Robots & Rights:
Will Artificial Intelligence Change The Meaning Of Human Rights?

PEOPLE POWER FOR THE THIRD MILLENNIUM: TECHNOLOGY, DEMOCRACY AND HUMAN RIGHTS :: SYMPOSIUM SERIES 2008

Symposium Report

Edited by
Matt James & Kyle Scott

BioCentre

Edited by:

Matt James
BioCentre, Associate Director

Kyle Scott
BioCentre, Project Administrator

BioCentre

51 Romney Street
London
SW1P 3RF

t: 020 7227 4706
e: info@bioethics.ac.uk
w: www.bioethics.ac.uk

Contents

Introduction

In 2006 the UK Office of Science and Innovation's Horizon Scanning Centre commissioned research into the future of robotics. The resulting paper, entitled "Utopian dream or rise of the machines?", examined the developments in artificial intelligence and how this may impact on law and politics. The report predicted a "monumental shift" that could occur should robots develop to the extent where they are able to reproduce, improve themselves or develop artificial intelligence.

With Bill Gates heralding 'a robot in every home' and a future in which robotic devices will become a part of our day-to-day lives, the 'Three Laws of Robotics' drawn up by Isaac Asimov could possibly shift from just being a good piece of science fiction to helping to shape the basis of future public policy. Is the future of robotics fact of fiction? What is artificial intelligence truly capable of producing? What are the challenges to humanity and how could this impinge on our human rights?

In response to these questions, on 15th January 2008 BioCentre hosted a symposium on 'Robots and Rights: Will artificial intelligence change the meaning of human rights?". This was the first symposium in the centre's 'People Power for the Third Millennium: Technology, Democracy and Human Rights' symposium series, held at the Royal Society of Medicine, London.

Professor Nigel M. de. S. Cameron, Executive Chairman of BioCentre, opened the symposium with some introductory remarks before Rachel Bell, Director of BioCentre, introduced the first speaker of the afternoon, Dr. Nick Bostrom. Dr. Bostrom is the Director of the Future of Humanity Institute, Faculty of Philosophy & James Martin 21st Century School, University of Oxford. Further presentations were made by Prof. Dr. Kerstin Dautenhahn, Professor of Artificial Intelligence, University of Hertfordshire and Prof. Steve Torrance, Professor of Cognitive Science, University of Middlesex and Visiting Senior Research Fellow at the Centre for Research in Cognitive Science, University of Sussex.

A question and answer session concluded the afternoon which allowed the audience to engage with the speakers on various issues arising from each of the presentations. Professor Nigel M. de S. Cameron returned to make some closing remarks on the afternoon before the event closed with a drinks reception.

Ethics of Artificial Minds

Dr. Nick Bostrom

Dr Nick Bostrom, gave an overview of different kinds of artificial intelligence, and the ethical issues that they would raise if they were ever created. He began by drawing a distinction between four kinds of artificial minds: the industrial robot, or domain specific AI algorithms, which is a kind of artificial intelligence that we find in society today; sentient or conscious artificial intelligence which we would consider to have moral status; artificial intelligence with unusual or strange properties; and finally super-intelligence. These four classes formed the basis of his talk and Dr Bostrom outlined how each class could raise its own unique set of moral questions.

Dr Bostrom pointed out that the ethical implications of artificial intelligence were the primary issues, over issues to do with public policy; he therefore geared his talk towards that side of the debate. The first class of artificial intelligence, the domain specific algorithm, are basically tools; it therefore raises no fundamentally new moral issues. As with any other tool there are issues surrounding the ways in which we use them and about who has responsibility when things go wrong. However, the tools themselves have no moral status, and similarly today, robots have no moral status. This means that if you do not like your robot you can hammer it to pieces, melt it down or change it into something different. You would not be harming the tool if you did any of these things. If it is wrong, then it is wrong for some other reason, such as it belongs to someone else, or you are polluting the environment by doing so, or it is somehow hurting another human being or animal. Nevertheless, there are a number of principles that should guide the use of and the development of domain specific AI algorithms. As such technology becomes more and more prevalent in society, these issues will become more important. There are problems that can arise from incorporating AI technology into the social environment. Dr Bostrom described a hypothetical situation where a bank used an advanced AI algorithm in order to check applications for mortgages, and a law suit had been brought against the bank because it is alleged that the bank has been discriminating racially by turning down a disproportionate number of black applicants. The bank could argue that no discrimination has taken place because the algorithm does not know the race of any of the candidates. Nevertheless, the statistics show that there is an unacceptably high

proportion of black applicants who are being turned down. Depending on how this algorithm has been constructed it may not be possible to identify what has gone wrong. It may be that the algorithm has detected that people who grew up in poor areas are less likely to pay back their mortgages or perhaps it is an accident in the coding that causes people with certain sorts of names to be rejected. It is possible that the algorithm would be so complex that you would never be able to detect the reason. A better design would be one that is transparent, so that someone could check exactly what it is that the algorithm is basing its decisions on. Another consideration would be as to how difficult it is to corrupt or manipulate the Artificial Intelligence. If such technology is to be widely used then we need to make sure that it cannot be easily fooled by someone else.

The subsequent categories lead us into new areas, and raise questions that have not been raised by technology before. If something has moral status, then that means that there are certain things that it would be wrong to do to that object in virtue of the thing itself and not due to its relation to anything else. We believe that all humans have moral status, and also that most animals have moral status. Most people would also recognise that there are degrees of moral status. Dr Bostrom said that the generally accepted view was that computers do not presently have moral status; however, it is possible that in the future they will. He then went on to consider how sentience might underpin moral status. Sentience is the capacity for phenomenal experience, or qualia, such as the capacity to feel pain or to suffer. If robots ever reached the cognitive ability and versatility level of a mouse or some other animal, then people would begin to ask whether they had also achieved sentience, and if they concluded that they had then they would have moral status as well. There would be certain things that it would be inappropriate to do, such as torture, or destroy it without good reason. If robots advanced further to the cognitive level of humans – developing the ability to reason and make long term plans – then we would also attribute moral agency to them, and they would be entitled to full human rights. Dr Bostrom outlined some principles for deciding when we should attribute rights to a being. He pointed out that over time we have changed our concepts about who has moral status, for example, we no longer assign moral status according to race or social status – we consider these factors to be irrelevant. What really matters is functionality, it is the cognitive experiences that you have and what you can do, and not what you are made of that decides your moral status. This means that moral principles that we have already developed for the human and animal domain can be transferred over to the domain of artificial intellect.

So far, the two groups looked at have not raised any fundamentally new moral questions, however, the third group would: artificial intelligence with strange properties. New questions would arise if sentience came apart from personhood. We believe that all healthy, normally functioning humans have sentience, however, it is possible that artificial intelligence could be functionally equivalent to a human but lack any inner experience, and so sentience and personhood would not necessarily always go together. Dr Bostrom pointed out that this was a controversial claim, but that it was worth considering the possibility. Such a situation would raise a number of interesting questions: Would a being without inner experience have the same moral status as one with inner experience? Would it matter how we treated such a

being if it could not feel pain? Other issues would arise due to ease of copying and other unusual properties that artificial intelligence may exhibit. If artificial intelligence was able to replicate itself millions of times very rapidly, then we might not think it is ok for such an individual to decide when to reproduce like we would with humans.

The final group that Dr Bostrom considered was artificial intelligence that achieved greater than human intelligence: super-intelligence. The issues with super-intelligence become not just how we should treat them, but how they treat us, due to their level of power. This means that we ought to build 'friendly' super-intelligence – one that does not harm us, or try to kill us. The problem of creating friendly artificial intelligence is much more important than the problem of creating super-intelligence because it is not like any other invention that we have ever made. Once this intelligence has come into existence it may not be possible for us to change it because it is so powerful. Therefore, we must first solve the problem of friendliness before we begin creating any super-intelligence. The problem of friendliness can be broken down into two components. Firstly, it should be stable, such that it continues to produce the desired effect even after iterated implementations and different contexts. Dr Bostrom gave two examples of scenarios where super-intelligence could create undesired consequences. Suppose that a computer was created with the aim of solving some very difficult mathematical problem, then this computer might decide to turn parts of the universe into a larger processor in order to give it more processing power with which to solve this problem. Another example was of a neural network that is programmed to maximise the number of smiling faces that it sees so that it will increase happiness. However, it may decide to pave the entire planet with molecular sized smiley faces in order to maximise this. Secondly, there is the issue of how we should define friendliness, and what it is that we want super-intelligence to do. Over time humanity has continued to redefine what it means by good, and new issues have come to light that have changed our understanding of it, so we would want friendliness to be sufficiently open to account for this. Also, there is the possibility that the way we have defined friendliness has unintended consequences. Dr Bostrom called this the genie in the bottle problem. He likened it to the person who has been given three wishes by a genie, but is concerned that the things they may wish for could have unforeseen and disastrous consequences.

Dr Bostrom concluded that although only the first class of artificial intelligence raised immediate practical problem, many of the possible future questions are so important that it is worth considering them now.

Dr. Nick Bostrom *is Director of the Future of Humanity Institute at Oxford University. He previously taught at Yale University in the Department of Philosophy and in the Yale Institute for Social and Policy Studies. He has more than 130 publications to his name, including three books. His writings have been translated into 16 different languages. Dr. Bostrom has a background in physics, computational neuroscience, and mathematical logic as well as analytic philosophy. Bostrom is a leading thinker on the consequences, ethical dimensions, and risks of anticipated future technologies and on "big picture questions" for humanity.*

Robots as Artificial Beings? A Human-Robot Interaction Viewpoint

Prof. Dr. Kerstin Dautenhahn

Professor Dautenhahn commenced her presentation by briefly outlining her work at the University of Hertfordshire and Adaptive Systems Research Group, trying to develop ways in which robots can behave intelligently. Within the discussion of robots and rights, Prof. Dautenhahn expressed the fact that she is more concerned with robots and people and how the two interact with one another. She perceives there is confusion over what robots are really capable of achieving. In her opinion, currently available robots are machines which can be fun, interesting and quite useful, yet nothing more than this.

Prof. Dautenhahn then proceeded to show some video footage of forty years ago which advertised the robotic housemaid and then contrasted this with a recent video clip of the work currently being carried out at the University of Hertfordshire on robotic interaction. Having set the scene, the professor made three key points concerning the subject of 'social robots'. Firstly, much of the work Prof. Dautenhahn is engaged with is focused on the robot as a companion which helps around the home and assisting those that are housebound. The professor acknowledged the fact that the notion of the robot companion has been around for a long time. Consequently, the question is often asked as to why technology has not achieved this given much of the advances in technology. In response, she pointed out that the concept of the robot companion involves multi-tasking as opposed to just one task; something that to date has been the primary focus of robotic developments. Therefore the emergence of the true robotic companion is still to be achieved.

Secondly, she made the point that there is often confusion and ambiguity surrounding the use of the terms 'social' and 'robot' being used together. Prof. Dautenhahn made reference to over twenty different definitions as to what this could mean from within the literature in this area. She highlighted the two ends of the spectrum in this regard. At one of the spectrum, some would call a robot social 'evocative' in the way that it evokes people's tendency to express empathy towards it regardless of what it does or whether it moves or not. In other words, they may respond to it by saying how cute it is or that they would really love to have one of their own. At the other end of the spectrum, there is the artificial intelligence (A.I.) perspective

where people are concerned with the issue of just how much social intelligence and competency a robot should have. Thirdly, within the literature emphasis may be placed on one of two aspects, either the A. I. (or robot cognition) centred view or the robot centred view. Those who take the A. I. perspective are concerned with how the perfect A.I. architecture can be designed and built. On the other hand, the robot centred view places the robot at the very centre of enquiry and considers it as a creature on its own. The robot is believed to have a human-like system consisting of moods, desires and emotions which need to be fulfilled. Prof. Dautenhahn noted that this was something which the first presentation of the afternoon addressed more specifically. Finally, there is the human-centred viewpoint which does not look at the robot in isolation but rather within the social context. Humans are invited to interact with the robot to participate in and evaluate the design of these systems. Given these three points, Prof. Dautenhahn illustrated the fact that the term "social robots" can mean different things in different contexts.

To further illustrate this concept of difference in interaction, she cited the example of a robot cleaning a sewage system and a robot being used in a therapy context. The robot cleaning the sewers involves no interaction; it is simply fulfilling an automated task. However, the robot used in a therapy context depends heavily on interaction, requiring it to understand and response to a variety of situations and stimuli as opposed to just one. The context determines the meaning of the interaction essentially. The professor proceeded to talk about the frustrations which can result as a result of human-robot interactions. When humans are interacting with robots which perform the same task every day (for example asking you whether you wanted a cup of tea), once the novelty of having a robot do this for you has worn off, the repetitive cycle will inevitably become annoying and frustrating. Therefore, there is the need for robots to be programmed with a lot more than just programmed social behaviour. Robots need to understand social norms, politeness and manners. This led Prof. Dautenhahn to briefly address the question of what kind of social interaction behaviour is required of robots before addressing her main area of specialisation that of the practical perspective of robots and their use, especially with regard to humanoids.

Humanoids have some human like characteristics for example their eyelids, arms and limbs can move. Drawing upon the work and experience of the Adaptive Systems Research Group, she spoke of the example of KASPER. The team adopted a minimal design for KASPER which incorporated some human elements but nevertheless remained robot-like in appearance. Dautenhahn pointed out that although the research team gave it certain features so human could interact with it (such as eyes, a mouth and arms), the fact that KASPER could smile was due to the research staff making it smile. Consequently, KASPER is not a sentient being. This then led to the professor to discuss the issue of attraction between humans and robots. There is something different between a human liking another human and a human liking or expressing a love of say KASPER, or a laptop, car or other machine. The fact that humans treat certain machines and other interactive technologies in a certain way from a social perspective, tells us more about the nature of a sentient being than it does about the machine such as KASPER. Just because people treat and engage with KASPER on a social basis does not make KASPER social. The professor expressed her opinion that

robots would not achieve any level of "true" sentience at any point in the near future.

Continuing on this theme of relationships and interaction with robots, Prof. Dautenhahn conjectured that should developments continue robots could well end up caring for vulnerable people (such as elderly and those with special needs) in the future. Such advances present many benefits as well as many problems, which are largely centred around the principle question of what kind of relationship they will form with robots. One of the primary concerns which Prof. Dautenhahn expressed was concerning the reciprocation of social interaction. In the context of a robot-human relationship, it is believed that robots can give back the same as another human would be able to in a typical human-human relationship. This is something that in Dautenhahn's opinion simply cannot happen and which is unachievable.

Turning to addressing the development of humanoids more broadly, over the years many have been fascinated with the idea of building robots with human-like qualities. In relation to this, Dautenhahn referred to the work of the Japanese in this area and the idea of the "Uncanny valley". This is an idea created by the Japanese which refers to the time when in the development of humanoids they become so human-like that humans increasingly cannot tell the difference and distinction between the two is blurred. In Prof. Dautenhahn's view, currently robots are still very much on the mechanistic level of appearance. However, Japanese work in this area is pushing the boundaries in this respect, perceived very much to be a model for reaching cognitive development of robots. In the professor's opinion, this is concerning since at present this work is reaching the 'uncanny valley' stage of development.

Seeking to expand upon these concerns, Prof. Dautenhahn conjectured that robots are relational artefacts. Machines can pretend that they can have a relationship with you by pretending to understand what you say and do. However, the fact that you can speak to a robot and it replies "I understand" has more to do with a customized set of responses rather than the robot's ability to listen and respond accordingly (in a human-like manner) to what you have just said to it. This is where the real danger lies; that people think they have been understood by the robot and have entered into a meaningful relationship with it. Proceeding with this line of argument, the questions arises as to whether or not the way in which humans view human-human relationships will be adversely affected, for example, should the facility become available whereby you can pause or stop the robot when you no longer want to interact with it? Will it change and alter what you want to come from a relationship with a fellow human? Despite the fact that many of us have been in situations where we wished humans had the facility, the fact remains humans do not have a pause button! This therefore begins to question the value of interactions in relationships and whilst Dautenhahn and others have no specific answers to these questions at present, they are nevertheless concerned with what the possible consequences could lead to. What does it mean in the long term that humans are interacting with machines who give the illusion they understand but in reality do not?

Proceeding to talk on human nature and the social nature, Prof Dautenhahn made reference to the work of Alison Jolly and the fiftieth anniversary of the social intelligence hypothesis. The hypothesis points to the origins of the primate nature

as being social. In the first instance, primates become social beings and become intelligent by realising skills which are gained through being social. Stemming from this work, Dautenhahn and her team at the University of Hertfordshire have developed the companion paradigm in pursuing their work with robots and care for people. They have sought to explore how a robot can be considerate, taught manners and be unobtrusive. In their work with IROMEC, they have developed robot tools for children with autism with the robot fulfilling the role of a mediator. To begin with the research team took the idea that the robot attracts the attention of the children and stimulates interest. The robot is then placed in a social context where both autistic and other children are present. The work focuses on how the robot can relate to the child, bearing in mind that children with autism have specific impairments in their communication and relationships with others, making social interaction very difficult. Therefore, the team seeks to exploit the fact that they are like any other child and allows them to interact with the robot. Naturally, this requires extensive work and research in being able to develop a robot which is able to interact successfully. The starting point is that children with autism can find social environments very overwhelming but that robots can help to create a safe, predictable and enjoyable environment for them to engage in. This is further aided by an emphasis on play which is a crucial part of the process as it plays a significant role in intellectual, social and emotional development. What is more, when working with young children it is important that they enjoy the process as well. Having established the fact that they are enjoying the process and consider it fun, the next step in the programme's development is to increase the educational and therapeutic aspects of the robot's interaction with the child.

From an experimental perspective, this kind of work with robots and children necessitates certain parameters being put in place. There must be safety procedures in place with regard to the use and operation of the robot as well as ethical issues surrounding the participation of the children. Prof. Dautenhahn noted the success of the work had come through collaboration with schools and working with small groups of children noting their likes and dislikes, trialling different designs of robots with them and ensuring that they learn and take something away with them from being part of the experimental process. Essentially, the approach adopted amounts to an experimental model based on long term case studies. On the issue of interaction, Prof. Dautenhahn noted that it had been important to compare the robots with other toys to ensure the distinction between the two. Children interact differently with the robot as opposed to children simply playing with toys. Furthermore, due to the complexity of learning and 'reading' all the subtle varieties of human expression (something which autistic children find difficult), KASPER had been designed with minimal expressions so as to make interaction between the child and the robot much easier. Based upon this work, Prof. Dautenhahn reiterated her belief that robots can be very useful when used in very specific areas. However, interaction with robots remains mechanical, unless the interaction takes place within a context which provides the meaning. A final video clip was shown relating to research work into robot companions and the use of robots in the home. The video showed work being carried out into how robots should approach people, interpret visual cues and measure distances.

In conclusion, Professor Dautenhahn pre-

sented some issues which in her opinion are worthy of further consideration in relation to the development of robots. She highlighted the need to think through the practical issues of human-robot interactions. If robots are going to be in the home there are confidentiality issues which need to be considered. Will humans be comfortable with robots being privy to arguments and disagreements which take place within the home? Will humans be happy for robots to hear what is said 'privately' on the telephone? If humans are uncomfortable with these ideas *now*, then research needs to be conducted *now* which will hopefully resolve these concerns as opposed to in the future when it might be too late. On the issue of human-robot interaction, Prof. Dautenhahn reiterated her concerns as to how truly rewarding and satisfying such interactions are and how authentic they can be in comparison to the more typical human-human interactions. Could the robot be tricking you into thinking it understands you when in reality it does not? The aspects of love and trust in any human-human relationships are vital so what will come of them in our interactions with robots? All in all, Prof. Dautenhahn concluded that robots are machines which in turn have the potential of being useful tools. However, anything beyond this tends to be more fiction rather than fact.

Prof. Dr. Kerstin Dautenhahn *is Research Professor in the School of Computer Science at University of Hertfordshire in U.K. where she coordinates the Adaptive Systems Research Group. She received her Ph.D. degree from the Biological Cybernetics Department of the University of Bielefeld, Bielefeld, Germany, in 1993. She has published more than 150 research articles on social robotics, robot learning, human-robot interaction and assistive technology. Human-Robot Interaction studies are one of her current key research interests.*

Robot Ethics: Fantasy or Necessity?

Prof. Steve Torrance

Professor Steve Torrance spoke about the necessity of some sort of robot ethics due to the inevitability of robotics becoming more and more common in our society.

Prof Torrance began by outlining two views of robot ethics: one is based on the view that robots will never achieve any kind of genuine moral agency, whereas the other view is that robots will soon become so widespread that a new ethics for them will be required. He also said that he believed that these two views were not incompatible, and that he hoped this would become apparent during his talk.

Prof Torrance then drew a distinction between deep and shallow robotics. He described deep robotics as being more academic and less practical, more concerned with possible advances in robotics in the future; whereas, shallow robotics is concerned with current robotics and the practical issues that these raise. Both these approaches take a human centred view of robotics, whereas Prof Torrance advocated taking a look at a robot centred view of robotics to consider under what circumstances we would have responsibilities towards robots and they might have responsibilities towards us, and each other. He pointed out that a human centred stance usually stems from an instrumental view of robots; and a robot centred stance from an intrinsic view, however, he thought that the boundary between these was a rather blurry indistinct one.

There are two key aspects of robot centred ethics that Prof Torrance commented on. One was of robots as moral producers, and the other was of robots as moral consumers. Viewing robots as moral producers means viewing them as having duties or obligations; viewing them as being subject to some sort of moral appraisal. The other way of viewing robots as moral agents is to see them as moral consumers, having needs or interests. This would mean that they will generate moral responsibilities in us, or in other robots. Moral production and consumption do not go hand in hand, for example, we would consider chickens to be moral consumers because we are arguably obliged not to hurt them unnecessarily, whereas they are not moral producers. Some robots on the other hand could fulfil the conditions for moral production but not those for moral consumption.

Two possible properties that could lead to moral agency were described by Prof Torrance: Autonomy and consciousness. There are at least three kinds of autonomy. The first is functional autonomy and it is connected with how much input is required from other sources in order for the robot to function, so just as a car is more functionally autonomous than a wagon because it does not require a horse, a self navigating car would be even more functionally autonomous because it did not require a person. Organic autonomy is the second kind, and it means not being reliant on another agent for maintenance or continued exist. The final kind of autonomy is moral autonomy, and this refers to the ability to make moral decisions. Some people might be tempted to see moral autonomy as naturally arising out of functional autonomy but they are conceptually distinct. This is particularly important, for instance in discussions of battlefield robots, where increasing functional autonomy of such robots may tempt people to assume that they can be held to be morally autonomous and therefore let their human makers and controllers 'morally off the hook'.

Consciousness is often thought to be a crucial element in the discussion of robot ethics because it is thought that this is a necessary requirement for being a moral agent. Very few people believe that robots will ever achieve consciousness, and so robots-as-moral-agents is often not taken seriously. This is because most people associate being a moral agent with being able to empathise with others, and this seems to require some sort of consciousness. However, there has been attempts to program ethical robots that do not rely on consciousness at all, for example, by programming them with deontic logic systems that cause such robots to act in ethical ways despite the lack of any kind of consciousness or sentience.

Prof Torrance then argued that there are possible cases where a robot centred ethics could come about even without consciousness. He gave the example of 'The Bicentennial Man' by Isaac Asimov where a robot gains the right to accumulate wealth because it is able to make and sell objects. If such a situation were to come about then there would be legal obligations surrounding such a robot. For example, it would be illegal for such a robot to acquire wealth through illegal means, and it would be illegal to steal from a robot. If we had a situation where robots had legal obligations and we had legal obligations to them then this would also generate moral obligations.

In conclusion, Prof Torrance claimed that true moral agency for robots would not be achievable unless a new approach was taken to robotics; perhaps a more biological approach. However, there are other ways that a robot centred ethics could still be required. This means that there is a very real need for further research into robot ethics that will direct the incorporation of ethics into functionally autonomous agents.

Prof. Steve Torrance *has researched in the philosophy of AI and cognitive science over two and a half decades. He has published many articles and collections of papers on issues concerning the conceptual and ethical foundations of artificial personhood. He is Professor Emeritus in Cognitive Science at Middlesex University, and a visiting Senior Research Fellow in the Philosophy of Artificial Intelligence and Cognitive Science group at the University of Sussex. He holds a doctorate in moral philosophy from the University of Oxford.*