ELSEVIER

Robotics and Autonomous Systems 20 (1997) 251–256

Robotics and Autonomous Systems

Cog as a thought experiment

Daniel C. Dennett

Tufts University Medford, Department of Philosophy, Medford, MA 02155, USA

Keywords: Cog; Robot; Sentience; Philosophy; Methodology

In her presentation at the Monte Verità workshop, Maja Mataric showed us a videotape of her robots cruising together through the lab, and remarked, aptly: "They're flocking, but that's not what they think they're doing". This is a vivid instance of a phenomenon that lies at the heart of all the research I learned about at Monte Verità: the execution of surprisingly successful "cognitive" behaviors by systems that did not explicitly represent, *and did not need to explicitly represent*, what they were doing. How "high" in the intuitive scale of cognitive sophistication can such unwitting prowess reach? All the way, apparently, since I want to echo Maja's observation with one of my own: "These roboticists are doing philosophy, but that's not what they think they're doing". It is possible, then, even to do philosophy – that most intellectual of activities – without realizing that that is what you are doing. It is even possible to do it well, for this is a good, new way of addressing antique philosophical puzzles.

Then why on earth do I point this out? Why do I want to make these thinkers self-conscious about their activities? Won't they thereby run the usual risks of self-consciousness: a sudden deterioration in performance, diminished spontaneity, awkward re-working of already graceful competences? Yes, I might unleash monsters: roboticists who fancy themselves philosophers – an ugly prospect! But an underappreciated fact – underappreciated by this "intelligence without representation" gang – is that *sometimes* the deliberate and accurate representation, and even re-representation, of one's activities does yield huge increments in competence, in comprehension. I think the gains are worth the risks, but then I would; I'm a philosopher. In what follows, I will address those who engage in this research directly, rather than speaking about their work in the third person, since what I am offering is not just disinterested commentary, but advice – take it or leave it.

Why do I say you're doing philosophy? Because both your topics and your methods are those of philosophy – except where they are improvements thereon. Topics first. You are asking *very* abstract, general questions about the conditions under which perception, action, intelligence, and yes, even consciousness, can emerge in the world. This is a point that is often noted in passing by people in AI or ALife: by looking at deliberately simplified and artificial cases, you get to explore the fundamental requirements, the minimal conditions under which various necessary components of cognition can be obtained. As I have pointed out [4], these are versions of Immanual Kant's questions about the transcendentally necessary conditions of experience, but a major difference between traditional AI approaches, and the ALife and especially autonomous robots approaches, is that the latter take

PII S0921-8890(96)00065-6

advantage of Braitenberg's (1984) law of uphill analysis and downhill synthesis: it is much easier to deduce the behavioral competence of a system whose internal machinery you have synthesized than to deduce the internal machinery of a black box whose behavioral competence you have observed [7].

As for methods, you share with philosophy and traditional AI a fundamental reliance on thought experiments. In fact, one might say that the whole field consists of nothing but thought experiments. Not "regular" experiments? What about all the data-gathering on your robots' behaviors? Do mere thought experiments ever yield graphs? The key difference is that when the data don't come out the way you expected, you get to tinker with the robots, tuning them into conformity with the point you were setting out to demonstrate. The improvement over traditional, philosophical methods is that your thought experiments are prosthetically controlled and enhanced by the requirement that you actually make your models and demonstrate their competences. In philosophers' thought experiments, the sun always shines, the batteries never go dead, and the actors and props always do exactly what the philosophers' theories expect them to do. There are no surprises for the creators of the thought experiments, only for their audience or targets. As Ronald de Sousa has memorably said, much of philosophy is "intellectual tennis without a net". Your thought experiments have nets, but they are of variable height. "Proof of concept" is usually all you strive for, though sometimes that's all you get even though you are striving for more.

Don't change! I think that the tactic of varying the degree of difficulty, the degree of ambitiousness, of your demonstrations until you can find a feat that you can get to work is a perfectly acceptable practice. It is not "unprincipled"; it is shrewd, resourceful, opportunistic in a good sense. But outsiders often have difficulty seeing this. Workers in more traditional fields are dubious of demos that seem to help themselves at every turning to whatever simplifications are imposed on the demonstrator by the hard realities of practice. But how else are we going to find paths through this foggy world of cognition? The abstemious routes (one neuron at a time, or one rat at a time, or one day of field observations at a time, or . . .) are myopic, slow-motion trudges that are manifestly in need of some guidance and inspiration from high-flying scouts who are willing to live dangerously. At the same time, the advice I give to my philosopher colleagues and students is that they ought to consider flying at a somewhat lower altitude, taking on some of the problems of implementation, some of the real-world difficulties that you address. Their speculations are too easy, too unconstrained, too abstract to be trustworthy. The way to find the right level is to do some floating and see what works.

Traditional philosophical methodology does offer one practice that is not very apparent in your work, and might be, on occasion, a major corrective to your ways of thinking: historical scholarship. In the "hard" sciences, cumulative progress more or less obviates the need for contemporary students to have a detailed knowledge of the history of their field. There are historians of mathematics, physics, or chemistry who will dispute this vigorously, but they have a hard sell. Does today's biochemist really need to retrace the stumbling steps of the alchemists of yore, or re-create the long and arduous path of specific arguments that unified organic and inorganic chemistry? Aside from the sheer drama of it, does today's gene sequencer need to have an accurate knowledge of just how Crick and Watson reasoned their way beyond the confusions of their day about the chemistry of the vehicles of heredity? But in philosophy, to move to the other extreme, the only discernible dimension of progress is the replacement of one set of mistakes by another: thanks to our appreciation of the works of Plato, Aristotle, Descartes, Kant, etc., we don't make *their* mistakes any more – or at least not all of them. But since those philosophers were not dummies, their mistakes were not obvious, and are typically still enticing to the uninitiated, because they concern questions that are still non-routine, still unclear. These perennially tempting bad ideas are, you might say, slop basins of attraction that continue to exert their pull. The only way to protect yourself from these tempting errors is to study them in situ. Or get a good philosopher to explain them to you, in terms you can appreciate. You might like to supplement your current methods, then, with a little traditional philosophical investigation – reading a few books and articles, for instance – but not all that many. (A brief list of recommended reading in philosophy of mind is appended.)

One contribution of philosophy to an enterprise like this is simply to put your questions in the larger context

From “mere sensitivity” to consciousness?

artificial		*natural*
	transducers	
bimetallic spring		rhodopsin molecule
photocell		cone cell
	pseudo- (or micro-) agents	
thermostat		temperature maint. syst.
camera		VOR, vergence control, etc
	agents	
“robocteria”		bacteria, spermatazoa, . . .
“animats”		amoebas, jellyfish, . . .
. . . ?		plants?
. . . ?		fish, reptiles, . . .
. . . ?		birds, mammals, . . .
Cog		people

Fig. 1.

of human curiosity, both lay and professional. For instance, in his 1995 course notes, Rolf Pfeifer announces: “Our main goal is to relate behavior to internal mechanisms”, which is fine, of course, but many bystanders are going to say, “Behavior is all very well, but what about consciousness? Where does that come in”? It is worth remembering that to the average layperson, “conscious robot” is an oxymoron, a contradiction in terms. They are supremely confident that no mere “automaton” will ever be conscious. Meanwhile, our professional colleagues in cognitive science want to see more cognitive behaviors than mere phototaxis and herding, as Phil Husbands said in his presentation. They recognize that one must start with something simple, but they are skeptical that “more of the same” will ever add up to the sorts of cognitive competences they study from a more top-down perspective. Putting these two sorts of curiosity together, we can join our bystanders in wondering if there is a distinction between mere “sentience” on the one hand and fancier cognition (and human-style consciousness) on the other, and we may also want to address the question of whether even simple sentience is beyond our current models.

Fig. 1 shows a putative table of increasing sophistication, with natural entities lined up opposite their artificial counterparts, starting with parts-of-agents (transducers, pseudo- or micro-agents) and arriving via paths of increasing complexity and sophistication of both living and non-living agents at the (current) summit: Cog on the left and conscious human beings on the right. If we grant that all these entities, on both the artificial and natural side of the ledger, are equipped with varying degrees of *sensitivity*, we may ask whether some distinct phenomenon, *sentience*, makes its appearance somewhere on this trajectory, and if so, at what level of sophistication? Concentrate first on the right hand, natural side: Surely a cone cell in the retina is not sentient all by itself, whatever sentience is, nor is the vestibular ocular reflex machinery or the immune system or the temperature-maintenance system. And moving to whole agents, are jellyfish sentient, or are they merely sensitive and adaptive? And plants? Perhaps most people would reserve sentience for animals somewhere higher up the complexity ladder (fish yes, insects maybe – that sort of thing). If naive intuition puts the emergence of sentience fairly high up on this scale, it is no wonder that few if any observers are comfortable with the claim that any existing robots on the left exhibit sentience (let alone conscious), since for all their cleverness, they are surely at or below the level of unicellular organisms in their sophistication. We should not be bound by naive intuition, however, or feel particularly obligated to answer the questions posed on its behalf. We do well to recognize, nevertheless, that this is the mindset of the onlookers, and if we are misunderstood in our pronouncements, it may well be because we haven't taken that mindset into account.

What is sentience? In my new book, *Kinds of Minds*, 1996, I argue that the widely shared idea that *there is* a basic, animal sort of consciousness ("sentience") which some animals have and plants lack is an illusion, but it is undeniable that naive intuition suggests that sentience is something more than sensitivity, that

$$\text{sentience} = \text{sensitivity} + X.$$

And what is X? Are we leaving something out on the left-hand side? Rod Brooks spoke, amusingly, of "the Juice". We might well suspect, with Brooks, that we haven't got the Juice *yet*, but I gather we would also all declare that as far as we can see, we don't need any radically new breakthroughs (quantum gravity, psi-forces, morphic resonances, *élan vital*, ectoplasm) to add the Juice, or X, at some later stage. How will we ever test our common conviction? We shouldn't be impatient for a "scientific proof", but if we want some sanity checks along the way, Cog is a project that ought to provide insight, if not an outright answer, by attempting to model at the highest level (on the pretheoretical, intuitive scale of Fig. 1).

We want the behaviors (internal and external) exhibited by Cog to parallel those of a human infant, and eventually of course, an adult. We want first protolanguage and later language to crown our efforts; we want Cog to manifest curiosity, insight, fear, hope, pleasure, comprehension, friendship, . . . the works. Also we want to do this by building Cog out of "more of the same", proceeding just as evolution has proceeded, piling complexity on complexity. Cog must be always a going bodily concern, in which the particularly human competences – and tell-tale pathologies – can *emerge* from the interaction of all this new growth. We would love to see Cog exhibit paranoia or left neglect or obsessive-compulsive disorder as a result of a naturally arising imbalance or breakdown. Other pathological symptom clusters would not be welcome: for instance, coma, or autism.

Autism is a particularly automaton-like condition, as the term suggests, so providing Cog with the wherewithal to avoid autism, to establish and maintain normal contact with human beings, is an important priority. How should we do it? By installing what Alan Leslie [15] has called a TOMM or Theory of Mind Mechanism? This can be understood in a strong or a vacuous sense. In the vacuous sense, the TOMM is simply whatever features of Cog's brain prevent Cog from being autistic; in the strong sense, it suggests a Fodorian "module", a GOFAI organ equipped with axioms of belief and desire expressed with the use of multi-place predicates, a theorem-prover, and capable of deriving predictions along these lines:

$$\begin{aligned}
&\mathrm{Bel}_{\mathrm{ego}}\ \{\text{the candy is in the box}\}\\
&\mathrm{Bel}_{\mathrm{ego}}\ \{\mathrm{Bel}_{\mathrm{x}}\ [\text{the candy is in the jar}]\}\\
&\mathrm{Bel}_{\mathrm{ego}}\ \{\mathrm{Des}_{\mathrm{x}}\ [\text{that x obtain the candy}]\}\\
&\vdots\\
&\textit{ergo}:\\
&\mathrm{Bel}_{\mathrm{ego}}\ \{\text{x } \textit{will look in the jar}\}
\end{aligned}$$

Adding a GOFAI "module" of this sort at Cog's "summit" to handle what George Bush might call the vision thing is literally the last thing the Cog team would do. The GOFAI methodology, and GOFAI structures, are just too brittle, too unbiological. They are wrong as process models even if they are sometimes valuable ways of characterizing the competence (under idealized conditions). If we want to build a TOMM in the evolutionary, behavior-based spirit shared by the Monte Verità participants, how might we proceed? A few steps can be seen. Consider Elaine Morgan's [18, p. 99] comment:

> The heart-stopping thing about the new-born is that, from minute one, there is somebody there. Anyone who bends over the cot and gazes at it is being gazed back at.

As an observation about how we human observers instinctively react to eye contact, this is right on target, but it thereby shows that we can be easily misled. Cog's video camera eyes, unseeing as they still are, will saccade to focus on a newly arrived person who enters the room, and then track that person as he or she moves. Being tracked in this way is an oddly unsettling experience even for those in the know. Also staring into Cog's eyes while Cog stares mindlessly back can be quite "heart-stopping" to the uninitiated. Not surprisingly, this natural – indeed involuntary – tendency to draw the conclusion that "there is somebody there" has, in the natural world, a powerful element of truth. A built-in capacity for good gaze monitoring is a natural enabler of (not quite a logical prerequisite for) *unthinking* second-order intentionality, of the sort exhibited by piping plovers when they lead the predator away from the vulnerable nest by feigning injury [19]. Gaze monitoring is also a natural enabler of *shared attention*, which in turn plays a crucial role, as Baron-Cohen [2] shows, in developing language and other interpersonal skills. (I have begun collecting observations on eye contact from primatologists and ethologists. They note a striking difference between, for instance, the great apes and all other primates. To what extent does this explain only our *intuitive sense of greater kinship* – that "there is somebody there" inside the chimpanzee-suit – and to what extent does it mark a theoretically important difference? If and when Cog's eye-contact and gaze-monitoring skills are put to work in creating higher levels of shared understanding between Cog and its human companions we will surely get a better grip on this question.

I have argued [6] that linguistic skills, especially their proto-versions in such phenomena as infant babbling and semi-understood self-commentary (self-admonition, self-description) probably play a crucial role in permitting the infant brain to develop skill at "labeling" and then "manipulating" some of its own internal representations, representations that had heretofore been "embedded" [3] in the sorts of computational architectures that are fine for insects and simple animals (and human infants), but not for mature human cognition. Alan Turing [24], as so often before, points to one of the keys to progress: "If the untrained infant's mind is to become an intelligent one, it must acquire both discipline and initiative". The initiative must be an outgrowth of Cog's innate curiosity (or what I call epistemic hunger, which is trivially present in transducers, but must be an active feature of larger sub-systems), while the discipline Turing speaks of comes, I suspect, as a byproduct of the talents for speaking and listening, to oneself and others. That is a contentious claim much in need of further defense and investigation. (These sketchy ideas are further developed in [8].

I recommend the following books for reading. First, I think it would help to have some clarity about associationism and behaviorism, in its many varieties. If you go back to David Hume, his *Enquiry Concerning Human Understanding* of 1748 is the short, user-friendly version of the *Treatise*, which is a sprawling masterpiece. A good analysis of Hume's views is Barry Stroud's *Hume* (1977), in the useful Arguments of the Philosophers series. A fine anthology of historical and contemporary work is David Rosenthal's *The Nature of Mind* (1991). The two most important recent philosophical books are *The Concept of Mind* (1949) by Gilbert Ryle, and *Philosophical Investigations* (1958) by Ludwig Wittgenstein. Both Ryle and Wittgenstein were quite hostile to the idea of a scientific investigation of the mind, and the standard wisdom in the "cognitive revolution" is that we have seen through and beyond their ruthlessly unscientific analyses of the mental. Not true. One has to tolerate their often frustrating misperception of good scientific questions, and their almost total ignorance of biology and brain science, but they still managed to make deep and important observations which most of us are only now getting into position to appreciate. A fine (if cranky) book on the history of behaviorism in psychology by a philosopher is Charles Taylor, *The Explanation of Behaviour* (1964). For some balance, see the two chapters on the topic in my *Brainstorms* (1978), "Skinner Skinned", and "Why the Law of Effect Will Not Go Away". An excellent book on the "opposite" tradition in philosophy and its relevance to cognitive science is Patricia Kitcher's *Kant's Transcendental Psychology* (1990). For a window into contemporary work on mental representation and "propositional attitudes", I recommend my 1987 book, *The Intentional Stance*, especially Chapters 4–6. The philosopher whose work is most directly relevant to evolutionary roboticists is surely Ruth Garrett Millikan, whose *Language, Thought and Other Biological Categories* (1984) is a very difficult read, but worth it. So is her *White Queen Psychology and Other Essays for Alice* (1993). All this is just the tip of the iceberg, of course, but life is short, and none of these will waste your time.

For reliable quick probes into the ice, consult any of the flurry of recent encyclopedias on the topics, such as Samuel Guttenplan's *A Companion to the Philosophy of Mind* (1994), or Robert Audi's *Cambridge Dictionary of Philosophy* (1995) or Ted Honderich's *Oxford Companion to Philosophy* (1995). A still excellent older work is Paul Edwards' four-volume *Encyclopedia of Philosophy* (1967). A remarkably wide-ranging and fascinating encyclopedia, covering the history of psychology and neuroscience as well, is Richard Gregory's *Oxford Companion to the Mind* (1987).

References

[1] R. Audi, ed., *Cambridge Dictionary of Philosophy* (Cambridge University Press, Cambridge, MA 1995).

[2] S. Baron-Cohen, *Mindblindness: An Essay on Autism and Theory of Mind* (MIT Press, Cambridge, MA, 1995).

[3] A. Clark and A. Karmiloff-Smith, The cognizer's innards: A psychological and philosophical perspective on the development of thought, *Mind and Language* 8 (1993) 487–519.

[4] D. Dennett, *Brainstorms* (Bradford Books, Montgomery Vermont; later MIT Press, 1978).

[5] D. Dennett, *The Intentional Stance* (MIT Press, Cambridge, MA, 1987).

[6] D. Dennett, Learning and labeling (Commentary on Clark and Karmiloff-Smith), *Mind and Language* 8 (1993) 540–548.

[7] D. Dennett, Cognitive science as reverse engineering: Several meanings of 'Top-Down' and 'Bottom-Up', in: D. Prawitz, B. Skyrms and D. Westerst†hl, eds., *Logic, Methodology and Philosophy of Science* IX (Elsevier, Amsterdam, 1994) 679–689.

[8] D. Dennett, *Kinds of Minds* (Basic Books, New York, 1996).

[9] P. Edwards, ed., *Encyclopedia of Philosophy* (MacMillan/Free Press, London/New York, 1967).

[10] R. Gregory *Oxford Companion to the Mind* (Oxford University Press, Oxford, 1987).

[11] S. Guttenplan, ed., *A Companion to the Philosophy of Mind* (Blackwell, Oxford, 1994).

[12] T. Honderich, ed., *Oxford Companion to Philosophy* (Oxford University Press, Oxford, 1995).

[13] D. Hume, *Enquiry Concerning Human Understanding* (available in many paperback editions).

[14] P. Kitcher, *Kant's Transcendental Psychology* (Oxford University Press, New York, 1990).

[15] A. Leslie, TOMM, ToBy, and Agency: Core architecture and domain specificity, in: L. Hirschfeld and S. Gelman, eds., *Mapping the Mind: Domain Specificity in Cognition and Culture* (Cambridge University Press, Cambridge, MA, 1994).

[16] R.G. Millikan, *Language, Thought and Other Biological Categories* (MIT Press, Cambridge, MA, 1984).

[17] R.G. Millikan, *White Queen Psychology and Other Essays for Alice* (MIT Press, Cambridge, MA, 1993).

[18] E. Morgan, *The Descent of the Child: Human Evolution from a New Perspective* (Oxford University Press, Oxford, 1995).

[19] C. Ristau, Aspects of the cognitive ethology of an injury-feigning bird, the piping plover, in: Carolyn Ristau, ed., *Cognitive Ethology* (Erlbaum, Hillsdale, NJ, 1991) 91–126.

[20] D. Rosenthal, *The Nature of Mind* (Oxford University Press, New York, 1991).

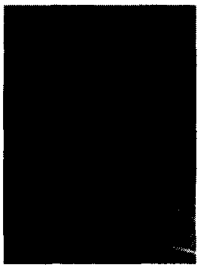
[21] G. Ryle, *The Concept of Mind* (Hutchinson, London, 1949).

[22] B. Stroud, *Hume* (Routledge and Kegan Paul, London, 1977).

[23] C. Taylor, *The Explanation of Behaviour* (Routledge and Kegan Paul, London, 1964).

[24] A. Turing, Intelligent Machinery, in: B. Meltzer and D. Michie, eds., *Machine Intelligence 5* (Halste Press, New York, 1970).

[25] L. Wittgenstein, *Philosophical Investigations* (Blackwell, Oxford, 1958).



Doniel C. Dennett is Distinguished Professor of Arts and Sciences and Director of the Center for Cognitive Studies at Tufts University. His books include *Consciousness Explained* (1991), *Darwin's Dangerous Idea* (1995) and *Kinds of Minds* (1996).