J. EXPT. THEOR. ARTIF. INTELL. 3(1991)163–170

Solipsism, individualism and cognitive science[1]

SAUL TRAIGER

Cognitive Science Program, Occidental College, Los Angeles, California 90041, USA
INTERNET: traiger@oxy.edu

Abstract. 'Artificial Intelligence cannot ignore philosophy' (McCarthy 1988) I shall challenge the claim that good old-fashioned artificial intelligence, or GOFAI (Haugeland 1985) is solipsistic while more recent neural or 'brain-style' approaches to AI are not (Rumelhart *et al.* 1986). After distinguishing GOFAI from connectionism, I will first show that GOFAI is not committed to solipsism but rather to what is more properly called individualism. I argue that the feature of GOFAI which entails individualism is shared by connectionism. Individualism is a metaphysical assumption of both types of AI, one which may indeed be pernicious. It is an assumption which must be located and understood.

Keywords: solipsism, philosophy, connectionism

Received: 25 July 1990

1. Varieties of artificial intelligence

Construed most broadly, artificial intelligence attempts to build intelligent systems using computational tools. AI begins with the specification of an architecture. Implementation involves a functional account of some aspect of cognition and the specification of that account in computations supported by the architecture. Two varieties of AI are often contrasted. Good old-fashioned artificial intelligence employs computations which can be carried out on von Neumann machines. Connectionism makes use of an architecture of abstract neurons which are connected in networks. The connections among neurons or units communicate values which effect the activation of units. The behaviour of a unit is a function which takes input from other units and sends output to the units to which it is connected. (Rumelhart 1989)

It is important to distinguish these minimal accounts of types of AI, in terms of architecture, from philosophical commentaries on them. The philosophy of mind most closely associated with GOFAI is functionalism. AI programs are theories of the functional organization of cognition. According to proponents of GOFAI, minds are literally computers, albeit computers we don't yet fully understand. The philosophical commitments of connectionism are less well established, perhaps because connectionism itself has only recently gained

Presented at the SUNY Binghamton AI Workshop 'Artificial Intelligence: an emerging science or a dying art form?'

substantial support from the AI community. Most connectionists see themselves as materialists, but not as functionalists, when functionalism is construed as the view that cognition can be understood in abstraction from the hardware in which it is realized (Churchland 1986). In this paper I shall be concerned primarily with philosophical claims for the two varieties of AI.

2. Solipsism

It is as rare to encounter an avowed solipsist as it is to locate philosophical works which endorse solipsism.[2] Instead, it is common to find the remark that a position is compatible with solipsism or that a position entails solipsism. The charge of solipsism is always an objection to a theory. Descartes, in his *Meditations on First Philosophy*, realizes that his own view before the Third Meditation is solipsistic (Descartes, Cottingham *et al.* trans. 1984). The fact that my mental states may not accurately mirror an external world is compatible with the extreme possibility that there is nothing outside my mental states for me to represent. This is not Descartes' final position. He subsequently argues that the character of one special mental item, that which represents an all-perfect god, forces the anti-solipsistic conclusion that the mind which has that idea is not alone.[3]

Understood as the ontological position that there is only one thing in the world, namely oneself, solipsism is not held by GOFAI. Traditional AI attempts to explain the mind in terms of mental representations understood as rule-governed systems of active physical symbols. There is an ontological motivation for this view, to retain what is important about the mental, particularly intentionality and mental representation, without invoking non-physical entities. But GOFAI's ontological strictures don't call into question the existence of things which are outside the cognitive agent. On the contrary, GOFAI accepts the existence of the causal order in which minds operate.

The ontological version of solipsism traditionally gets its punch from the putative primacy of introspective awareness. Philosophical behaviourism, by denying that there are mental states to introspect, is the opposing view. Functionalism occupies the middle ground: the mental isn't eliminated by reduction to public behavioural phenomena, though it is in some sense reduced to it. Thus traditional AI rejects the solipsist's introspective route to the mental (Lycan 1987).

There is a related sense in which GOFAI may appear to be solipsistic, and this might be called epistemological solipsism. This is the view that the only things I can know are states of myself, though there may very well be things outside of me. Traditional AI is not solipsistic in this sense for two reasons. First, as we've just seen, it doesn't accept introspection as the fundamental source of knowledge, and second, it refuses to participate in the epistemological enterprise which leads to epistemological solipsism.

Traditional epistemology involves justifying knowledge claims by appeal to a foundation of self-justifying propositions. Epistemological solipsism results when one denies that there are good inferences from a foundation of direct self-knowledge to the existence of external objects (Chisholm 1982). Artificial intelligence is concerned with knowledge, but not primarily in terms of the credentials of particular knowledge claims. Rather, AI investigates the way systems of beliefs are organized and stored in memory, and the way such information is accessed. The knowledge representation problem demonstrates that AI takes the existence of external objects and our knowledge of them for

granted (Minsky 1975). Philosophical GOFAI endorses the naturalistic approach to epistemology (Goldman 1986).

There is a third variety of solipsism; Jerry Fodor calls it 'methodological solipsism' (Fodor 1981). This is the view that for the purpose of constructing a model of human intelligence, we can act as if (ontological) solipsism is true. That is, what is relevant to understanding the mind and mental representation is not the world in which the cognitive agent represents, but the internal mental realm in which the representations are constructed. One can investigate mental states by looking at those internal states. GOFAI is solipsistic in this sense.

Fodor's description of Winograd's classic AI program, SHRDLU, nicely illustrates methodological solipsism. The program manipulates blocks, and answers questions about blocks in a constrained 'block world'. But there really are no blocks at all. The programmer presents the data to the machine in such a way that the machine has the formal representations it would have if it were a robot actually manipulating blocks. Fodor (1981) says: 'In effect, the device is in precisely the situation that Descartes dreads; it's a mere computer which dreams that it's a robot.'

A similar point has also been made by Putnam, who argues that the Turing test does not provide a test for reference (Putnam 1981). Putnam maintains that even if a machine passed the test, we would still deny that its linguistic output achieves reference. Here's the reason: suppose two machines play the imitation game with each other. We could imagine them continuing to play the imitation game together happily, even if the rest of the world disappeared! The machines, like ants in the sand who miraculously draw a picture which bears a striking resemblance to Winston Churchill, fail to refer. Their sentences are not 'about' anything. The machines which Putnam imagines passing the Turing test lack what he calls 'language entry' and 'language exist' moves. The moves are all internal; they don't depend on the external world at all.

It may appear that methodological solipsism results from GOFAI's representationalism and its commitment to what Newell has called 'the knowledge level' (Newell *et al.* 1989). The argument is this: if one is out to account for how mental representations and their logical transformations bring about high level cognitive behaviour, then one need not pay attention to the relation of those representations to the outside world. So there could be a gap between the representations and the way the world is. This gap provides a foothold for solipsism, one which AI creates when it concerns itself with such processes as problem solving and expertise, while ignoring the causal mechanisms which, in humans, presumably give rise to cognitive states.

3. Solipsism and individualism

Consideration of the varieties of solipsism suggests that there is a sense in which GOFAI is solipsistic. It is methodologically solipsistic, not ontologically or epistemologically solipsistic. I will now characterize methodological solipsism more fully, and identify it with individualism. It will turn out that despite the appearance reported above, GOFAI's concern with the knowledge level has nothing to do with its commitment to methodological solipsism.

Individualistic theories of the mind are those which share a strategy for individuating mental states. Such theories hold that mental states can be individuated or cared up by looking at the goings-on in the individual to whom

mental states have been attributed. Individualism claims that we can distinguish mental states from one another while ignoring matters outside the skin. What happens outside the individual is irrelevant to the determination of that individual's mental state.

The mind-brain (type) identity theory, proposed in the 1950s by J.J.C. Smart and others is an example of an individualistic view (Smart 1959). It says that mental states are states of the brain. So it individuates mental states by what goes on literally 'in the head' of a person. The neurological activity in the head *fixes* the mental state. Even if everything outside the person were different, the mental state is the same as long as the brain state remains the same. Mental states *supervene* on physical states. That is, if we fix the physical state of an individual, we've fixed the person's mental state. Supervenience requires only that the mental be determined by the physical; it allows different physical states to determine the same mental states. The supervenience thesis, then, is weaker than the identity theory, since the identity theory imposes strict identity, i.e. that mental states determine physical states as well.

Functionalism, in contrast to the mind-brain identity theory, admits that a cognitive state can be realized by more than one physical state. That is, there could be two different physical state types which can realize the same mental state type. There's a range of physical state configurations which can instantiate any particular functional architecture, such as the von Neumann architecture. Functionalism, then, explains the relevance of the physical to the mental without demanding a full reduction of the mental to the physical.

The multiple realizability or medium-independence of formal systems is a familiar and deep point (Haugeland 1985). It shouldn't blind us, however, to the very important way in which cognition is tied to the physical system which implements it. That's simply this: for the functionalist, like the type identity theorist, once the physical states of a system are fixed, the mental states of that system, if any, are fixed as well. Mental states supervene on physical states.

GOFAI is individualistic because it is committed to supervenience, not because it tends to focus on mental representation or on high-level cognitive functions such as problem-solving, linguistic competence, and reasoning. There could be theories of these features of cognition which are not individualistic. Any theory which is committed to supervenience about the mental, however, will be committed to individualism.

4. Problems with individualism

Traditional artificial intelligence is individualistic or methodologically solipsistic, because it embraces supervenience. There are powerful arguments against individualism, however, and I will now briefly formulate two.[4]

Tyler Burge gives the following example which calls supervenience into question (Burge 1979). He asks us to imagine two possible worlds, one just like the actual world, containing a person (let's call him Ignat_1) and a 'twin' world which is similar in almost every respect to the actual world. The twin world contains Ignat_2, who, up to time t, is indistinguishable from Ignat_1. Both Ignats suffer from a painful condition in the thigh, and both go around believing that the pain is arthritis. Ignat_1 is, of course, wrong. In the actual world arthritis is a condition of the joint, and a belief that one has arthritis in the thigh is false. The only difference between the two worlds is that in the twin world, experts use the word

'arthritis' exactly the way people like $Ignat_1$ in the actual world misuse the word 'arthritis'. In the twin world when $Ignat_2$ believes that he has arthritis in his thigh, he has a true belief.

$Ignat_1$ and $Ignat_2$ are physically exactly alike until time *t*. No one corrects $Ignat_1$'s use of the word 'arthritis' in the actual world until *t*, and no one corrects $Ignat_2$'s usage in the twin world (since in the twin world $Ignat_2$'s usage is correct). Although $Ignat_1$ and $Ignat_2$ are physically identical, down to their micro-structure, before *t* they have different beliefs. It is clear that the two beliefs are different: $Ignat_1$ has a false belief while $Ignat_2$'s belief is true. This is a case where the mental does not supervene on the physical, where two individuals in identical physical states differ in their mental states.

Burge's counterexample strikes some as implausible. One common intuition is that the beliefs of the two Ignats are the same. The beliefs have different truth values due to the difference in the two world, outside the skins of the Ignats. On this defence of supervenience, the facts about the way the word 'arthritis' is used in a world can't have any effect on the content of Ignat's belief, (Fodor 1987).

Though my central purpose is not to defend Burge's argument, I want to support Burge's counterexample by putting some pressure on this last assertion. As I see it, the crucial question is whether facts about the way a term is used by a community can have an effect on the content of a belief (or other propositional attitude) when the believer (or holder of the propositonal attitude) is unaware of those facts.

Mildred believes that a Coast Live Oak tree (CLO) would look good in her yard. She desires that there be a CLO in her yard. In taking steps to fulfill that desire, Mildred travels to her local nursery, where the following conversation ensues:

> *Mildred*: 'I'd like to order a Coast Live Oak. Could you deliver it and install it?'
> *Nurseryperson*: 'Sure, we can have it planted by Thursday.'
> *Mildred*: 'I'll take it.'

The tree is promptly delivered and planted. Mildred is thrilled. Her pleasure, however, is short-lived. A friend visits, and when Mildred shows off her new tree, the following dialogue transpires:

> *Mildred*: 'How do you like my beautiful new CLO?'
> *Friend*: 'Mildred, you've been had. That's not a CLO! It's a California Scrub Oak.'
> *Mildred*: 'Are you sure? The nurseryperson promised to provide the tree I desired.'
> *Friend*: 'As you know, Mildred, I'm the nation's leading expert on native oaks.' [And he really is.]

Mildred is furious. Returning to the nursery she complains:

> *Mildred*: 'There's some mistake. You didn't plant a CLO in my yard.'
> *Nurseryperson*: 'I agree that we didn't plant a CLO in your yard. But I maintain that there has been no mistake.'
> *Mildred*: (perplexed) 'How could that be? I asked for a CLO.'
> *Nurseryman*: 'We knew when you ordered your tree that you couldn't tell the

> difference between a CLO and other oaks. So we knew that your desire for a CLO and your belief that a CLO would look well in your yard was really the desire for any old oak and your belief that a CLO would look good in your yard was really the belief that an oak tree would look nice in your yard. Since you couldn't distinguish a CLO from any other oak *in your own head*, it makes no sense to say that you really desired a CLO. So we satisfied *your* desire. We pride ourselves on knowing our customers and on giving them exactly what they have *in mind*.'

Should Mildred accept this argument? If you try to defend supervenience against Burge's counterexample on the grounds that the two Ignats have the same belief, then you must say that the nurseryperson's argument is a good one. But the fact that we are inclined to reject that argument suggests that we must reconsider the view that what matters is what is in the head of the two Ignats and Mildred. Beliefs and desires involve meanings which are expressed in a common language. This fact allows us to hook onto meanings of which we have an imperfect grasp. I maintain that Mildred is owed a CLO, that she had a specific belief and desire for a CLO, regardless of what was in her head.

How could the nurseryperson be so confident about the content of Mildred's belief? He needs a theory of mind, one which enables him to individuate beliefs. As hypothesized, the story has the odd feature that the nurseryperson can confidently individuate Mildred's beliefs; in fact, he thinks he is better at individuating Mildred's beliefs than is Mildred. The nurseryperson has an individualistic theory: he individuates beliefs by what's going on in Mildred's head. We've seen that the GOFAI view is individualistic, and so the nurseryperson could be a researcher in traditional AI. But he could be anyone committed to supervenience. As long as Mildred's beliefs are fixed by her internal physical state, it doesn't matter whether one is an identity theorist, a functionalist, or as I shall argue, a connectionist. Each is committed to methodological solipsism, i.e. individualism.

A methodological solipsist might try to accommodate these examples by arguing that beliefs are still in the head, and that AI can explain how the cognitive engine can deal with the wider social context on which the examples depend.[5] In the case of Mildred, the methodological solipsist would agree that the nurseryperson does not correctly account for Mildred's belief. The belief is something like 'I want a CLO and a CLO is whatever the experts say is a CLO'. Thus the trees picked out by the belief are only CLOs. The idea is that Mildred's cognitive system contains, at least implicitly, the machinery to produce the right belief, and all of that machinery is internal. On this view, one could grant that the belief involves an appeal to experts, yet hold that we can pull that appeal 'inside'.

If this is right, then we should be able to get the fully explicit belief by looking only at what's inside Mildred. Now suppose that all oak expertise in the world is lost before Mildred's friend appears on the scene. Mildred is never corrected. Could we determine the content of Mildred's belief based on what is left, namely, just Mildred and the non-experts? I think not. We would not be able to individuate her belief or her desire. So the *individuation* of the belief depends on the existence of a community of experts who provide the content of the belief.[6]

The point can be made another way. Does Mildred's knowledge representation system contain the resources to deal with the CLO situation? In some sense it does, but in another it doesn't. It doesn't because the system can't determine

what counts as a CLO. It does in the sense that the system knows when to look for outside information to get answers it can't provide itself.

5. Anti-individualism and connectionism

The connectionist approach to AI differs from GOFAI in important respects. Connectionist architectures enable AI researchers to investigate the subsymbolic aspects of cognition, the existence of which even the die-hard cognitivist accepts (Pylyshyn 1984). While GOFAI first isolates functional structures such as beliefs and desires, and then explains them in computational terms, many connectionists have devoted their energies to the study of more basic cognitive activities, such as pattern recognition and vision, and they've argued that these phenomena are best studied at the subsymbolic level. Recent work on perceptual processing of information from the environment of the cognitive agent suggests that the external environment is important, and that an account of cognition which includes it may overcome some of the problems of the solipsistic alternatives.

I argued above that GOFAI's endorsement of individualism has nothing to do with its focus on 'higher' cognitive phenomena, but only with the endorsement of supervenience. But connectionists are committed to supervenience. They hold that cognitive states of an individual are determined by internal physical states, described in terms of a proposed architecture. As different as this architecture is from that of classical AI, its proponents believe, with the functionalists, that the physical states which instantiate the architecture determine mental states. So connectionists are committed to individualism.

It might help to attempt to place connectionist approaches on the larger philosophical map. One possibility is that connectionists are identity theorists: mental states are just brain states. But if full blown reductionism seemed overly parochial to functionalists, it should to connectionists as well, since their work demonstrates that intelligent systems can be built from a variety of hardware (though there will certainly be constraints imposed by the architecture). Another is that connectionists are simply functionalists with a new architecture. In any case, both alternatives are committed to supervenience, and are thus solipsistic in the methodological sense.

There is another possibility. Among philosophers who advocate connectionist AI, some are eliminative materialists. Like identity theorists, they identify the mind with the brain. But unlike identity theorists, they believe that the conceptual scheme of folk psychology will turn out to be a radically false theory of cognition. So we shouldn't expect to bridge our everyday psychological concepts with the new connectionist science. Rather, we'll just give up our belief/desire psychology (Churchland 1981).

Eliminitivists may be able to deny that they are individualists. Beliefs and desires are not 'in the head' because there are no beliefs and desires at all! It's beyond the scope of this paper to evaluate the merits of the eliminativist strategy. I'll simply observe that it would be unattractive to those who, I believe, rightly, don't want to throw out the folk psychological baby with the solipsistic bath water.[7]

To understand cognition we must place the cognitive agent in its environment. Connectionists may place more emphasis on this aspect of cognition than has traditional AI. Further, there may be other grounds for preferring connectionist approaches to AI over the traditional ones. But the philosophical commitment

to individualism/methodological solipsism is as strong for connectionism as it is for GOFAI. My view is that the philosophical objections considered are not fatal for either approach. Instead they suggest that the unit of investigation in cognitive science is wider than the individual cognitive agent.

Notes

1. Presented at the workshop: Artificial Intelligence: Emerging Science or Dying Art Form?, SUNY at Binghamton, June 21–23, 1990. I'd like to thank James Hendler, Terry Nutter, Jerold Aronson and an anonymous referee for this Journal for helpful comments and suggestions. Any difficulties which remain are my own.
2. Ludwig Wittgenstein flirted with solipsism in his early work. See Wittgenstein 1961 and McGinnis 1988.
3. I hesitate to use the term 'anti-solipsistic'. If solipsism is the view that one is the only thing in the universe, then anti-solipsism could be the (likely incoherent) view that everything in the universe except oneself exists!
4. Another set are due to Putnam, *op. cit.* I'm influenced by his views in what follows, though I only discuss Tyler Burge's counterexamples.
5. This defence of methodological solipsism was suggested by Terry Nutter in conversation.
6. It should be emphasized that the arguments presented against individualism do not depend on the question of determining the truth-value of the beliefs in question. Talk of truth value comes up only to illustrate, in the Ignat case, that the beliefs in the two worlds are different.
7. James Hendler endorses 'hybrid' approaches and thus supports this point (cf. Hendler 1989).

References

Burge, T. (1979) Individualism and the Mental. In *Midwest Studies in Philosophy*, IV (Minneapolis: University of Minnesota Press).

Chisholm, R. M. (1982) *The Foundations of Knowing* (Minneapolis: University of Minnesota Press).

Churchland, P. S. (1986) *Neurophilosophy* (Cambridge: MIT Press).

Churchland, P. (1981) Eliminative Materialism and Propositional Attitudes. *Journal of Philosophy*, **78** (2): 67–90.

Descartes, R. (1984) *The Philosophical Writings of Descartes*, Volume II, John Cottingham, Robert Stoothoff and Dugald Murdoch, trans. (Cambridge: Cambridge University Press).

Fodor, J. A. (1981) Methodological Solipsism Considered as a Research Strategy in Psychology. In *Representations: Philosophical Essays on the Foundations of Cognitive Science* (Cambridge: MIT Press).

Fodor, J. A. (1987) *Psychosemantics* (Cambridge: MIT Press).

Goldman, A. I. (1986) *Epistemology and Cognition* (Cambridge: Harvard University Press).

Graubar, S. R. (ed.) (1988) *The Artificial Intelligence Debate: False Starts, Real Foundations* (Cambridge: MIT Press).

Haugeland, J. (1985) *Artificial Intelligence: The Very Idea* (Cambridge: MIT Press).

Haugeland J. (ed.) (1981) *Mind Design* (Cambridge: MIT Press).

Hendler, J. (1989) On the Need for Hybrid Systems. *Connection Science*, **1** (3): 227–229.

Lycan, W. (1987) *Consciousness* (Cambridge: MIT Press).

McCarthy, J. (1988) The logistic approach to artificial intelligence. In S. R. Graubar (ed.) *The Artificial Intelligence Debate: False Starts Real Foundations* (Cambridge: MIT Press).

McGinnis, B. (1988) *Wittgenstein: A Biography* (Berkeley: California University Press).

Minsky, M. (1975) A Framework for Representing Knowledge. Reprinted in J. Haugleand (ed.) (Cambridge: MIT Press).

Newell, A., Rosenbloom, P. S. and Laird, J. E. (1989) Symbolic architectures for cognition. In M. I. Posner (ed.) *Foundations of Cognitive Science* (Cambridge: MIT Press).

Posner, M. I. (ed.) (1989) *Foundations of Cognitive Science* (Cambridge: MIT Press).

Putnam, H. (1981) *Reason, Truth and History* (Cambridge: Cambridge University Press).

Pylyshyn, Z. (1984) *Computation and Cognition* (Cambridge: MIT Press).

Rumelhart, D. E. *et al.* (1986) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* (Cambridge: MIT Press).

Rumelhart, D. E., McClelland, J. L. and the PDP Research Group (1986) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* (Cambridge: MIT Press).

Smart, J. J. C. (1959) Sensations and Brain Processes. *Philosophical Review*, **68**: 141–156.

Wittgenstein, L. (1961) *Tractatus Logico-Philosophicus* D. F. Pears and B. F. McGuinness, trans. (London: Routledge and Kegan Paul).