Two pieces on robots

By Bart Nooteboom

179. Moral robots?

published 05-01-2015

There is talk of robots developing beyond human cognitive capacity. There is fear that they will even come to dominate humans. In particular, the fear is that they will behave like sociopaths, with superior rationality but without human morality, without human values of empathy, compassion, and without practical wisdom of action.

If there is any ground to this fear, it is important to develop moral sense in the design and development of robots. How is that to be done? It requires, I propose, collaboration between engineers, software developers, cognitive and neural scientists, and philosophers.

But first, let us consider how robots could equal and surpass human cognition. That would require, among other things, a capability to develop tacit knowledge and associative, creative thought.

In the preceding item of this blog I discussed how cognition might develop in the brain, in *neural Darwinism*. I see no reason why that could not be emulated in robot brains. In fact, present robots already have been programmed to have an evolutionary learning capacity, where more or less random trials are reinforced or weakened in performance.

How about morality, then? I argued earlier that in humans moral instincts have evolved in a long process of biological evolution, with rival instincts of self-interest and altruism. For robotics to reproduce this, the development of robots, with some selection process in their functioning, would have to be speeded up enormously to match the long evolution of the human brain. Perhaps that can be done. But who defines and sets the selection conditions for survival of conduct? What if those are somehow set to *prevent* the selection of altruism?

In item 46 of this blog I argued that in human evolution an instinct for altruism might have developed from a benefit in group selection, under certain conditions, and that in-group loyalty probably arose at the price of out-group discrimination. Would this also have to be reproduced in the evolution of robots? And couldn't humans then be seen as the out-group, suffering all the more from robots?

Alternatively, could robots be programmed to act morally according to the multiple causality of moral action discussed previously? They would then have to be made to adequately perceive situations in a morally relevant way (material cause), to interpret their moral import and match them to moral principles (formal cause), taking into account situational conditions (conditional cause), depending on the agent and its position and role (efficient cause).

Would they also need to take into account their own interests (final cause), such as their own survival or 'health', not to self-destruct too easily?

And what moral principles would be programmed, according to what ethical system? Utilitarianism, Kantian duty ethics, or some form of virtue ethics?

Would there be exemplary robots for robots to imitate (exemplary cause)? Or could they learn to imitate their human teachers? Or could humans at some point learn from exemplary robots?

I argued earlier that in the exercise of practical wisdom intersubjective debate is needed, in *debatable ethics*, between different moral perspectives and assessments of situations, to fine-tune, moderate or revise moral perspectives. Would such debate need to occur between robots and humans, or robots among themselves? Or would robots help in debates between humans?

To sharpen their moral sense and become morally more adept, would it help to let robots read literature (see items 92, 120 of this blog)? Could they produce literature for humans to sharpen their moral sense?

231. Will robots have voice? published 13-12-2015

What will happen when robots take on more and more tasks, with increasing intelligence? What if a robot is opinionated, its views going against the established order, or against the will of its maker or owner?

Presently, an intellectual, scientist, or worker on a shop floor with contrary views cannot easily be silenced, in democracies. But robots may be simply switched off, or re-programmed to conform.

What will this do to people, if with regard to robots they no longer need to defend their views, and can bend the views of robots to their own? Would people then prefer to consort with robots, for the ease and comfort of it? Would that make them more self-involved, narcissistic even, turning robots into mirrors?

In this blog I argued that one needs the opposition form the other to detect one's own myopia, to nourish a flourishing life. This is needed, I argued, to achieve the highest form of freedom, which includes freedom from the bias of the self.

If robots are self-learning, by adapting their intelligence to what is successful, more rigorously and perfectly than humans, will this be a source of contrariness, defiance? People have a variety of sources of experience, in jobs, families, friendship, sports, travel and chance encounters, to feed their cognition and morality. Will robots have access to such diversity of experience? Will the owner of the robot, having invested in it, be willing to grant it unproductive time, in a range of private activities?

Next to his notions of 'exit' and 'voice', Albert Hirschman recognized the possibility of 'loyalty', which is acceptance, surrender to a faulty relationship.

Robots may undergo forced exit, being switched off, or may be programmed for loyalty. Will they be allowed to raise voice, or even be programmed for it? Or will they ever be self-generative enough to grasp voice, or even to impose loyalty? How moral will they be? And how would they acquire morality? I discussed that in item 179 of this blog.