Philosophy and Theory of Artificial Intelligence

Can we make machines that think and act like humans or other natural intelligent agents? The answer to this question depends on how we see ourselves and how we see the machines in question. Classical AI and cognitive science had claims that cognition is computation, and can thus be reproduced on other computing machines, possibly surpassing the abilities of human intelligence. This consensus has now come under threat and the agenda for the philosophy and theory of AI must be set anew, re-defining the relation between AI and Cognitive Science.

We can re-claim the original vision of general AI from the technical AI disciplines; we can reject classical cognitive science and replace it with a new theory, e.g. embodied; or we can try to find new ways to approach AI, for example from neuroscience or from systems theory. To do this, we must go back to the basic questions on computing, cognition and ethics for AI. The 36 papers in this volume provide cutting-edge work from leading researchers that define where we stand and where we should go from here.
Studies in Applied Philosophy, Epistemology and Rational Ethics

Editor-in-Chief
Prof. Dr. Lorenzo Magnani
Department of Arts and Humanities
Philosophy Section
University of Pavia
Piazza Botta 6
27100 Pavia
Italy
E-mail: lmagnani@unipv.it

Editorial Board
Prof. Atocha Aliseda
Instituto de Investigaciones Filosóficas, Universidad Nacional Autónoma de México (UNAM), Ciudad Universitaria, Coyoacan, 04510, Mexico, D.F.
E-mail: atocha@filosoficas.unam.mx

Prof. Giuseppe Longo
Laboratoire et Departement d’Informatique, CREA, CNRS and Ecole Polytechnique, LIENS, 45, Rue D’Ulm, 75005 Paris, France
E-mail: Giuseppe.Longo@ens.fr

Prof. Chris Sinha
Centre for Languages and Literature, P.O. Box 201, 221 00 Lund, Sweden
E-mail: chris.sinha@ling.lu.se

Prof. Paul Thagard
Department of Philosophy, Faculty of Arts, Waterloo University, Waterloo, Ontario,
Canada N2L 3G1
E-mail: pthagard@uwaterloo.ca

Prof. John Woods
Department of Philosophy, University of British Columbia, 1866 Main Mall BUCH E370,
Vancouver, BC Canada V6T 1Z1
E-mail: john.woods@ubc.ca

For further volumes:
http://www.springer.com/series/10087
Philosophy and Theory of Artificial Intelligence
Dedicated to the memory of John Haugeland who remains an inspiration to us all
Introductory Note:
Philosophy and Theory of Artificial Intelligence

Vincent C. Müller
Anatolia College/ACT & University of Oxford
e-mail: www.sophia.de
16.05.2012

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

2 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.

3 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.)

4 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. 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 & methods, cognition and ethics & society – and this is what the papers in this volume address.

The papers published here have passed two high hurdles: they have been blind peer reviewed as long abstracts and those who passed were reviewed a second time as full papers. A list of the distinguished members of the program committee can be found on our website.

Further work on these issues is to be found in the companion volume to this book, which will appear as a special volume of the journal Minds and Machines in 2012. We expect to hold further events and other activities in this field – watch pt-ai.org!
Contents

Computation and Method

Machine Mentality? .................................................. 1
Istvan S.N. Berkeley, Claiborne Rice

‘Quantum Linguistics’ and Searle’s Chinese Room Argument ........ 17
John Mark Bishop, Slawomir J. Nasuto, Bob Coecke

The Physics and Metaphysics of Computation and Cognition ....... 29
Peter Bokulich

The Two (Computational) Faces of AI ................................ 43
David Davenport

The Info-computational Nature of Morphological Computing ....... 59
Gordana Dodig-Crnkovic

Limits of Computational Explanation of Cognition .................. 69
Marcin Młkowski

Of (Zombie) Mice and Animats ...................................... 85
Slawomir J. Nasuto, John Mark Bishop

Generative Artificial Intelligence .................................... 107
Tijn van der Zant, Matthijs Kouw, Lambert Schomaker

Cognition

Turing Revisited: A Cognitively-Inspired Decomposition ............ 121
Tarek Richard Besold
<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>The New Experimental Science of Physical Cognitive Systems:</td>
<td>133</td>
</tr>
<tr>
<td>AI, Robotics, Neuroscience and Cognitive Sciences under a New Name</td>
<td></td>
</tr>
<tr>
<td>with the Old Philosophical Problems?</td>
<td></td>
</tr>
<tr>
<td>Fabio Bonsignorio</td>
<td></td>
</tr>
<tr>
<td>Toward a Modern Geography of Minds, Machines, and Math</td>
<td>151</td>
</tr>
<tr>
<td>Selmer Bringsjord, Naveen Sundar Govindarajula</td>
<td></td>
</tr>
<tr>
<td>Practical Introspection as Inspiration for AI</td>
<td>167</td>
</tr>
<tr>
<td>Sam Freed</td>
<td></td>
</tr>
<tr>
<td>“Computational Ontology and Deontology” in Vincent C. Müller</td>
<td>179</td>
</tr>
<tr>
<td>“Theory and Philosophy of Artificial Intelligence”</td>
<td></td>
</tr>
<tr>
<td>Raffaela Giovagnoli</td>
<td></td>
</tr>
<tr>
<td>Emotional Control–Conditio Sine Qua Non for Advanced Artificial</td>
<td>187</td>
</tr>
<tr>
<td>Intelligences?</td>
<td></td>
</tr>
<tr>
<td>Claudioius Gros</td>
<td></td>
</tr>
<tr>
<td>Becoming Digital: Reconciling Theories of Digital Representation</td>
<td>199</td>
</tr>
<tr>
<td>and Embodiment</td>
<td></td>
</tr>
<tr>
<td>Harry Halpin</td>
<td></td>
</tr>
<tr>
<td>A Pre-neural Goal for Artificial Intelligence</td>
<td>215</td>
</tr>
<tr>
<td>Micha Hersch</td>
<td></td>
</tr>
<tr>
<td>Intentional State-Ascription in Multi-Agent Systems: A Case Study</td>
<td>225</td>
</tr>
<tr>
<td>in Unmanned Underwater Vehicles</td>
<td></td>
</tr>
<tr>
<td>Justin Horn, Nicodemus Hallin, Hossein Taheri, Michael O’Rourke,</td>
<td></td>
</tr>
<tr>
<td>Dean Edwards</td>
<td></td>
</tr>
<tr>
<td>Snapshots of Sensorimotor Perception: Putting the Body Back into</td>
<td>237</td>
</tr>
<tr>
<td>Embodiment</td>
<td></td>
</tr>
<tr>
<td>Anthony F. Morse</td>
<td></td>
</tr>
<tr>
<td>Feasibility of Whole Brain Emulation</td>
<td>251</td>
</tr>
<tr>
<td>Anders Sandberg</td>
<td></td>
</tr>
<tr>
<td>C.S. Peirce and Artificial Intelligence: Historical Heritage and (New)</td>
<td>265</td>
</tr>
<tr>
<td>Theoretical Stakes</td>
<td></td>
</tr>
<tr>
<td>Pierre Steiner</td>
<td></td>
</tr>
<tr>
<td>Artificial Intelligence and the Body: Dreyfus, Bickhard, and the</td>
<td>277</td>
</tr>
<tr>
<td>Future of AI</td>
<td></td>
</tr>
<tr>
<td>Daniel Susser</td>
<td></td>
</tr>
<tr>
<td>Introducing Experion as a Primal Cognitive Unit of Neural</td>
<td>289</td>
</tr>
<tr>
<td>Processing</td>
<td></td>
</tr>
<tr>
<td>Oscar Vilarroya</td>
<td></td>
</tr>
</tbody>
</table>
Contents

The Frame Problem: Autonomy Approach versus Designer Approach .................................................... 307
Aziz F. Zambak

Ethics and Society

Machine Intentionality, the Moral Status of Machines, and the Composition Problem .................................................... 321
David Leech Anderson

Risks and Mitigation Strategies for Oracle AI .................................................. 335
Stuart Armstrong

The Past, Present, and Future Encounters between Computation and the Humanities .................................................. 349
Stefano Franchi

Being-in-the-AmI: Pervasive Computing from Phenomenological Perspective .................................................. 365
Gagan Deep Kaur

The Influence of Engineering Theory and Practice on Philosophy of AI .................................................. 375
Viola Schiaffonati, Mario Verdicchio

Artificial Intelligence Safety Engineering: Why Machine Ethics Is a Wrong Approach .................................................. 389
Roman V. Yampolskiy

What to Do with the Singularity Paradox? .................................................. 397
Roman V. Yampolskiy

Author Index .................................................. 415