Müller, Vincent C. (2016), 'New developments in the philosophy of AI', in Vincent C. Müller (ed.), *Fundamental Issues of Artificial Intelligence* (Synthese Library; Berlin: Springer). – http://www.sophia.de - http://orcid.org/0000-0002-4144-4957

New developments in the philosophy of AI

Vincent C. Müller

Anatolia College/ACT
www.sophia.de

February 2015

Abstract: The philosophy of AI has seen some changes, in particular: 1) AI moves away from cognitive science, and 2) the long term risks of AI now appear to be a worthy concern. In this context, the classical central concerns – such as the relation of cognition and computation, embodiment, intelligence & rationality, and information – will regain urgency.

1.1. Getting interesting again?

We set the framework for this conference broadly by these questions: "What are the necessary conditions for artificial intelligence (if any); what are sufficient ones? What do these questions relate to the conditions for intelligence in humans and other natural agents? What are the ethical and societal problems that artificial intelligence raises, or will raise?" – thus far, this was fairly similar to the themes for the 2011 conference (Müller 2012, 2013).

This introduction is also a meditation on a remark by one of our keynote speakers, Daniel Dennett, who wrote on Twitter: "In Oxford for the AI conference. I plan to catch up on the latest developments. It's getting interesting again." (@danieldennett 19.09.2013, 11:05pm). If Dennett thinks "it's getting interesting" that is good news, and it is significant that he remarks that this interest appears *again*.

In the following year, the AAAI invited me to speak about "What's Hot in the Philosophy of AI?" (their title) – so the organization of AI researchers around the world also thinks it might be worthwhile to have a look at philosophy *again*. And indeed, one of the major topics in the AAAI plenary discussion was the social impact of AI; the president of AAAI has now made ethics an 'official' topic of concern (Ditterich and Horowitz 2015).

So, there are indications that 'philosophy of artificial intelligence' might have an impact, again. I think there are two major changes that make this the case: The changing relation to cognitive science and the increasing urgency of ethical concerns.

1.2. AI & CogSci – a difficult marriage

The traditional view of AI and Cognitive Science has been that they are two sides of the same coin, two efforts that require each other or even the same effort with two different methods. The typical view in the area of 'good old fashioned AI' (GOFAI, as Haugeland called it) until the 1980ies was that the empirical discipline of cognitive science finds how natural cognitive systems (particularly humans) work, while the engineering discipline of AI tests the hypotheses of cognitive science and uses them for progress in its production of artificial cognitive systems. This marriage was thus made on the basis of a philosophical analysis of joint assumptions – so philosophy served as the 'best man'.

This collaboration was made possible, or we at least facilitated, by the by classical 'machine functionalism, going back to (Putnam 1960) and nicely characterized by Churchland: "What unites them [the cognitive creatures] is that […] they are all computing the same, or some part of the same abstract <<sensory input, prior state>, <motor output, subsequent state>> *function*." (Churchland 2005: 333). If cognition is thus a computational process over symbolic representation (this thesis is often called 'computationalism') then computation can be discovered by cognitive science and then implemented by AI in an artificial computational system. This was typically complemented by a view of cognition as central 'control' of an agent that follows a structure of sense-model-plan-act; that rationally 'selects' an action, typically given some utility function. – All these components have been the target of powerful criticism over the years.

1.3. After GOFAI, "What's Hot in the Philosophy of AI?"

Two factors are different now from the way things looked only 10 or 20 years ago: a) We now have much more impressive technology, and b) we have a different cognitive science. The result is, or so I will argue, that we get a new theory of AI and new ethics of AI.

It currently looks like after the cold 'AI winter' in the 1980ies and 90ies we are already through a spring and staring a nice and warm summer with AI entering the mainstream of computing and AI products – even if much of this success does not carry the name of 'artificial intelligence' any more. This is a version of the well-known 'AI curse': in the formulation known as 'Larry Tesler's Theorem' (ca. 1970): "Intelligence is whatever machines haven't done yet." What is successful in AI takes a different name (e.g. 'machine learning'), and what is left for the old name are the currently impossible problems and the long–term visions.

A lot of classical AI problems are now solved and even thought trivial (e.g. real-life character recognition); robotics is now moving beyond the classic DDD problems (dirty, dangerous, dull). It appears that this is largely due to massively improved computing resources (processing speed and the ability to handle large data sets), the continued 'grind' forward towards better algorithms and a certain focus on feasible narrower problems. It does not seem to be due to massive new deep insight.

What does this mean for our marriage? Is there a divorce in the offing? The cognitive science side has largely learned to live with in separation – not quite a divorce but a more independent life. There is no more the assumption that cognition must be algorithmic (computational) symbol processing but rather a preference for broadly computational models. A strong emphasis on empirical work supports a tendency of cognitive science to undergo a metamorphosis from a multidisciplinary enterprise to another word for cognitive psychology. Cognitive science now involves embodied theories, dynamic theories, etc. – and it tends to find its own path now, not as adjunct to AI.

1.4. Ethics (big and small)

It has always been clear that AI, esp. higher level AI, will have an 'impact on society' (e.g. surveillance, jobs, weapons & war, care, …) and that some of that impact is undesirable, perhaps requiring policy interference. There is also the impact on the self-image of humans that makes AI, and especially robotics, have such a powerful impression on people who care a lot less about other new technologies. This is what I would call 'small ethics', the kind of ethics that concerns impact on a relatively small scale.

There is also 'big ethics' of AI that asks about a very large impact on society, and on the human kind. A discussion of this issue is relatively new in academic circles. Stuart Russell, one of our keynote speakers, had called it the question "what if we succeed?" (at IJCAI 2013) - (Bostrom 2014; Russell et al. 2015).

If the results of the paper by Bostrom and myself in this volume are to be believed, then experts estimate the probability of achieving 'high level machine intelligence' to go over 50% by 2040-2050, over 90% by 2075. Broadly, this will happen soon enough to think about it now, especially since 30% of the same experts think, that the outcome of achieving such machine intelligence will be 'bad' or 'very bad' for humanity.

I expect that this theme will create much discussion and interest, and that its speculation about what can be and what will be forces a return to the 'classical' themes of the philosophy of AI, including the relation of AI and cognitive science.

References

Bostrom, Nick (2014), *Superintelligence: Paths, dangers, strategies* (Oxford: Oxford University Press).

Churchland, Paul M. (2005), 'Functionalism at forty: A critical retrospective', *Journal of Philosophy,* (102/1), 33-50.

Ditterich, Tom and Horowitz, Eric 'Benefits and risks of artificial intelligence', *medium.com* <https://medium.com/@tdietterich/benefits-and-risks-of-artificial-intelligence-460d288cccf3%3E, accessed 23.01.2015.

Müller, Vincent C. (ed.), (2012), *Theory and philosophy of AI* (Minds and Machines, 22/2- Special volume: Springer).

— (ed.), (2013), *Theory and philosophy of artificial intelligence* (SAPERE, 5; Berlin: Springer).

Putnam, Hilary (1960), 'Minds and machines', *Mind, Language and Reality. Philosophical Papers, Volume II* (Cambridge: Cambridge University Press 1975), 362–85.

Russell, Stuart; Dewey, Daniel and Tegmark, Max (2015), 'Research priorities for robust and beneficial artificial intelligence', <http://futureoflife.org/static/data/documents/research_priorities.pdf%3E.