THE PHILOSOPHY OF ARTIFICIAL INTELLIGENCE

Time and Place: Tuesdays 16:15-17:45, 23.02/U1.22
Instructor: Dr. Ioannis Votsis
E-mail: votsis@phil-fak.uni-duesseldorf.de
Office hours (Room Geb. 23.21/04.86): Wednesdays 11:00-12:00

This course aims to probe the philosophical presuppositions that underlie discussions of artificial intelligence. We will start with a semi-historical introduction to several central issues and notions in the debate on artificial intelligence, e.g. notions of intelligence, the definition of an algorithm, and the Turing test. We will then proceed to examine to what extent the debate has moved forward since. For example, it was once argued that a computer would never be able to match humans in complicated reasoning tasks, including chess-playing at a world-class level. As is well known, computers have now surpassed that milestone. How much more can we expect computers to be able to achieve in the future? Will computers be able to think? If not, can their outputs at least be indistinguishable from those of human beings? What about consciousness? Will there ever be computers that are conscious and have human-like feelings? To take this line of reasoning to the limit, will there ever be computers that exceed humans in all these respects, i.e. that become 'superhuman'? In pursuing answers to these and other related questions we will delve into discussions of behaviourism, consciousness, reductionism, intentionality, the Chinese room argument, simulation, neural networks, expert systems, the frame problem, etc.

Useful Anthology:
Boden, Margaret (ed.) (1990) *The Philosophy of Artificial Intelligence*, Oxford: Oxford University Press.

Coursework:
- One presentation (about 20 minutes) on one of the main readings. [3 credits]
- One essay (about 2,500 words), **deadline 07/07/09**. [3 credits]

NB: Presentations will be assigned on the second week. Suggested essay topics will be distributed in May.

WEEK 1: Introduction and Presentation Assignments

WEEK 2: Computability and the Turing Test

Main Reading:
Turing, A. M. (1950) 'Computing Machinery and Intelligence', *Mind*, vol. 59, 433-460.

Further Reading:
Block, N. (1981) 'Psychologism and Behaviorism', *Philosophical Review*, vol. 90: 5–43.
Copeland, B. (2000) 'The Turing Test', *Minds and Machines*, vol. 10 (4):519-539.
Harnad, S. (1991) 'Other Bodies, Other Minds: A Machine Incarnation of an Old Philosophical Problem', *Minds and Machines*, vol. 1(1):43-54.

WEEK 3: A Modern Philosophical Theory of Mind

Main Reading:

Levin, J. (2009) 'Functionalism', *Stanford Encyclopedia of Philosophy*, http://plato.stanford.edu/entries/functionalism/

Further Reading:

Braddon-Mitchell, D. and F. Jackson (1996) *Philosophy of Mind and Cognition*, Oxford: Basil Blackwell.

Heil, J. (2004) *Philosophy of Mind*: *A Contemporary Introduction*, London: Routledge.

Rey, G. (1997) *Contemporary Philosophy of Mind*, Cambridge, MA: Blackwell.

WEEK 4: Qualia

Main Reading:

Jackson, F. (1986) 'What Mary Didn't Know', *Journal of Philosophy*, vol. 83: 291-5.

Further Reading:

Block, N. (1980) 'Troubles with Functionalism', in N. Block (ed.), *Readings in the Philosophy of Psychology*, Cambridge, MA: Harvard University Press, pp. 268-305.

Chalmers, D. J. (1995) 'Absent Qualia, Fading Qualia, Dancing Qualia', in *Conscious Experience*, T. Metzinger (ed.), Thorverton: Imprint Academic.

Nagel, T. (1974) 'What Is It Like To Be a Bat?', *Philosophical Review*, 83: 435–450.

WEEK 5: Connectionism

Main Reading:

Smolensky, P. (1988) 'On the Proper Treatment of Connectionism', *Behavioral and Brain Sciences*, vol. 11(1): 1-23.

Further Reading:

Churchland, P. M. (1989) *A Neurocomputational Perspective*: *The Nature of Mind and the Structure of Science*, Cambridge, MA: MIT Press.

Horgan, T., and J. Tienson (1996) *Connectionism and the Philosophy of Psychology*, Cambridge, MA: MIT Press.

Ramsey, W., S. Stich and J. Garon (1991) 'Connectionism, Eliminativism and the Future of Folk Psychology', in W. Ramsey, S. Stich and D. Rumelhart (eds.), *Philosophy and Connectionist Theory*, Hillsdale NJ: Lawrence Erlbaum.

WEEK 6: Classical Theories vs. Connectionism

Main Reading:

Fodor, J. A. and Z. W. Pylyshyn (1988) 'Connectionism and Cognitive Architecture', *Cognition*, vol. 28:3-71.

Further Reading:

Chalmers, D. J. (1993) 'Connectionism and Compositionality: Why Fodor and Pylyshyn were Wrong', *Philosophical Psychology*, vol. 6 (3):305-319.

Fodor, J. and B. McLaughlin (1990) 'Connectionism and the Problem of Systematicity: Why Smolensky's Solution doesn't Work', *Cognition*, vol. 35:183-205.

Hadley, R. F. (1994) 'Systematicity in Connectionist Language Learning', *Mind and Language*, vol. 9:247-72.

WEEK 7: The Chinese Room Argument

Main Reading:

Searle, J.R. (1980) 'Minds, Brains and Programs', reprinted in *The Philosophy of Artificial Intelligence*, M. Boden (ed.), Oxford: Oxford University Press, 1990, pp. 67-88.

Further Reading:

Boden, M. (1988) 'Escaping from the Chinese Room', reprinted in *The Philosophy of Artificial Intelligence*, M. Boden (ed.), Oxford: Oxford University Press, 1990, pp. 89-104.

Churchland, P.M and P.S. Churchland (1990) 'Could a Machine Think?', *Scientific American*, pp. 32-37.

Searle, J.R. (1990) 'Is the Brain's Mind a Computer Program?', *Scientific American*, vol. 262(1):26-31.

WEEK 8: The Chinese Room Revisited

Main Reading:

Block, N. (2002) 'Searle's Arguments Against Cognitive Science', in J. Preston and M. Bishop (eds.) *Views into the Chinese Room*: *New Essays on Searle and Artificial Intelligence*, Oxford: Oxford University Press, pp. 70-79.

Further Reading:

Chalmers, D. J. (1992) 'Subsymbolic Computation and the Chinese Room', in J. Dinsmore (ed.), *The Symbolic and Connectionist Paradigms*: *Closing the Gap*, Hillsdale NJ: Lawrence Erlbaum, pp. 25-48.

Copeland, B. (1993) 'The Curious Case of the Chinese Gym', *Synthese*, vol. 95 (2):173-86.

Searle, J.R. (2002) 'Twenty-one Years in the Chinese Room', in J. Preston and M. Bishop (eds.) *Views into the Chinese Room*: *New Essays on Searle and Artificial Intelligence*, Oxford: Oxford University Press, pp. 51-69.

WEEK 9: The Limits of Formalisation

Main Reading:

Dreyfus, H. L. (1992) *What Computers Still Can't Do*, Cambridge (MA): MIT Press, introductory chapter, pp. 67-90.

Further Reading:

Dennett, D. C. and H. L. Dreyfus (1997) 'Artificial Intelligence', e-mail debate at: http://www.slate.com/id/3650/entry/23905/

Dreyfus, H. L. and Dreyfus, S. E. (1988) 'Making a Mind versus Modeling the Brain: Artificial Intelligence Back at a Branch Point', reprinted in *The Philosophy of Artificial Intelligence*, M. Boden (ed.), Oxford: Oxford University Press, 1990, pp. 309-333.

Haugeland, J. (1996) 'Body and World: A Review of What Computers Can't Do', *Artificial Intelligence*, vol. 80: 119-128.

WEEK 10: The Mathematical Mind

Main Reading:

Lucas, J. R. and Redhead, M. (2007) 'Truth and Provability', *British Journal for the Philosophy of Science*, vol. 58 (2):331-2.

Raatikainen, P. (2005) 'Truth and Provability: A Comment on Redhead', *British Journal for the Philosophy of Science*, vol. 56 (3):611-613.

Redhead, M. (2004) 'Mathematics and the Mind', *British Journal for the Philosophy of Science*, vol. 55 (4):731-737.

(NB: Read all of the above in chronological order)

Further Reading:

Lucas, J. R. (1961) 'Minds, Machines and Gödel', *Philosophy*, vol. 36:112-127.

Raatikainen, P. (2002) 'McCall's Gödelian Argument is Invalid', *Facta Philosophica*, vol. 4:167-69.

Schurz, G. (2002) 'McCall and Raatikainen on Mechanism and Incompleteness', *Facta Philosophica*, vol. 4:171-74.

WEEK 11: Embodied Cognition

Main Reading:

Clark, A. (2003) *Natural-Born Cyborgs*: *Minds, Technologies and the Future of Human Intelligence*, Oxford: Oxford University Press, ch. 3.

Further Reading:

Lakoff, G., and M. Johnson (1999) *Philosophy in the Flesh*: *The Embodied Mind and its Challenge to Western Thought*, New York: Basic Books.

Noë, A. (2004) *Action in Perception*, Cambridge MA: MIT Press.

Shapiro, L. (2007) 'The Embodied Cognition Research Programme', *Philosophy Compass*, vol. 2(2): 338-346.

WEEK 12: Consciousness

Main Reading:

Dennett, D. C. (1997) 'Consciousness in Human and Robot Minds' in *Cognition, Computation and Consciousness*, M. Ito, Y. Miyashita and E. T. Rolls (eds.), Oxford: Oxford University Press.

Further Reading:

Harnad, S. (2003) 'Can a Machine be Conscious? How?', *Journal of Consciousness Studies*, vol. 10 (4):67-75.

McDermott, D. (2007) 'Artificial Intelligence and Consciousness', in P. D. Zelazo, M. Moscovitch and E. Thompson (eds.) *The Cambridge Handbook of Consciousness*, Cambridge: Cambridge University Press, pp. 117–150.

Noë, A. (2009) *Out of our Heads*, New York: Hill and Wang.

WEEK 13: Representation

Main Reading:

Chalmers, D., R.M. French and D.R. Hofstadter (1992) 'High-level Perception, Representation, and Analogy: A critique of Artificial Intelligence Methodology', *Journal of Experimental and Theoretical Artificial Intelligence*, vol. 4 (3):185- 211.

Further Reading:

Brooks, R. (1991) 'Intelligence without Representation', *Artificial Intelligence*, vol. 47:139-159.

Clark, A. and J. Toribio (1994) 'Doing without Representing', *Synthese*, vol. 101 (3):401-31.

Van Gelder, T. (1995) 'What might cognition be if not computation?', *Journal of Philosophy*, vol. 92 (7):345-81.

WEEK 14: Expert Systems

Main Reading:

Gillies, D. (1996) *Artificial Intelligence and Scientific Method*, Oxford: Oxford University Press, ch. 2.

Further Reading:

Dennett, D. (1997) 'When HAL Kills, Who's to Blame? Computer Ethics', in *HAL's Legacy*: *2001's Computer as Dream and Reality*, D. G. Stork (ed.), Cambridge, MA: MIT Press.

Dreyfus, H. L. (1985) 'From Socrates to Expert Systems: The Limits and Dangers of Calculative Rationality', in C. Mitcham and A. Huning (eds.), *Philosophy and Technology II*: *Information Technology and Computers in Theory and Practice*, Dordrecht: Reidel, pp. 111-130.

Russell, S. and P. Norvig (2003) *Artificial Intelligence*: *A Modern Approach*, Englewood Cliffs, NJ: Prentice Hall.

WEEK 15: The Frame Problem

Main Reading:

Dennett, D. C. (1984) 'Cognitive Wheels: The Frame Problem of AI', reprinted in *The Philosophy of Artificial Intelligence*, M. Boden (ed.), Oxford: Oxford University Press, 1990, pp. 147-170.

Further Reading:

Fodor, J.A. (2000) *The Mind Doesn't Work That Way*, Cambridge, MA: MIT Press.

McCarthy, J. and P. Hayes (1969) 'Some Philosophical Problems from the Standpoint of Artificial Intelligence', in B. Meltzer and D. Michie (eds.), *Machine Intelligence*, vol. 4. Edinburgh: Edinburgh University Press.

Pylyshyn, Z. (1996) 'Introduction: The Frame Problem Blues. Once More, with Feeling', in K. M. Ford and Z. W. Pylyshyn (eds.), *The Robot's Dilemma Revisited*: *The Frame Problem in Artificial Intelligence*, Norwood, NJ: Ablex, pp. xi-xvii.

WEEK 16: Superhuman Intelligence

Main Reading:

Minsky, M. (1994) 'Will Robots Inherit the Earth?', Scientific American, Oct. 1994, http://web.media.mit.edu/~minsky/papers/sciam.inherit.txt

Further Reading:

Bostrom, N. (2006) 'How long before Superintelligence', *Linguistic and Philosophical Investigations*, vol. 5(1): 11-30.

Drexler, E. (1986) *Engines of Creation*, New York: Anchor Books.

Moravec, H. (1994) 'Robots Inherit Human Minds', talk found at: http://www.frc.ri.cmu.edu/~hpm/project.archive/general.articles/1995/RobotMind.talk.html