J. EXPT. THEOR. ARTIF. INTELL. 12(2000)247–250

Introduction to the special issue on philosophical foundations of artificial intelligence

VAROL AKMAN

Department of Computer Engineering, Bilkent University, Bilkent, Ankara 06533, Turkey
e-mail: akman@cs.bilkent.edu.tr

> Philosophy, like piloting, is mostly about figuring out where you are. The basic principle of both is much the same: find an object whose position is known and locate yourself with respect to it.
> (Fodor, 1998: ix)

The web site of Taylor & Francis, the publisher of the *Journal of Experimental and Theoretical Artificial Intelligence*, notes that the foundations of the company were laid in 1798 when Richard Taylor launched the Philosophical Magazine. It thus gives me a thrill to be able to write, more than two centuries later, the introduction to a *JETAI* special issue highlighting philosophical foundations of AI. The project basically took off during the early days of 1999 when I started to invite the authors who contributed the seven thought-provoking papers to the issue you are holding in your hands. The initial call for papers solicited contributions treating one or more of the following themes:

(1) Embodied cognition
(2) The nature of computation and the mind
(3) Dynamic systems and computation
(4) The nature of representation
(5) Physical symbol systems vs connectionism
(6) The symbol grounding problem (excluding discussions of the Chinese Room)
(7) Intelligence without representation
(8) Perception and action; space and time; causation
(9) Quantum physics and AI
(10) The phenomenological agenda
(11) Situation theory
(12) Modelling knowledge, belief, intention, desire, etc.
(13) The frame problem (excluding logic solutions)
(14) Contextual reasoning

Perhaps not surprisingly, I had to invite many in order to get the final seven, which I think are representative of some of the best work that is being done today at the interface of philosophy and AI. Let me quickly add that only contributions that bear a strongly research-oriented flavour have been considered and all papers have gone through a careful revision process.

The premise that gave shape to the project was simply this: philosophy has been, right from the early days of AI, a close and dependable ally of researchers working on the foundations of AI. One only needs to recall a landmark article (Turing 1950)

Journal of Experimental & Theoretical Artificial Intelligence ISSN 0952-813X print/ISSN 1362-3079 online
http://www.tandf.co.uk/journals

published half a century ago in a famous philosophy journal. It is also comforting to know that the cross-fertilisation of the two areas has been going since the fifties and that the most encyclopedic AI textbook of our time (Russell and Norvig 1995) devotes a reasonable space—of approximately 25 pages—to the philosophical foundations of AI.

Sloman and McCarthy, two major proponents of philosophical research in AI, eloquently argued on several occasions that realistic AI requires arming a programme with a philosophy. Their IJCAI-95 papers (McCarthy 1995, Sloman 1995) consolidate the need for deeper inquiries into philosophical foundations by suggesting numerous pathways for the growth of philosophical AI. In the words of McCarthy, 'AI needs many ideas that have hitherto been studied only by philosophers.'

Inspired by Sloman and McCarthy's advice, this special issue brings together articles describing strong foundational work in AI with an unmistakable philosophical flavour. In addition to papers approaching essential AI problems via existing philosophical techniques or concepts, there are contributions that give rise to fresh appraisals of contemporary issues in the philosophies of mind and language (and even ethics as the next paragraph shows).

Allen *et al.*'s paper, 'Prolegomena to any future artificial moral agent', breaks new ground by considering the possibility of a Turing test treating moral issues. As technological advances are carrying us all to a future populated by fully autonomous robots, important ethical questions arise. For instance, what if a robot decides to make its living by cheating, theft, or even murder? How can we make sure that such agents—while greatly differing in their preferences—lead a plausible life not harmful to humans (and each other)? It may be argued that human-like morality, with all its inclinations towards immoral action, should not be a model for robots.[1]

How could a material entity have conscious states? Much ink has been spilled on this question and the most complete contemporary account of proposed answers can be found in (Block *et al.* 1997). Aydede and Güzeldere's 'Consciousness, intentionality, and intelligence: some foundational issues for artificial intelligence' emphasizes this question once again, citing its seemingly evasive nature. They first explain why consciousness poses puzzling problems and then advance a convincing proposal that can explain its tricky aspects. In a nutshell, their recipe for phenomenal consciousness is to construct systems whose sensory and cognitive/introspective architectures parallel their proposal.

'Contextual reasoning distilled', one of the two technical papers in this issue, is due to Benerecetti *et al.* When people act in a certain way or make a particular statement, they do so in a context. Consequently, in all the things they do or say, there are background assumptions detectable only through the context. The authors' analysis in this paper suggests that the dimensions of contextual representation is probably threefold, *viz.* The limited part of the world—also known as a 'situation'—covered by the representation, the level of detail (granularity) at which this part is described, and the cognitive perspective from which the representation is given. Contextual reasoning thus becomes essentially the problem of determining how different contexts relate to each other.

Bringsjord and Xiao's highly technical contribution, 'A refutation of Penrose's Gödelian case against artificial intelligence', is written as a thorough examination of Penrose's anti-computationalist stance, elucidated in his best selling The Emperor's New Mind and then in a more recent sequel to it (Penrose 1994). Their painstaking analyses of Penrose's arguments show that there are a number of defects of logical disposition. It is engaging to note that the authors themselves also believe that Gödelian results can be used to demonstrate that the mind is beyond computation; their criticism is just that Penrose's attempt fails in this endeavour.

French's 'Peeking behind the screen: the unsuspected power of the standard Turing Test' is along the lines of his influential paper (French 1990). He argues that the Turing Test—in its standard version (Turing 1950)—is already very difficult and that coming up with even more involved versions of it would be an overkill. French's crucial observation is that it is easy to devise a set of subcognitive questions to obliquely probe the human mind, which houses, among other things, reminiscences of experiences with the world. In order to answer these questions a machine would have to experience the world in the same fashion as a human being does.[2]

McCarthy's contribution 'Free will—even for robots' regards human free will as an outcome of evolution and predicts that robots of the future will need it. He believes that free will does not necessitate a knotty system. Rather, it is a graded capability (a robot can have more free will than another can) and his bold standpoint is that even average computer systems can represent free will and act accordingly. The distinction he draws between having choices and being conscious of these choices—nicely captured by the formula I can, but I won't—is an important one.[3]

In his inventively titled contribution, 'Producing mind', Torrance studies the cognitive and phenomenal aspects of mind. Like Aydede and Güzeldere, he foresees great promise in realizing a subjective (experiential) and productive (cognitive) mind computationally. The paper scrutinises the interaction between mental productivity and phenomenal consciousness, which results in semantic content.

With this brief appraisal of the papers making up this special issue, it is time for me to leave the floor to the authors themselves. McCarthy 1995) notes that 'many philosophical problems take new forms when thought about in terms of how to design a robot' and issues a warning: 'some approaches to philosophy are helpful and others are not'. I do hope that you'll find the approaches here useful in whatever philosophical study of AI you may be currently occupied with. In the spirit of Fodor's remark, it is my sincere belief that the illuminating works in the sequel will help you figure out where you are Happy piloting!

Acknowledgements

I am grateful to Eric Dietrich, the editor-in-chief of JETAI, for invaluable advice and support. Without his encouragement, this special issue would not be possible.

Notes

1. There is a demand here for a universal non-harm credo, together with a reasonably effective mutual-aid principle (Narveson 1997), both specified in computationally tractable terms.
2. Interestingly, Dennett (1998: 13) remarks that '[Turing's] point was that we should not be species-chauvinistic, or anthropocentric, about the insides of an intelligent being, for there might be inhuman ways of being intelligent'.
3. It is instructive to note that some of the ideas McCarthy treats have been introduced originally in a milestone paper of philosophical AI (McCarthy and Hayes 1969).

References

Block, N., Flanagan, O., and Güzeldere, G. (eds), 1997, *The Nature of Consciousness: Philosophical Debates* (Cambridge MA: MIT Press).

Dennett, D. C., 1998, *Can Machines Think? In his Brainchildren: Essays on Designing Minds* (Cambridge MA: MIT Press), pp. 3–20. Originally appeared in 1985 in M. Shafto (ed.), *How We Know* (San Francisco: Harper & Row).

Fodor, J., 1998, *In Critical Condition: Polemical Essays on Cognitive Science and the Philosophy of Mind* (Cambridge MA: MIT Press).

French, R., 1990, Subcognition and the limits of the Turing test. *Mind*, **99**: 53–65. Available at; ftp://forum.fapse.ulg.ac.be/pub/techreports/turing.pdf

McCarthy, J., 1995, What has AI in common with philosophy? Paper given at a panel at IJCAI-95: 'A philosophical encounter: An interactive presentation of some of the key philosophical problems in AI and AI problems in philosophy', Montreal. Available at; http://www-formal.stanford.edu/jmc/aiphil.html

McCarthy, J., and Hayes, P. J., 1969, Some philosophical problems from the standpoint of artificial intelligence, In B. Meltzer and D. Michie (eds) *Machine Intelligence 4* (Edinburgh: Edinburgh University Press), pp. 463–502. Available at; http://www-formal.stanford.edu/jmc/mcchay69/mcchay69.html

Narveson, J., 1997, Egalitarianism: partial, counterproductive, and baseless *Ratio* (New Series) **X**: 280–295.

Penrose, R., 1994, *Shadows of the Mind* (Oxford: Oxford University Press).

Russell, S., and Norvig, P., 1995, *Artificial Intelligence: A Modern Approach* (Englewood Cliffs NJ: Prentice-Hall).

Sloman, A., 1995, A philosophical encounter, Paper given at a panel at IJCAI-95: 'A philosophical encounter: an interactive presentation off some of the essential philosophical problems in AI and AI problems in philosophy', Montreal. Available at; http://www.cs.bham.ac.uk/~axs/cog_affect/ijcai95.txt

Turing, A., 1950, Computing machinery and intelligence. Mind, **59**: 433–460. Available at; http://dangermouse.uark.edu/ai/Turing.html