The Naive Psychology Manifesto

Stuart Watt
Human Cognition Research Laboratory
The Open University, Milton Keynes, MK7 6AA
Email: `S.N.K.Watt@open.ac.uk`

Abstract

This paper argues that artificial intelligence has failed to address the whole problem of common sense, and that this is the cause of a recent stagnation in the field. The big gap is in common sense—or naive—psychology, our natural human ability to see one another as minds rather than as bodies. This is especially important to artificial intelligence because AI must eventually enable us humans to see computers not as grey boxes, but as minds.

The paper proposes that AI study exactly this—what is going on in people's heads that makes them see others as having minds. To illustrate this, it describes three models for common sense psychology; each of which illustrates some aspect of the human ability to see people as minds. The paper concludes by drawing some conclusions about where and how AI can adapt to deal with these issues and to provide a new and stronger foundation for future research.

1 Introduction: from naive physics to naive psychology

Ten years ago, Hayes published the "Second naive physics manifesto" [24]. Hayes proposed that we "put away childish things by building large-scale formalisations," beginning with "our knowledge of the everyday physical world." He, and others, have since put a lot of effort into developing models of our common sense understanding of the physical world.

But life in the common sense community isn't a bed of roses. Common sense has generally been a big problem for artificial intelligence, and despite the attempts of many brave souls (e.g. Hayes, McCarthy, McDermott, and Lenat) it hasn't really yielded at all: "the common sense knowledge problem has blocked all progress in theoretical artificial intelligence" [17]. For twenty years there has been only slow progress in the field, and although many important technologies have been found while looking for ways to deal with common sense (e.g. nonmonotonic reasoning,) the core of the problem is still almost untouched.

The problems for common sense have generally come in two classes, technical and methodological. On the technical side, for example, there is the 'frame problem' [32], which Dennett [15] describes as the "smoking pistol" behind a lot of the attacks on artificial intelligence, and as a major philosophical problem for everybody. The methodological criticisms for the project have already been adequately aired, but none of these really attack the logicist position as a descriptive project, merely as a causal account of common sense reasoning [8, 33].

This paper will present a position orthogonal to that of Hayes. Hayes' initiative clarified many of the issues associated with common sense, and other developments in comparative and developmental psychology have further highlighted the apparently fundamental nature of naive physics [6, 40], but they have also revealed a deeper and bigger problem than that of naive physics—naive psychology.

Naive psychology [7, 24] can be thought of as the natural human ability to infer and reason about other people's mental states—in short, to see one another as minds rather than bodies. This is an issue that artificial intelligence must also address. Although people see one another as minds not simply bodies, they just don't see computers as minds in the same way. If artificial intelligence is to overcome this barrier, we humans must be able to see minds in artifacts, to ascribe mental states to artificial intelligences in the same way that we do to people.

An example from Haugeland [22] will make this clearer. Imagine that somebody presents you with an object and tells you that it can play chess. How can they convince you that it really does? You can look at the behaviour of the system, but that might be open to different interpretations depending on how the system communicates its moves. You decide by looking at its behaviour, interpreting it in a consistent way, and checking whether the moves that it communicates are legal and sensible with respect to the rules of chess as you know them. Haugeland's point is that the meaning of the behaviour isn't an intrinsic property of the behaviour, but is "attributed in an empirically justified interpretation" [22]. But this ability to make the interpretation is itself in part a property of you, the judge.

Now expand the problem to the whole of human interaction, which is similarly game-like. The 'moves' can again be interpreted in many different ways, but if the moves are to tell a coherent story, they must be compatible with the laws and heuristics of other people's interpretation. Human interaction, like chess, isn't usually learned from books of rules, but by people learning from one another. It depends more on human agreement than adherence to sets of explicitly specified rules. Just as learning chess from watching only one player's moves would be almost impossible, the same is true of human social interaction. And any good chess program will depend on a player's *understanding* of the moves, in terms of (often social) constructs such as 'attacking' and 'supporting.'

Following on from this game metaphor for human interaction, there is evidence that a lot of human intelligence is 'Machiavellian' [4] in the way people use it to outwit each other and to recognise and manipulate one another's mental states; our social environments are considerably more complex than our physical ones. Survival in these social environments require us to become "natural psychologists" [27], capable of recognising and reasoning about one another's mental states from their behaviour. Humphrey [27] even suggests that naive physics may, in part, be itself derived from a leaky naive psychology; he suggests that the transactional nature of human social interaction is so persistent that even in situations involving physical actions, we often see these actions as though they were *trans*actions. Naive psychology is at the core of our understanding of the world.

At the heart of this proposal is a methodological inversion. Usually, artificial intelligence is thought of as the 'science of smart behaviour'—building systems which behave in a way that seems 'intelligent.' This naturally leads to all sorts of navel-oriented definitions and operational interpretations of the word 'intelligent,' none of which help to find intelligence. The reason they don't help is they miss the point: a definition of intelligence becomes part of science, but doesn't have any impact where it counts, which is on everyday human naive psychology.

So this is only part of the story. Artificial intelligence also needs to study naive psychology to find out what is going on in my head to make me see other people as having minds—and it is this that is an inversion of the conventional artificial intelligence approach. There should be two parts to the study of intelligence: the smart behaviour we're all familiar with, but also our ability to *recognise* that behaviour as smart. The sin of artificial intelligence is a sin of omission—it hasn't properly addressed the second part of the problem, that of naive psychology.

2 A brief history of naive psychology

Naive psychology isn't new to artificial intelligence, which has already tried a number of approaches to the problem. Perhaps the most successful have been the axiomatic formalisms (e.g. [9, 36].) These provide a representation of naive psychology as the ability to make inferences about a set of beliefs, desires, and intentions, corresponding to an agent's mental states.

These axiomatic formal approaches to naive psychology are usually based on some kind of modal logic, with knowledge and belief, desires and intentions as various different operators. These logics enable representation and reasoning about someone's mental states by making these states unobservable, so these 'opaque' agents can believe something which other people know to be false, for example.

But representational approaches to naive psychology also have their critics (e.g. [33, 35].) McDermott's [33] criticism is that using axiomatic formalisms and assuming deductive reasoning leaves out something—McDermott argues that the representations and the use of those representations cannot be truly separated, as they are separated in formalisms based on pure logic. But besides this methodological criticism, there are deeper ones—are mental states really best described in terms of beliefs, desires, and intentions? [16, 42]. This, again, is an assumption, and while it works well a lot of the time, there are some mental states which fit uneasily in this model (e.g. moods, hostility.) A lot of these are dispositional à la Ryle [41], who shows how a considerable part of typical human conversation is not in the indicative mood of propositional statements.

The most widely voiced criticisms of explicit representations in artificial intelligence (e.g. [2]) haven't really had much impact in this field, because they have to deal with something rare in physical environments, this opacity that people have. Both situated and connectionist approaches break down with the opaque nature of other agents. Although it is possible for such a system to have states which can be called 'beliefs,' 'desires,' and 'intentions,' they can have no direct perception of these states in other agents. While these approaches are apparently very good at dealing with observables, they are less good at dealing with the unobservable nature of other agent's mental states. This really does seem to require something which does the job of a theory—and this is what representations are good at. If representations had to become situated to deal with the physical world, situated approaches have to become representational to deal with the psychological one.

The conflict between logical and psychological approaches to artificial intelligence is not a new one [28, 29], but there is little to add to this controversy beyond what has already been written (e.g. [28, 33, 35].) The logicist project as it stands is unlikely to succeed in producing a causal account—but for a descriptive project [8] this doesn't matter.

So there has been significant work in this field, but perhaps artificial intelligence just hasn't realised the scale or importance of the problem. Naive psychology is *the* pressing problem. All the major philosophical stumbling blocks of artificial intelligence (e.g. consciousness, intentionality) can be traced to our inability to understand when and how to ascribe mental states to computers or other artifacts. This doesn't mean that naive psychology is logically prior to these problems, but that it is methodologically prior. If artificial intelligence is to tackle consciousness and subjectivity, it must study the complex of intuitions involved in our *recognition* of subjectivity—it is not enough to play operationalist games with it [34].

There is another source of evidence in naive psychology. In developmental psychology in particular, there has been a recent rapid growth in interest in how people—and some animals—acquire the ability to perceive and reason about one another's mental states (e.g. [38, 49].) This often builds on the beliefs-desires-intentions nexus used in representational naive psychology, but it questions it in other ways. An understanding of beliefs, for instance, doesn't seem to be innate, but is developed between the ages of two-and-a-half and four [38, 48]. These psychological studies are starting to probe the regularities and structures which underlie the development of human naive psychology.

In addition to the theoretical models common in representational approaches to naive psychology, there are a few more practical examples. In one, Davis [13] described a model by which one agent could infer the limits on another agent's knowledge from knowledge about its perceptual limits—this shows a close affinity with some of the 'false belief' work in developmental psychology (e.g. [1, 48].) And in another, Shultz [45] described a model for the ascription of intentions to other agents, using rules which relate behaviour and mental states. These models both show that it is possible to successfully construct working models for aspects of naive psychology.

This proposal is not one of "psychologising," making "the artificial imitate the natural" [28], although this might turn out to be necessary after more research. The point is that we—the people who are designing and evaluating these machines—are people with relatively uniform cultures, societies, and biologies—at least when compared to machines. Perhaps, as Searle [44] claims, these factors affect human mental phenomena. If so, it would be surprising if they didn't also affect our recognition and interpretation of those phenomena.

We can, of course, take McCarthy's stance: "this is artificial intelligence and so we don't care if it's psychologically real" [29]. But as soon as we talk about minds we are, whether we like it or not, involved in something psychological, so to compare minds and computers will inevitably be partly a psychological question. The actual nature of the distinction between human intelligence and artificial intelligence does matter.

3 Human and alien intelligence

Within artificial intelligence, there is often an assumption that there is something which can be called 'intelligence,' but which is very different from what we call 'intelligence' in people. We can call this the 'alien intelligence hypothesis.'

It is entirely possible that the alien intelligence hypothesis is just false. If the complex bag of phenomena we call 'intelligence' is something that people use to interact with each other in human societies, an alien intelligence which didn't interact in the same way in these societies might not be seen by us as intelligence. Perhaps systems are just seen as intelligent in proportion to how well we can understand their patterns of behaviour.

There are three possibilities for artificial intelligence. First, it might be possible to build a system with 'alien intelligence' which doesn't require our naive psychology to see it as intelligent. Second, it is possible that 'intelligence' implicitly means 'human intelligence,' and appeals to our human naive psychology to recognise it as such. A third possibility is that intelligent systems can never be built (for either logical or pragmatic reasons) in which case naive psychology isn't relevant. Dealing fully with this third possibility is outside the scope of this paper.

A simple reply to the first option would be that supposing alien intelligence did exist, how would we recognise it as such *without* appealing to our human recognition of intelligence? The idea that alien intelligence exists but isn't recognisable is unfalsifiable. We could call this the 'Vegetarian's Dilemma,' lettuce might be conscious and suffering immensely when we eat it, but if it is incapable of communicating that fact to us, should we ignore the problem? On this principle, computers (along with lettuce, beer cans, and rocks) could already be intelligent, we just aren't able to recognise them as such. And if intelligence is recognisable, it must be recognisable by us humans, and then it probably doesn't count as alien any more.

This is because human intelligence isn't restricted to any particular narrow set of behaviours. That which we recognise as intelligence can change—indeed it is pretty certain that computers are introducing such a change. One of Turing's [46] points was that much hangs on "the use of words and general educated opinion." Artificial intelligence isn't just a case of making machines smarter, but also of making us better at seeing their smartness.

Perhaps the case for alien intelligence is poorer than it seemed at first. The onus is on those who claim systems to have it to demonstrate (not define) its independence from human intelligence. If there is no such thing as "intelligence in general" [20]; if the principles of thinking for artificial intelligence are the principles of thinking for human intelligence, we are reduced to the second possibility, and should take the term 'intelligence' to mean culturally-oriented human intelligence. After all, this is what the Turing test really tests for [20], which is why its results are sometimes so strange [10, 11, 26, 47]. The Turing test is, in many ways, a paradigm example of the effects of naive psychology [5], in that it is, in effect, a test of the ability of one system to ascribe mental states to another. But most of the interesting stuff in the Turing test is going on in the head of the observer, not only in the system under test [11]. It is not, as Searle [44] takes it to be, "third person or 'objective.' " We might even suggest an inverted Turing test: putting a system in the role of the observer and looking at its ability to distinguish between humans and programs could be a useful way of evaluating its naive psychology.

The problem is this persistent anthropocentricity—we can't step outside our humanity although we perpetually see things as if they are independent of us, and while for physics that doesn't usually matter, for psychological concepts such as 'intelligence,' we must remember that we are human. We need to discover what it is to be human before we can truly know where the differences between people and machines are.

4 Models for naive psychology

In looking at what is going on our heads when we see people as minds rather than as bodies, some of the most useful tools are models of the process of ascribing mental states to other systems. In this section, three candidate models will be examined in a little more detail, anthropomorphism, the simulation model, and the theory model.

Model one: anthropomorphism

One way of ascribing mental states to a system is just to anthropomorphise it—to ascribe it human mental characteristics without reference to their real competences, but anthropomorphism is a complex and subtle phenomenon [5, 18] and not one that has been studied much. Eddy et al. [18] looked at people's tendency to anthropomorphise animals, and suggest that there are two primary mechanisms involved: "people are likely to attribute similar experiences and cognitive abilities to other animals based on (1) the degree of physical similarity between themselves and the species in question (e.g. primates,) and (2) the degree to which they have formed an attachment bond with a particular animal (e.g. dogs and cats)" [18].

Computers don't score too well on physical similarity, so this is likely to form a persistent bias against people ascribing mental states to them, unless we build them with a physical resemblance to us (e.g. [3].) Familiarity, fortunately, offers us a way out of this trap—we can in principle learn to see computers as minds.

Anthropomorphism can be interpreted several ways. Caporael [5] suggests that it is a "'default schema' applied to non-social objects, one that is abandoned or modified in the face of contradictory evidence," but the evidence is against either animals or computers really being 'non-social' [21, 37] and familiarity can increase rather than decrease the tendency to anthropomorphise [18]. Alternatively, perhaps our tendency to anthropomorphise is really a disposition to take the "intentional stance" [14], to see others as minds rather than as bodies. If, instead of taking the intentional stance, the physical stance is taken, the very different faculty of naive physics will be deployed. Anthropomorphism, then, is a part of naive psychology in that it determines whether or not an intentional stance will be taken, but it is not truly part of the stance itself. It plays the role of the rationality assumption in Dennett's model—although clearly anthropomorphism isn't the same thing as rationality—and the suggestion that the rationality assumption is "pre-theoretic" [16] does allow us to interpret it as non-philosophical.

Model two: simulation

Sometimes prediction of other people's mental states is better modelled by 'simulating' the other person, by pretending to *be* them, and to look at the world from their point of view. Clark [8] suggests that a similar simulation process could even account for naive physics—perhaps Hayes' [23] paper on liquids could be recast as a kind of simulation, and as far as the predictions are concerned, viewed externally, there needn't be any difference. For naive psychology, there is evidence that for some predictions—particularly those involving affective states—an ability to simulate other people works well [25, 38]. Representational artificial intelligence does simulation all the time—it's just another kind of hypothetical reasoning. Simulation, or taking another person's role, is a way that we can understand some aspects of another's mental states; for instance, to recognise somebody's ignorance [13].

So simulation is another way that we can reason about another's mental states. It is a way that works rather better for affective than for cognitive states [25] but it doesn't deal with everything: there are some tasks which children actually answer differently, but which they ought to answer the same if they use simulation to get the answer [39]. Something is left over, and that something is a 'theory' of mind—not a theory in the scientific sense [7, 44]—simply a theory in the sense of a set of tools for thinking about the unobservables of another person's mental states.

Model three: theory

This theory aspect of prediction is that aspect which is most similar to the representational artificial intelligence. Some (e.g. [19]) even take it as the complete answer to naive psychology, but this stretches it too far; a strong representational theory of mind is subject to too many philosophical and evolutionary objections [16], and fails to account for all phenomena [25, 39]. But just because a representational theory of mind can't provide a complete naive psychology doesn't mean that it doesn't form part of a complete naive psychology. The theory theory, as it is currently interpreted in psychology, describes naive psychology as a set of rules for dealing with the unobservable mental states of others. Its best analogue in artificial intelligence, therefore, would be a body of laws and heuristics for guessing at one other's mental states.

Young children, for example, can't properly ascribe beliefs they know to be false to someone. Their ascription can be described by a rule like:

$$believes(self, X) \rightarrow believes(self, believes(agent, X))$$

In time, the patterns of ascription change, so that the child becomes aware of the other person's perceptions. So now the ascription can be described:

$$\begin{aligned} believes(self, perceived(agent, X)) \wedge believes(self, \neg believes(agent, \neg X)) \\ \rightarrow believes(self, believes(agent, X)) \end{aligned}$$

These rules are steps towards a simple descriptive account of how an agent might come to pass the false belief test [1]. Note that both of these rules are intentional, that is, they only apply when taking the intentional stance to $agent$.

Whether or not this account is correct or not isn't really the issue. These rules, which a close affinity to those of Davis [13], are a first step to modelling Wellman's [48] account of belief progressing from a copy theory to a representational theory, but there are many other possible models (e.g. [1, 38].) What is actually needed is a way of comparing and combining the different models.

In artificial intelligence, all the best programs for playing games like chess (and games are often a good metaphor for human social interaction) use a subtle mixture of simulation (look ahead) and theory (heuristics.) because neither on their own is ever sufficient. In principle, of course, a heuristic theory can generate a simulation [12], and in practice an actual system—such as a trained connectionist network—might show aspects of theory and simulation under different circumstances, just as electrons can behave like particles or like waves. These are valid points, but are more relevant to a causal account than to a descriptive one.

These three models—anthropomorphism, simulation, and theory— represent different aspects of naive psychology rather than the whole, but they can be combined to create a complex composite model. When trying to predict or reason about the behaviour of a system, a complex of dispositions, one of which is anthropomorphism, selects a stance with respect to that system. These stances deploy natural faculties—so when dealing with a physical system, naive physics is applied, but for a psychological system, naive psychology is applied. Often, both stances could in principle be taken to the same system (even a thermostat, [31],) and although in practice there seems to be a mutual exclusion between the different stances, this is where individual differences in the dispositions and the social context can influence how different people see the same system.

A mind will only be seen in the system from the intentional stance—that selected by anthropomorphism—and within the intentional stance as a whole there are different sub-stances which depend on the access to the other's mental states that is required. If we are to 'simulate' it—to see what it is like to *be* the system—that can only happen if the system is believed to have the right kind of mechanism. The theory stance, on the other hand, is better at dealing with external, behavioural, questions.

How well do these models do? Although they barely hint at the true complexity of naive psychology, they do have some predictive power—this can be shown with Searle's [43] "Chinese Room." This is not the place for a new rebuttal or any analysis of Searle's argument, but the intuitive part of Searle's experiment is attractive because he appeals to plausibility and insists on us taking the first person stance when we look at the scenario. This is a big hint that he is appealing to our naive psychology as we try to understand his model.

First, Searle claims the system can be assumed to have passed the Turing test, so as far as external appearances are concerned, it has already been ascribed mental states. It should be difficult to anthropomorphise a room on the similarity factor, but maybe it can be done. But then Searle adds something—he tells us what is going on inside the room. In fact, he tells us that inside there is an agent, Searle-in-the-room, manipulating slips and scripts according to a rule book written in English. Now the story is different: we can anthropomorphise either the room, Searle-in-the-room or the rules and scripts, but Searle-in-the-room (being human) is by far the most similar to us, so he acts like a magnet to our naive psychology and we immediately take his stance—which is forbidden by Searle's rules. We are then trapped, as Hofstadter was, asking "which level does Searle wish us to *identify* with?" [26, added emphasis]. And variations on the theme change the identification: the story changes if we put the room in the head of a robot, partly because robots are a lot easier to anthropomorphise than rooms.

This point mustn't be pushed too far—enough has been written on this topic already—but it shows how sensitive thought experiments are to naive psychology, and that even a fairly simple model can begin to make rough predictions about naive psychology—even a philosopher's naive psychology. We might learn a lot developing a model for naive psychology which would match Searle's intuitions, and, accommodating individual differences, those of his critics—in fact, this is work in progress. The idea of a computational model of a philosopher is perhaps rather bewildering, but if we borrow from McCarthy the idea that a part of philosophy is "the science of common sense" [31] perhaps a computational version of this part isn't so strange after all.

Another example of the effects of naive psychology is Woolgar's [50] description of a device which bolts on to a video recorder and splices out advertisements during recording. On one level this is clearly intelligent behaviour, but if you then read the instructions, and they tell you that it actually works by detecting a particular signal in the transmission, this changes the ascription of intelligence, and "redefines and thus reserves the attribute of 'intelligence' for some future assessment of performance" [50]. Again, the change in our knowledge affects the stance that we take—affects whether or not we see the system from the intentional stance.

This integrated model shows a sensitivity to physical form and our knowledge of the system's design which is perhaps rather distressing for strong artificial intelligence. It seems to show not that it is impossible in principle, but that it is just very hard for people to see things which don't physically, structurally, and behaviourally resemble people as being intelligent. Perhaps Brooks and Stein [3] are right to build Cog with a humanoid form, not for any technical reason, but simply because it will make it easier for us to see Cog as an intelligent system.

5 Naive psychology for machines

In the "Second naive physics manifesto," Hayes [24] concluded with a discussion on the importance of common sense for artificial intelligence. The reasons for this proposal are almost completely orthogonal to the reasons for his, so the justifications are different.

Of all the naive disciplines proposed by Hayes and others, naive psychology is the only one that is obviously specifically human, but in all this work there is an implicit anthropocentricity. Right back to McCarthy's [30] proposal of common sense, it was assumed that the common sense to be used is human common sense.

It is entirely possible that intelligent behaviour is distinguished not by an objective criterion of success, rationality, adaptiveness, or what have you, but by a subjective criterion of compatibility with our human naive psychology. If this is true, there are two ways to build intelligent systems. The first is to build such systems, and then to look at them to see whether we think they are intelligent or not. A second approach is to first look at how this naive psychology construes more familiar systems. The problems of the first strategy are simply that, without any principled understanding of how we construe the behaviour of such systems, we stumble across all sorts of biases. These biases can make us completely misinterpret the behaviour of systems, so an understanding of the biases is required before we can be certain about our understanding of the behaviour of systems.

There is no strong methodological component to this proposal, because the project is just too important to be dismissed as a project merely on methodological grounds—and the same goes for naive physics. Dreyfus and Dreyfus claim, for example, that "the problem of finding a *theory* of common sense physics is insoluble because the domain has no theoretical structure" [17, original emphasis]. Well, this depends on what you want from the theory. Even if, as Dreyfus and Dreyfus claim, naive physics can't be described fully by reference to "abstract laws," that doesn't mean that we should just give up. In the real world, theories aren't just right or wrong, but provide a greater or lesser measure of predictive competence—and even a partially correct theory is better than none.

This proposal is, like Hayes', a descriptive one: the construction of broad models of naive psychology, along the lines of the aspects sketched here—the attempts to build models of small parts of naive psychology have not been so successful as to indicate that they are necessarily on the right track. At least to start with, a broad and shallow approach is needed to sketch out naive psychology; it is not yet anywhere near as clearly structured into topics as Hayes presents naive physics. Pushing hard on one topic, like an air bubble under the wallpaper, might just move the problems somewhere else. We need to look at the whole problem and at the relationships between the different topics.

But this proposal is methodologically different from Hayes' in that working models of naive psychology are to be constructed. There are several reasons for this: first, simply because the psychologists and philosophers have asked for one and lament artificial intelligence having "lost much of its ambition in this area" [42]. Secondly, in the absence of a working model, theories become too hard to evaluate, especially for naive psychology which depends on comparison with people.

A third and more fundamental reason is that for the higher orders of mental states, a compatibility between psychologies is essential. A statement like 'I believe that she believes that he believes that…' depends on my recognising someone else's beliefs as like my beliefs. If we are to be able to make statements like this about machines—and the Turing test surely permits this—then there must be a *compatibility* between human and machine competence at recognising beliefs in others. Evaluating this compatibility requires working models; paper theories just can't support the interaction required.

6 Conclusions

The problems that artificial intelligence is tackling are big ones—big enough to make some think that there are fundamental and possibly irretrievable flaws either in the discipline or even in the whole of science. This is an over-reaction; certainly our anthropocentricity is a big problem, but not one that is inaccessible in principle to science.

There is little evidence that, as Dreyfus and Dreyfus [17] claim, the problem of common sense cannot be solved in principle, although the apparent lack of progress in this area is disconcerting. While work is progressing, it is hampered by some rather big methodological problems.

At the core there was a simple problem: we forgot about anthropocentricity and took too much of what we intuitively felt to be right as being the truth. Stepping outside our humanity is something that perhaps we can never do in principle, but that doesn't mean that we shouldn't try—not by a regress to the Skinnerian vantage point (with apologies to Dennett [16]) denying human mentalistic terms completely, but by indirectly looking at the effects of the ultimate unobservable, our anthropocentric point of view.

Although there is no justification for a *definition* of intelligence as the behaviour that we recognise as intelligent—it may be that intelligence is identified principally by our ability to recognise it. Unfortunately, we don't really know—even informally—what intelligence is.

To find out, we need to look at how people (and perhaps animals, too) ascribe mental states to other people, to animals, to machines, and perhaps even to thermostats, all under a variety of different conditions. Artificial intelligence has a role to play here, too; we can use it to build models of this ascription process—models which can be compared to those of people [42]. This proposal sketches out how this might be started, but there is a lot more to be done.

At the end of the day, we can all recognise intelligent behaviour when we see it. When we see people, we see them as minds, not just as bodies. When we see computers, we don't see minds. The difference between people and computers lies in ourselves as well as in them, and if we are to overcome this fundamental anthropocentric asymmetry, artificial intelligence must join up with psychology at least the extent of finding when and how we see minds. It must begin to study naive psychology.

References

[1] Simon Baron-Cohen, Alan M. Leslie, and Uta Frith. Does the autistic child have a 'theory of mind'? *Cognition*, 21(1):37–46, 1985.

[2] Rodney A. Brooks. Elephants don't play chess. In Pattie Maes, editor, *Designing Autonomous Agents: Theory and Practice from Biology to Engineering and Back*. MIT Press, Cambridge, Massachusetts, 1991.

[3] Rodney A. Brooks and Lynn Andrea Stein. Building brains for bodies. AI Memo 1439, MIT AI Laboratory, August 1993.

[4] Richard W. Byrne and Andrew Whiten, editors. *Machiavellian Intelligence: Social Expertise and the Evolution of Intellect in Monkeys, Apes, and Humans*. Oxford University Press, Oxford, 1988.

[5] Linnda R. Caporael. Anthropomorphism and mechanomorphism: Two faces of the human machine. *Computers in Human Behavior*, 2(3):215–234, 1986.

[6] Susan Carey. *Conceptual Change in Childhood*. MIT Press, Cambridge, Massachusetts, 1985.

[7] Andy Clark. From folk psychology to naive psychology. *Cognitive Science*, 11:139–154, 1987.

[8] Andy Clark. Two kinds of cognitive science? In *Proceedings of the European Conference on Artificial Intelligence, ECAI*, pages 155–157, 1988.

[9] Philip R. Cohen and Hector J. Levesque. Intention is choice with commitment. *Artificial Intelligence*, 42:213–261, 1990.

[10] Kenneth M. Colby. Modeling a paranoid mind. *Behavioural and Brain Sciences*, 4:515–560, 1981.

[11] Harry M. Collins. *Artificial Experts: Social Knowledge and Intelligent Machines*. MIT Press, Cambridge, Massachusetts, 1990.

[12] Martin Davies. The mental simulation debate. In Christopher Peacocke, editor, *Objectivity, Simulation and the Unity of Consciousness*, pages 99–128. Oxford University Press for the British Academy, 1994.

[13] Ernest Davis. Inferring ignorance from the locality of visual perception. In *Proceedings of the Seventh National Conference on Artificial Intelligence, AAAI*, pages 786–790, 1988.

[14] Daniel C. Dennett. Intentional systems. *Journal of Philosophy*, 68:87–106, 1971.

[15] Daniel C. Dennett. Cognitive wheels: The frame problem of AI. In Christopher Hookway, editor, *Minds, machines, and evolution*, pages 129–151. Cambridge University Press, 1984.

[16] Daniel C. Dennett. *The Intentional Stance*. MIT Press, Cambridge, Massachusetts, 1987.

[17] Herbert L. Dreyfus and Stuart E. Dreyfus. Making a mind versus modelling the brain: Artificial intelligence back at a branchpoint. *Daedalus*, 117(1):185–197, 1988.

[18] Timothy J. Eddy, Gordon G. Gallup, and Daniel J. Povinelli. Attribution of cognitive states to animals: Anthropomorphism in comparitive perspective. *Journal of Social Issues*, 49(1):87–101, 1993.

[19] Jerry A. Fodor. Fodor's guide to mental representation: The intelligent auntie's vade-mecum. *Mind*, 94:55–97, 1985.

[20] Robert M. French. Subcognition and the limits of the Turing test. *Mind*, 99:53–65, 1990.

[21] Donna Haraway. Otherworldly conversations; terran topics; local terms. *Science as Culture*, 3(14):64–98, 1992.

[22] John Haugeland. The nature and plausibility of cognitivism. *Behavioural and Brain Sciences*, 1:215–226, 1978.

[23] Patrick J. Hayes. Naive physics I: Ontology for liquids. In Jerry R. Hobbs and Robert C. Moore, editors, *Formal Theories of the Commonsense World*, pages 71–108. Ablex, Norwood, New Jersey, 1985.

[24] Patrick J. Hayes. The second naive physics manifesto. In Jerry R. Hobbs and Robert C. Moore, editors, *Formal Theories of the Commonsense World*, pages 1–36. Ablex, Norwood, New Jersey, 1985.

[25] Peter Hobson. Understanding persons: The role of affect. In Simon Baron-Cohen, Helen Tager-Flusberg, and Donald J. Cohen, editors, *Understanding Other Minds: Perspectives From Autism*, pages 204–227. Oxford University Press, Oxford, 1993.

[26] Douglas R. Hofstadter and Daniel C. Dennett. *The Mind's I: Fantasies and Reflections on Self and Soul*. Basic Books, New York, 1981.

[27] Nicholas K. Humphrey. *Consciousness Regained*. Oxford University Press, 1984.

[28] David Israel. A short companion to the naive physics manifesto. In Jerry R. Hobbs and Robert C. Moore, editors, *Formal Theories of the Commonsense World*, pages 427–447. Ablex, Norwood, New Jersey, 1985.

[29] Gina Kolata. How can computers get common sense? *Science*, 217:1237–1238, 1982.

[30] John McCarthy. Programs with common sense. In *Proceedings of the Teddington Conference on the Mechanisation of Thought Processes*, pages 75–91, London, 1959. HMSO.

[31] John McCarthy. Ascribing mental qualities to machines. In Martin Ringle, editor, *Philosophical Perspectives in Artificial Intelligence*, pages 161–195. Harvester Press, Brighton, 1979.

[32] John McCarthy and Patrick J. Hayes. Some philosophical problems from the standpoint of artificial intelligence. In B. Meltzer and Donald Michie, editors, *Machine Intelligence 4*, pages 463–502. Edinburgh University Press, Edinburgh, 1969.

[33] Drew McDermott. A critique of pure reason. *Computational Intelligence*, 3:151–160, 1987.

[34] Donald Michie. Turing's test and conscious thought. *Artificial Intelligence*, 60(1), 1993.

[35] Marvin Minsky. *The Society of Mind*. Simon and Schuster, New York, 1985.

[36] Robert C. Moore. A formal theory of knowledge and action. In Jerry R. Hobbs and Robert C. Moore, editors, *Formal Theories of the Commonsense World*, pages 319–358. Ablex, Norwood, New Jersey, 1985.

[37] Clifford Nass, Jonathan Steuer, and Ellen R. Tauber. Computers are social actors. In *Proceedings of CHI'94*, pages 72–78. ACM Press, 1994.

[38] Josef Perner. *Understanding the Representational Mind*. MIT Press, Cambridge, Massachusetts, 1991.

[39] Josef Perner. The necessity and impossibility of simulation. In Christopher Peacocke, editor, *Objectivity, Simulation and the Unity of Consciousness*, pages 145–154. Oxford University Press for the British Academy, 1994.

[40] David Premack and Guy Woodruff. Does the chimpanzee have a 'theory of mind'? *Behavioural and Brain Sciences*, 4:515–526, 1978.

[41] Gilbert Ryle. *The Concept of Mind*. Hutchinson, 1949.

[42] Jerry Samet. Autism and theory of mind: Some philosophical perspectives. In Simon Baron-Cohen, Helen Tager-Flusberg, and Donald J. Cohen, editors, *Understanding Other Minds: Perspectives from Autism*, pages 427–449. Oxford University Press, Oxford, 1993.

[43] John R. Searle. Minds, brains, and programs. *Behavioural and Brain Sciences*, 3:417–424, 1980.

[44] John R. Searle. *The Rediscovery of the Mind*. MIT Press, Cambridge, Massachusetts, 1992.

[45] Thomas R. Shultz. From agency to intention: A rule-based computational approach. In Andrew Whiten, editor, *Natural Theories of Mind: Evolution, Development and Simulation of Everyday Mindreading*, pages 79–95. Basil Blackwell, Oxford, 1991.

[46] Alan M. Turing. Computing machinery and intelligence. *Mind*, LIX(2236):433–460, 1950.

[47] Joseph Weizenbaum. *Computer Power and Human Reason*. W. H. Freeman, 1976.

[48] Henry M. Wellman. *The Child's Theory of Mind*. MIT Press, Cambridge, Massachusetts, 1990.

[49] Andrew Whiten, editor. *Natural Theories of Mind: Evolution, Development and Simulation of Everyday Mindreading*. Basil Blackwell, Oxford, 1991.

[50] Steve Woolgar. Why not a sociology of machines? the case of sociology and artificial intelligence. *Sociology*, 19:557–572, 1985.