Minds & Machines (2012) 22:67–69
DOI 10.1007/s11023-012-9278-y

INTRODUCTION

Introduction: Philosophy and Theory of Artificial Intelligence

Vincent C. Müller

Received: 31 January 2012 / Accepted: 18 May 2012 / Published online: 6 June 2012

The conference and this volume are dedicated to the memory of John Haugeland, who remains an inspiration to us all.

What is PT-AI?

The theory and philosophy of artificial intelligence has come to a crucial point where the agenda for the forthcoming years is in the air. This special volume of *Minds and Machines* presents leading invited papers from a conference on the "Philosophy and Theory of Artificial Intelligence" that was held in October 2011 in Thessaloniki (www.pt-ai.org).

Artificial Intelligence is perhaps unique among engineering subjects in that it has raised very basic questions about the nature of computing, perception, reasoning, learning, language, action, interaction, consciousness, humankind, life etc. etc.—and at the same time it has contributed substantially to answering these questions (in fact, it is sometimes seen as a form of empirical research). There is thus a substantial tradition of work, both on AI by philosophers and of theory within AI itself.

The classical theoretical debates have centred on the issues whether AI is possible at all (often put as "Can machines think?") or whether it can solve certain problems ("Can a machine do x?"). In the meantime, technical AI systems have progressed massively and are now present in many aspects of our environment. Despite this development, there is a sense that classical AI is inherently limited, and must be replaced by (or supplanted with) other methods, especially neural networks, embodied cognitive science, statistical methods, universal algorithms, emergence,

V. C. Müller (✉)
Anatolia College/ACT, Thessaloniki, Greece
e-mail: vmueller@act.edu
URL: www.sophia.de

V. C. Müller
University of Oxford, Oxford, UK

behavioural robotics, interactive systems, dynamical systems, living and evolution, insights from biology & neuroscience, hybrid neuro-computational systems, etc.

After Classical Artificial Intelligence?

We are now at a point where we can see more clearly what the alternatives are. The classical 'computationalist' view was that cognition is computation over representations, which may thus take place in any computational system, natural or artificial. On this classical view, AI and Cognitive Science are two sides of the same coin—this view had fuelled a large part of the philosophical and theoretical interest in AI. However, most of the defining features of this old consensus are now under threat: computation is digital; representation is crucial for cognition; embodiment, action and interaction are not; the distinction between living and non-living agents is irrelevant; etc. So, should we drop the classical view, should we supplement it, or should we defend it in the face of modish criticism? These philosophical debates are mirrored in technical AI research, which has been moving on (for the most part), regardless of the 'worries' from the theorists; but some sections have changed under the impression of classical criticism while new developments try to shed the classical baggage entirely. In any case, the continued technical success has left an impression: We are now much more likely to discuss human-level AI (whatever that means) in machines as a real possibility.

Given where we stand now, the relation between AI and Cognitive Science needs to be re-negotiated—on a larger scale this means that the relation between technical products and humans is re-negotiated. How we view the prospects of AI depends on how we view ourselves and how we view the technical products we make; this is also the reason why the theory and philosophy of AI needs to consider such apparently widely divergent issues from human cognition and life to technical functioning.

What Now?

A bewildering mass of questions spring to mind: Should we repair classical AI, since intelligence is still input–output information processing? Drop the pretence of general intelligence and continue on the successes of technical AI? Embrace embodiment, enactivism or the extended mind? Revive neural networks in a new form? Replace AI by 'cognitive systems'? Look for alternative systems, dynamic, brain-inspired, …? And what about the classical problems that Dreyfus, Searle, Haugeland or Dennett had worked on; what about meaning, intention, consciousness, expertise, free will, agency, etc.? Perhaps AI was blind in limiting itself to human-level intelligence, so why not go beyond? What would that mean and what would its ethical implications be? What are the ethical problems of AI even now and in the foreseeable future?

The discussion on the future of AI seems to open three different directions. The first is AI that continues, based on technical and formal successes, while re-claiming the original dream of a universal intelligence (sometimes under the heading of 'artificial general intelligence'). This direction is connected to the now acceptable notion of the 'singular' event of machines surpassing human intelligence—it plays a central role in Bostrom's and Dreyfus' papers here.

The second direction is defined by its rejection of the classical image, especially its rejection of representation (as in Brooks' 'new AI'), its stress of embodiment of agents and on the 'emergence' of properties, especially due to the interaction of agents with their environment—O'Regan is a clear example of this direction.

A third direction is to take on new developments elsewhere. One approach is to start with neuroscience; this typically focuses on dynamical systems and tries to model more fundamental processes in the cognitive system than classical cognitive science did. Other approaches of more general 'systems' subvert the notion of the 'agent' and locate intelligence in wider systems.

Finally, there are many approaches that try to combine the virtues of the various approaches towards practical results, especially systems that are more autonomous and robust in real-world environments. These approaches are often pushed by funding agencies; the National Science Foundation (USA) supports 'Cybertechnical Systems' while the European Commission sponsors 'Artificial Cognitive Systems'. (I happen to coordinate "EUCog", a large network of researchers in this context.)

Reclaiming AI: Back to Basics

The basic problems of AI remain and ignoring them 'because our systems are getting better anyway' is a risky strategy, as Dreyfus explains in his paper where he warns of the many instances of the 'first step fallacy' (Dreyfus could have rested on his laurels and said, 'I told you so 40 years ago' but he takes the line to the present). The way to move forward in this context seems to go back to basics … and of course, philosophers are likely to do this in any case. There are a few basic notions that are fundamental for the decisions in this debate and also, the basic problems have significant backward relevance for philosophy (if we can say something about free will in machines, for example, this has direct repercussions on how we see free will in humans).

Unsurprisingly, the basic issues are *computation*, *cognition* and *ethics*—and this is what the papers in this volume address: Shagrir asks which physical systems implement a computation that is sufficient for cognition; Gomila/Travieso/Lobo explain the conditions for a mind to be 'systematic' and O'Regan explains his sensorimotor approach to the conscious 'feel', while Bostrom discusses the relation between intelligence and goals or motivations of an agent.

Further work on these issues is to be found in the companion volume to this journal issue, which will be published as "Philosophy and Theory of AI" with Springer in 2012. We expect to hold further events and other activities in this field—watch pt-ai.org!