Smith is not wealthy.
(*Assessing subject's reasoning ability; simulating* HIGH *level of reasoning ability*)
Premise: Every banker is rich.
Conclusion: Every banker is wealthy.
Premise: Every banker is rich.
Conclusion: Every banker is not poor.
Premise: Every banker is rich.
Premise: Smith is a banker.
Conclusion: Smith is rich.
Premise: Every banker is wealthy.
Premise: Smith is a banker.
Conclusion: Smith is wealthy.
Premise: Every banker is not poor.
Premise: Smith is a banker.
Conclusion: Smith is not poor.
(*Simulation completed; contradiction detected*)
PT's response: I cannot tell – the individual holds inconsistent beliefs.

Even though PT knows that Smith's not being poor is derivable from the propositions Green believes, she refrains from attributing belief in this conclusion to him because she encounters an inconsistency while attempting to replicate his inferences. In effect, PT assumes that had Green deployed his own reasoning abilities extensively enough to reach this conclusion, he would probably have detected the problem, and his set of beliefs would have undergone some sort of readjustment. Thus, PT models ways in which we try to make allowances not only for variations in people's reasoning abilities but also for their tendencies to try to maintain coherence within their belief systems.

The Simplicity Contest: Which Theory Wins?

Implementation of PT's belief-prediction abilities within the DOM-model required only a small amount of coding. Starting with the Core-model, I added a disengaged operation mechanism, DOM, to enable PT to employ inference replication in such predictions. DOM in effect extracts the propositional contents of metacognitive representations of the form "S believes that P," enables the inference engine to derive conclusions while operating in a disengaged and, if necessary, attenuated manner, searches for inconsistencies in the resultant set of propositions, and incorporates the relevant conclusions into metacognitive representations of the same form if no problems are encountered.

DOM enables PT to predict beliefs without utilizing any actual representations of psychological principles. Because PT's inference engine incorporates hundreds of inference rules that can be combined in numerous ways, DOM automatically endows PT with operative representations of thousands of psychological principles. For example, because PT's inference engine embodies the rule "Given that X is an A and that every A is a B, derive X is a B," DOM provides PT with an operative representation of the psychological principle discussed above:

> (P2) If S believes that X is an A and that every A is a B, then S tends to believe that X is a B.

Indeed, owing to PT's ability to detect conflicts within a set of beliefs, DOM provides her with an operative representation of the following, more sophisticated, version of (P2):

> (P3) If S believes that X is an A and that every A is a B, and S's inferring that X is a B would not tend to generate any conclusions that conflict with other propositions that S believes, then S tends to believe that X is a B.

Moreover, DOM enables PT to take account of variations in reasoning ability and of semantic entailments associated with terms involved in

belief statements – for example, the entailment between being rich and not being poor. This has the effect of providing PT with operative representations of numerous predictively powerful principles like the following one, which was implicated in the two sample conversations above:

> (P4) If S believes that X is an A and that every A is a B, and X's being a B semantically entails X's not being a C, and S has a high level of reasoning ability, and S's inferring that X is not a C would not tend to generate any conclusions that conflict with other propositions that S believes, then S tends to believe that X is not a C.

In undertaking the task of constructing a ToMM-model with belief-prediction capabilities equivalent to those in the DOM-model, I encountered a major problem. My initial efforts were guided by the standard theory-theorist view that mindreading is based on deployment of actual representations of psychological principles by means of ordinary reasoning processes. It quickly became apparent, however, that principles like (P3) and (P4) could be deployed only via inference replication, something that would require disengaged use of the model's inference engine.

After numerous unsuccessful attempts to dispense with inference replication, I concluded that ToMM needed its own inference engine, and a disengaged operation mechanism as well. I then realized that providing these would obviate the need for actual representations of psychological principles. A massive fund of operative representations could be supplied in a very compact form – I could provide ToMM with a replica of the DOM-model's inference engine, including its disengaged operation mechanism. This strategy enabled me to construct a ToMM-model that inherited the DOM-model's entire operative theory of belief. When a belief-prediction task is undertaken, the main inference engine sends all information about the target individual to ToMM. ToMM extracts the propositional contents of the individual's beliefs, uses its own inference engine to process them with a sophistication matching that of the DOM-model, and outputs appropriate belief predictions. This design achieves the goal of providing the ToMM-model with belief-prediction capabilities equivalent to those of the DOM-model.

There are several features of the ToMM-model's design that are advantageous from a theory-theory perspective. First, the model accords with the theory-theorist's view that mindreading is grounded in possession of a theory that could, in principle, turn out to be radically misguided. If extensive alterations were made in ToMM without modifying the main inference engine (or vice versa), ToMM would possess an operative theory of belief that was not even roughly descriptive of PT's own cognitive system.

Second, the model accords with Churchland's strategy for dealing with the following vexing problem confronting theory-theorists: "If one's capacity for understanding and predicting the behavior of others derives

from one's internal storage of thousands of laws or nomic generalizations, how is it that one is so poor at enunciating the laws on which one's explanatory and predictive prowess depends? It seems to take a trained philosopher to reconstruct them! How is it that children are so skilled at understanding and anticipating the behavior of humans in advance of ever acquiring the complex linguistic skills necessary to express those laws?" (Churchland 1988, 217). To address this problem, Churchland sketches a revised theory-theory that employs the notion of a prototype: "A normal human's understanding of the springs of human action may reside not in a set of stored generalizations about the hidden elements of mind and how they conspire to produce behavior, but rather in one or more prototypes of the deliberative or purposeful process. To understand or explain someone's behavior may be less a matter of deduction from implicit laws, and more a matter of recognitional subsumption of the case at issue under a relevant prototype" (Churchland 1988, 218). Because the ToMM module contains a replica of the core of PT's cognitive system, it functions as a prototype of a rational thinker and dispenses with the need for inferential deployment of representations of folk-psychological principles. The fact that the module contains "virtual" representations of thousands of folk-psychological principles indicates how Churchland's new theory-theory could follow the lead of the simulation theory in construing the sophisticated mindreader's actual representations of such principles and capacities to provide them with verbal garb, as "effects" rather than "causes" of basic mindreading capability.

Finally, the design of the ToMM-model suggests a way in which Churchland could handle a problem confronting his new prototype-based theory-theory. Churchland hypothesizes that the prototypes needed for mindreading can be modeled in "artificial neural networks, networks that mimic some of the more obvious organizational features of the brain" (Churchland 1988, 219). But to model the mindreader's capacity to engage in inference replication, which involves tracking logical relationships among propositions constituting the contents of mental states, such networks would probably have to implement symbol-processing systems, and Churchland generally eschews positing symbol-processing systems to explain human capabilities. By drawing inspiration from the ToMM-model, however, the theory-theorist could envision an artificial neural network that models mindreading via a separate symbol-processing module, a module that is not directly involved in other activities and does not presuppose commitment to realism regarding contentful mental states.[10]

[10] See Fodor and Pylyshyn 1988 and Ramsey, Stich, and Garon 1991 for discussion of limitations of artificial neural networks with respect to modeling cognitive-level representations without implementing symbol-processing systems. For a discussion of networks that are capable of implementing symbol-processing systems, see Horgan and Tienson 1996.

Thus, the ToMM-model not only serves to clarify and test the theory-theory but also suggests promising remedies for some of its problems. The computational resources employed for implementation of belief-prediction capability in this model, however, are hundreds of times greater in volume, and far more complex in structure, than those used for this purpose in the DOM-model. This outcome suggests that the simulation theory will turn out to enjoy a simplicity advantage over the theory-theory. If so, we shall have reason to think that simulationism provides the correct explanation for mindreading capability, and hence that folk psychology is fundamentally sound. Of course, exploration of alternative modeling strategies, utilizing neural-network architecture as well as the classical architecture exemplified by ProtoThinker, must be conducted before a warranted verdict can be reached. Meanwhile, however, the prospects for using simulationism to rescue Optimistic Preservationism appear bright.[11]

Department of Philosophical Studies
Southern Illinois University, Edwardsville
Edwardsville, IL 62026
USA
jbarker@siue.edu

References

Astington, J. W., P. L. Harris, and D. R. Olson, eds. 1988. *Developing Theories of Mind*. Cambridge: Cambridge University Press.
Barker, J. A. 1998. *ProtoThinker: A Model of the Mind*. DOS Version 3.1. Belmont, Calif.: Wadsworth Publishing.
———. 1999. *ProtoThinker: A Model of the Mind*. Windows Version 4.1. Belmont, Calif.: Wadsworth Publishing.
———. 2001. *ProtoThinker: A Model of the Mind*. Windows Version 5.1. Available for downloading from www.mind.ilstu.edu/research/ai/pt.
Barnden, J. A. 1995. "Simulative Reasoning, Commonsense Psychology, and Artificial Intelligence." In Davies and Stone 1995b, 247–73.
Baron-Cohen, S. 1995. *Mindblindness: An Essay on Autism and Theory of Mind*. Cambridge, Mass.: MIT Press.
Baron-Cohen, S., A. Leslie, and U. Frith. 1985. "Does the Autistic Child Have a Theory of Mind?" *Cognition* 21: 37–46.
Carruthers, P., and P. K. Smith, eds. 1996. *Theories of Theories of Mind*. Cambridge: Cambridge University Press.
Christensen, S. M., and D. R. Turner, eds. 1993. *Folk Psychology and the Philosophy of Mind*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

[11] I am indebted to Fred Adams, David Anderson, Gary Fuller, Robert Gordon, and James Moor for helpful discussions concerning this essay. An earlier version was presented at the 2000 meeting of the American Philosophical Association, Pacific Division.

Churchland, P. M. 1981. "Eliminative Materialism and the Propositional Attitudes." *Journal of Philosophy* 78: 67–90. Reprinted in Churchland 1989, 1–22.

———. 1988. "Folk Psychology and the Explanation of Human Behavior." *Proceedings of the Aristotelean Society*, suppl. vol. 62: 209–21. Reprinted in Churchland 1989, 111–27.

———. 1989. *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science*. Cambridge, Mass.: MIT Press.

———. 1995. *The Engine of Reason, the Seat of the Soul*. Cambridge, Mass.: MIT Press.

Churchland, P. M., and P. S. Churchland. 1998. *On the Contrary: Critical Essays, 1987–1997*. Cambridge, Mass.: MIT Press.

Clark, A., and P. J. R. Millican. 1996. *Connectionism, Concepts and Folk Psychology: The Legacy of Alan Turing*. Oxford: Oxford University Press.

Davidson, D. 1984. *Inquiries into Truth and Interpretation*. Oxford: Oxford University Press.

Davies, M., and T. Stone, eds. 1995a. *Folk Psychology and the Theory of Mind Debate*. Oxford: Blackwell Publishers.

Davies, M., and T. Stone, eds. 1995b. *Mental Simulation: Evaluations and Applications*. Oxford: Blackwell Publishers.

Dennett, D. 1984. *The Intentional Stance*. Cambridge, Mass.: MIT Press.

Fletcher, G. 1995. *The Scientific Credibility of Folk Psychology*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Fodor, J. 1983. *The Modularity of Mind: An Essay on Faculty Psychology*. Cambridge, Mass.: MIT Press.

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

———. 1992a. "A Theory of the Child's Theory of Mind." *Cognition* 44. Reprinted in Davies and Stone 1995b, 109–22.

———. 1992b. *A Theory of Content and Other Essays*. Cambridge, Mass.: MIT Press.

———. 1993. *The Elm and the Expert: Mentalese and Its Semantics*. Cambridge, Mass.: MIT Press.

———. 1998. *In Critical Condition: Polemical Essays on Cognitive Science and the Philosophy of Mind*. Cambridge, Mass.: MIT Press.

———. 2000. *The Mind Doesn't Work That Way: The Scope and Limits of Computational Psychology*. Cambridge, Mass.: MIT Press.

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

Goldman, A. 1989. "Interpretation Psychologized." *Mind and Language* 4: 161–85. Reprinted in Davies and Stone 1995a.

———. 1993. "Empathy, Mind, and Morals." *Proceedings and Addresses of the American Philosophical Association* 67. Reprinted in Davies and Stone 1995b, 185–208.

Gopnik, A. 1996. "Theories and Modules: Creation Myths, Developmental Realities, and Neurath's Boat." In Carruthers and Smith 1996, 169–83.
Gopnik, A., and H. M. Wellman. 1995. "Why the Child's Theory of Mind Really Is a Theory." In Davies and Stone 1995a, 232–58.
Gordon, R. M. 1986. "Folk Psychology as Simulation." *Mind and Language* 7: 11–34. Reprinted in Davies and Stone 1995a.
———. 1996. "'Radical' Simulationism." In Carruthers and Smith 1996, 11–21.
Gordon, R. M., and J. A. Barker. 1994. "Autism and the Theory of Mind Debate." In *Philosophical Psychopathology*, edited by G. Graham and G. Stevens, 163–81. Cambridge, Mass.: MIT Press.
Greenwood, J. D., ed. 1991. *The Future of Folk Psychology: Intentionality and Cognitive Science*. Cambridge: Cambridge University Press.
Haselager, W. F. G., ed. 1997. *Cognitive Science and Folk Psychology: The Right Frame of Mind*. London: Sage Publications.
Heal, J. 1986. "Replication and Functionalism." In *Language, Mind, and Logic*, edited by J. Butterfield, 135–50. Cambridge: Cambridge University Press.
———. 1995. "How to Think about Thinking." In Davies and Stone 1995b, 33–52.
Horgan, T., and J. Tienson. 1996. *Connectionism and the Philosophy of Psychology*. Cambridge, Mass.: MIT Press.
Leslie, A. 1987. "Pretense and Representation: The Origins of 'Theory of Mind.'" *Psychological Review* 94: 412–26.
———. 1994. "ToMM, ToBy, and Agency: Core Architecture and Domain Specificity." In *Mapping the Mind: Domain Specificity in Cognition and Culture*, edited by L. Hirschfield and S. Gelman. Cambridge: Cambridge University Press.
McCloskey, M. 1983. "Intuitive Physics." *Scientific American* (Apri): 122–30.
Ramsey, W., S. Stich, and J. Garon. 1991. "Connectionism, Eliminativism, and the Future of Folk Psychology." In Greenwood 1991, 93–119.
Stich, S., and S. Nichols. 1992. "Folk Psychology: Simulation or Tacit Theory?" *Mind and Language* 7: 35–71.
———. 1997. "Cognitive Penetrability, Rationality and Restricted Simulation." *Mind and Language* 12: 297–326.
Wimmer, H., and J. Perner. 1983. "Beliefs about Beliefs: Representation and Constraining Function of Wrong Beliefs in Young Children's Understanding of Deception." *Cognition* 13: 103–28.