Brian R. Duffy:

Fundamental Issues in Social Robotics

Abstract:

Man and machine are rife with fundamental differences. Formal research in artificial intelligence and robotics has for half a century aimed to cross this divide, whether from the perspective of understanding man by building models, or building machines which could be as intelligent and versatile as humans. Inevitably, our sources of inspiration come from what exists around us, but to what extent should a machine's conception be sourced from such biological references as ourselves? Machines designed to be capable of explicit social interaction with people necessitates employing the human frame of reference to a certain extent. However, there is also a fear that once this man-machine boundary is crossed that machines will cause the extinction of mankind. The following paper briefly discusses a number of fundamental distinctions between humans and machines in the field of social robotics, and situating these issues with a view to understanding how to address them.

Agenda

Author:

Dr. Brian R. Duffy:

- SmartLab Digital Media Institute, University of East London, Docklands Campus, 4-6 University Way, London E16 2RD
- ✉ brd@media.mit.edu, 💻 http://www.manmachine.org/brd/, http://www.smartlab.uk.com/
- Relevant publications:
  - B.R. Duffy, G. Joue, "The Paradox of Social Robotics: A Discussion", AAAI Fall 2005 Symposium on Machine Ethics, November 3-6, 2005, Hyatt Regency Crystal City, Arlington, Virginia
  - Duffy, B.R., Joue, G., I, Robot Being, Intelligent Autonomous Systems Conference (IAS8) 10-13 March 2004, The Grand Hotel, Amsterdam, The Netherlands
  - Duffy, B.R., " Anthropomorphism and The Social Robot", Special Issue on Socially Interactive Robots, Robotics and Autonomous Systems 42 (3-4), 31 March 2003, pp170-190

ISSN 1614-1687

Brian R. Duffy:

Fundamental Issues in Social Robotics

Introduction

Man and machine are rife with fundamental differences. Formal research in artificial intelligence and robotics has for half a century aimed to cross this divide, whether from the perspective of understanding man by building models, or building machines which could be as intelligent and versatile as humans. Inevitably, our sources of inspiration come from what exists around us, but to what extent should a machine's conception be sourced from such biological references as ourselves? Machines designed to be capable of explicit social interaction with people necessitates employing the human frame of reference to a certain extent. However, there is also a fear that once this man-machine boundary is crossed that machines will cause the extinction of mankind. The following paper briefly discusses a number of fundamental distinctions between humans and machines in the field of social robotics, and situating these issues with a view to understanding how to address them.

The Body Dilemma: Biological vs. Mechanistic

The fundamental difference between man and machine[1] is that of existence. Maturana and Varela [1] differentiate between the issue of animal systems versus mechanical systems by concentrating on the organisation of matter in systems (see also [2][3]) via the concepts of *autopoiesis* and *allopoiesis.* In essence this constitutes the fundamental distinction between natural systems embodiment and an artificial intelligence perspective of embodiment. *Autopoiesis* means self- (auto) –creating, –making, or –producing (poiesis). Animal systems adapt to their environment at both macro (behavioural) and micro (cellular) levels and are therefore autopoietic systems. Mechanical systems on the other hand primarily adapt at a behavioural level (with highly constrained physical adaptivity capabilities relative to natural systems) and are allopoietic.

Similarly, Sharkey and Ziemke highlight in [2], "[l]iving systems are not the same as machines made by humans as some of the mechanistic theories would suggest". The fundamental difference lies in terms of the organisation of the components. Autopoietic systems are capable of self-reproduction. The components of a natural system can grow and evolve, ultimately growing from a single cell or the mating of two cells. In such systems, the processes of system development and evolution specify the machine as a whole.

Allopoietic systems are, on the other hand, a concatenation of processes. Its constituent parts are produced relatively independent of the organisation of the machine. In such systems, the processes of producing/manufacturing each of the individual components of the system and the hard constraints in their integration define the machine and its limitations. This fundamental difference, in the context of artificial intelligence, has been highlighted in [2] where the notion of evolvable hardware is discussed. The designer of a robot is constrained by such issues as the physical and chemical properties of the materials used, by the limitations of existing design techniques and methodologies. The introduction of evolvable hardware could also help overcome to a certain extent, the inherent global limitations of the robot end product by facilitating adaptation and learning capabilities at a hardware level rather than only at a software level. This adaptability is often taken for granted in biological systems and likewise often ignored when dealing with such issues as robustness, survivability, and fault tolerance in robotic systems. Sharkey and Ziemke highlight the lack of evolvable capabilities in allopoietic systems as being directly related to its lack of autonomy. Unlike allopoietic systems, biological or autopoietic systems *are* fully autonomous[2].

[1] It is important to note that this paper discusses physical humanoid robotic systems. Purely virtual representations of robots including avatars are not considered in this work due to their constrained environmental integration in our physical world. The hard issues of sensor and actuator complexity in physical social environments are important aspects of the ideas discussed here.

[2] While there is considerable discussion over the meaning of the term autonomy, it is here used in the context of physical systems existing in real world unstructured environments. Autonomy refers to a system's ability to function independent of external control mechanisms.

ISSN 1614-1687

While the fields of evolutionary and bio-inspired robotics look to bridge the gap between natural and artificial systems, the fact that they still fundamentally involve the concatenation of digital processes, both at a hardware and software level, does not bridge the divide between allopoietic and autopoietic systems. The practical reality is that in order to realise a physical robotic system, a collection of actuators, sensors and associated control mechanisms must be integrated in some way. Their integration can simply not equal that of biological systems or else it would *be* a biological system and hence not an artificial machine[3].

While technological innovation and development will increase the resolution of the machine artefact in its behavioural and aesthetic similarities, it will fundamentally remain a machine. Given this most fundamental difference, the resemblance between a human and a machine will remain an analogy. However, should the intelligent social machine be constrained to resemble man in the first place?

Anthropomorphism: Balancing Function & Form

Unquestioningly, the holy grail for roboticists has been to realise a humanoid robot that is indistinguishable from ourselves. Intuitively, social robots may seem more socially acceptable if they are built in our own image. However, this should not necessarily imply that they must be indistinguishable from us. As yet, given the limitations of the state of the art in social robotics, we can easily feel more comfortable with a cartoon-like appearance than imperfect realism [4]. While anticipating that future technologies may allow us to achieve more humanlike function and form, is this really the ultimate design goal that we would like to achieve? From a technological standpoint, can building a mechanistic digital synthetic version of man be anything less than a cheat when man is not mechanistic, digital nor synthetic?

Our propensity to anthropomorphise and project humanness onto entities that may bear only the slightest resemblance to ourselves is well known [5]. Thus a successfully designed social robot may be one that maximises both its mechanical advantages and the minimum humanlike aspects required for their social acceptance. Our future interaction with robots will undoubtedly use alternate features than those we are currently familiar with in our interactions with people. From a robots perspective, its ability to garner bio information and use sensor fusion in order to augment its diagnosis of the human's emotional state to facilitate interaction through for example, a "techno handshake" and an infra-red vision system, illustrates how a machine can engage people in their social space without necessarily employing human-like frames of reference for sensing technologies. This strategy can effectively increase people's perception of the social robot's "emotional intelligence" without feeling alienated by it, and consequently its improved social integration. This also relies on the machine not garnering nor utilising knowledge about a person's emotional state that would generally be hidden from people they interact with, such as excitation states resulting in an increased heart rate. Social interaction involves as much our control of our perceivable emotional states as the expression of these states.

There is an interesting issue where the mechanical robotic system's "understanding" of the social situation and the consequent development of its social interaction with people is based on allopoietic mechanisms. The bridging of the digital divide between the biological and the artificial takes on a new dimension other than the issues more generally discussed in the literature regarding physical embodiment. The social embodiment of the robot effectively creates the illusion of bridging the real-vs.-artificial divide and is discussed in the following section.

The Power of the Fake

Intentionality, consciousness, and free will are important traits associated with human-kind. We have continually posed the question whether it would be possible to realise such properties in a machine. While progress to date has been impressive, few would argue that we are much closer to understanding these notions well enough to be able to artificially recreate them in a machine in some way. With the advent of the social machine, and particularly the social robot, where the aim is to develop an artificial system capable of socially engaging people according to standard social mechanisms (speech, gestures, affective mechanisms), the *perception* as to whether the machine has intentionality, consciousness and free-will will change. From a social interaction perspective, it becomes less of an issue whether the machine

[3] The merging of the biological and the artificial in the form of cyber-organisms is not discussed here.

actually has these properties and more of an issue as to whether it *appears* to have them. If the fake is good enough, we can effectively perceive that they do have intentionality, consciousness and free-will.

Our assessment of whether you or I are intelligent is generally based on our social interaction (assuming that one does not always have access to the intelligence metrics of IQ and EQ tests, which are also a source of controversy). Social robots become the important step in us coming to the conclusion that machines may *observably* possess intelligence and emotional capabilities. Whether they are in fact genuinely intelligent according to the human frame of reference becomes less of an issue (a measure of human intelligence is based on observation). Based on our social interaction with them, the associated communication mechanisms facilitate such a conclusion. They simply speak our language.

There still remains the fact that while the machine may have a different kind of energy coursing through its circuitry where we would clearly not classify it as alive, it is definitely ON.

The Decision Dilemma

Social interaction between people is a very complex problem, something that requires all our capacities in order to be able, on the whole, to succeed. Within this domain are mountains of complex social and physical data with intertwining contexts and conclusions. In the not too distant future, a dilemma will face man in how to negotiate the role of integrating social machines into our society, whether one supports the idea or not. Artificial reasoning mechanisms have demonstrated a strong capacity to navigate vast quantities of data and, through employing logic-based reasoning mechanisms, extract key features. IBM's Deep Blue has shown how a machine can very rapidly search areas of a large defined state space (approximately 10^43 legal positions), and has famously defeated the world champion chess grand master in a number of games. While it is possible that a machine could make a more informed and even better decision than a human *in certain situations*, it is important that this is not taken out of context. A machine *may* be able to make faster better decisions but only when it has enough accurate structured information about the problem. A key trait in human reasoning is the ability to make good decisions with incomplete information. This fundamental distinction is important.

As quoted from Arthur Conan Doyle's Sherlock Holmes, "[w]hen you have eliminated all which is impossible, then whatever remains, however improbable, must be the truth". While a person's cultural background has been shown to influence their preference for formal vs. intuitive reasoning [6], machines are fundamentally grounded in logical symbolic manipulation according to defined structures. While neural networks, fuzzy logic and others have looked to implement our ability to reason with incomplete information through such mechanisms as learning algorithms, pattern matching and statistics for example, their usability and robustness is dependent on the quality of the training data and is by no means "perfect". The source of such data, particularly if recovered through current sensor technologies (with inherent noise and error issues), provides additional error dimensions.

The use of logical reasoning plays an important role in how a machine can reach a conclusion given a set of premises. While often counter-intuitive (e.g. the birthday paradox: if there are 23 people in a room, there is a chance of more than 50% that at least two of them will have the same birthday yet this defies our common sense – see [7] for a comprehensive list of similar paradoxes), the process of formal logic and its conclusions is very difficult to refute. A machine may be more equipped than a human to make a decision that requires the processing of a massive volume of data, with its abilities may lead to a better solution. Recent natural disasters have necessitated the negotiation and coordination of major logistical efforts, a problem domain where a machine may be best equipped to navigate such complex problems. An interesting problem arises when, in a major public domain, a machine could provably make a better decision than man. One should just not forget the role of instincts in human survival. Will it become difficult to justify choosing the decision of a machine over the instincts of man if they are conflicting? Can we entrust moral decisions to a machine?

Moral Rights and Duties

The issue of moral rights and duties arises from two perspectives. The first is whether a machine should be programmed to be morally capable of assessing its actions within the context of its interaction with people (this includes the evolution of behavioural mechanisms and associated moral "values"). This involves defining, in a similar vein to developing the allowable behaviours of the system, the limitations of its actions through defining forms of *anti-*

behaviours, behaviours that are not to be realised. Serious issues of complexity arise from core embodiment issues such as sensor noise and the associated accuracy issues in environmental modelling, and inter-behaviour interference. The concept of bounded rationality argues against the capability of a system of being sufficiently aware of all the environmental implications of its actions either before it undertakes them, or after it has performed them.

The second perspective is whether it is necessary to have human capabilities in order to be able to assess morality. This also involves the notion of whether a human *perceives* the machine to have moral rights and duties, and incorporates the aesthetic of the machine (see [8] for human social perception studies based on attractiveness). Employing the human frame of reference for aesthetic and behavioural features loads a human's expectations of the robot having a degree of such human-centric values. An important issue also arises as to whether it is morally acceptable to build social robots, whether certain robots can be built but not others. Military and security robotics already pose this problem when they are designed for human occupied environments. Their interaction with people is inevitable, and is invariably negotiated at the programming and development stages of their construction. This remains a difficult issue as the link between technology and warfare is an age-old debate.

The success of employing the anthropomorphic metaphor is in fact grounded on maintaining human-centric expectations, whether being moral or immoral. If the robot is perceived as simply a functional machine and nothing more, the issue of robot morals and duties is irrelevant. It is just a question of whether it is programmed correctly and safely or not. If the robot draws on human-like features and behaviours to explicitly develop social interaction with humans, then moral rights and duties become part of the set of expectations associated with our interaction with something that looks to use our frame of reference, humanness.

If the form of the robot is highly human-centric where it has the potential to integrate itself too much in our social circle, is it right to build such robots that basically are "cheating" people to believing in their perceived humanness? This involves a betraying of our trust. The problem becomes even more complex if we consider the relationship that is important and not necessarily the physical robot itself. If the social machine employs contrived notions of humanness and helps a patient through difficult times by listening and "understanding" their problems, is that allowed? Does the end justify the means? These dilemmas are not new. The role of a pet and their status in some people's lives poses similar problems.

Conclusion

While the development of intelligent affective human-like robots will raise interesting issues about the taxonomic legitimacy of the classification *human*, the question of whether machines will ever approach human capabilities persist. Technology is now providing robust solutions to the mechanistic problems that have constrained robot development thus far, thereby allowing robots to permeate all areas of society from work to personal and leisure spaces. As robots become more integrated into our society, unresolved ethical issues of its existence and design become more imminent. Will, for example, the idea of introducing a subservient human-like entity into society rekindle debates on slavery? Age old problems and conundrums should help us negotiate some of the problems that will arise. The fact that it could be an autonomous reasoning human-like *machine* will add a new dimension to the problem.

The sensationalist perspective of machines taking over the world in the future tends to ignore the points raised in this paper. The key is to take advantage of these reasoning machines and their capabilities rather than constrain them [9]. By allowing machines become a very different form of "species", rather than constraining it too much to the human frame of reference, we can profit from its inherent capabilities as a machine without trying too hard to cross, and inevitably fall into, the chasm that separates man and machine. This also constitutes a step in avoiding a number of the ethical issues that the domain is in the process of introducing. It being a machine is not a flaw, it's a role.

References

[1] Maturana, H.R., Varela, F.J., "Autopoiesis and cognition – The realization of the living", D. Reidel Publishing, Dordrecht, Holland, 1980.

[2] Sharkey, N., Ziemke, T., "Life, mind and robots: The ins and outs of embodied cognition", Symbolic and Neural Net Hybrids, S. Wermter & R. Sun (eds), MIT Press, 2000

[3] Duffy, B.R., Joue, G. "Embodied Mobile Robots", 1st International Conference on Autonomous Minirobots for Research and Edutainment - AMiRE2001, Paderborn, Germany, October 22-25, 2001

[4] B.R. Duffy, Anthropomorphism and The Social Robot, Robotics & Autonomous Systems: Special Issue on Socially Interactive Robots, 42 (3-4), 31 Mar (2003), p170-190

[5] Reeves, B., & Nass, C., The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places, Cambridge University Press, 1996

[6] Norenzayan, A., Smith, E.E., Kim, B. J. & Nisbett, R. E. Cultural preferences for formal versus intuitive reasoning. (In press). Cognitive Science

[7] Clarke, Michael (2002). Paradoxes from A to Z. London: Routledge

[8] M. Alicke, R. Smith, M. Klotz, Judgements of physical attractiveness: the role of faces and bodies, Personality and Social Psychology Bulletin 12 (4) (1986) 381–389.

[9] Duffy, B.R., O'Hare, G.M.P., Martin, A.N., Bradley, J.F., Schön, B., "Future Reasoning Machines: Mind & Body", Kybernetes Journal, Vol.34, 2005