THE SOCIAL ROBOT PARADOX

Brian R. Duffy

Department of Computer Science,
University College Dublin, Belfield, Dublin 8, Ireland

Brian.Duffy@ucd.ie

For centuries, the idea of the robot (or automota as it was generally known) has always been intriguing. The possibility of bringing an artificial device created by man to "life" continues to drive the imagination, from films and novels, to engineering and scientific laboratories. Core to the original coining of the term "robot" (from Karl Capek's 1923 play "Rossum's Universal Robots") is a strong association with the human form, its appearance, its functionality, and even its failings. Although the term robot now includes the cold functionality of a car assembly machine, the intrigue as to whether man has the capacity to create artificial life remains.

The origins of robots or automata date back millennia. One theme that has always prevailed is the use of the human as the frame of reference in its design. The humanoid is seen as representing the pinnacle of robot development, and clearly the most complex. Whether explicit or not, employing the human form inherently defines the robot as having some degree of social functionality.

In fact, humanoid robots outside of science fiction, have thus far only been toys or research platforms with nebulous applications. Even Sony's SDR4X (QRIO) has been presented in recent international conferences as a toy, albeit an expensive one. Even throughout the millennia of robot history, the humanoid has continuously been demonstrated to entertain and amuse. It is intriguing to note that what is effectively seen as one of the most powerful paradigms for adaptivity and flexibility, the human, has, when modelled in the form of a machine, resulted in little more than a toy. Its usefulness is very limited. The social robot paradox begs the question that if non-humanoid robots have been so successful in industry as functional machines and humanoids on the other hand have only resulted in elaborate toys, then why build a social robot?

1 What happens when you give a robot a name?

Envisage the following two scenarios.

> **Scenario 1**: You encounter a blue box (figure 1) which as a camera (vision system), speaker and microphone (speech synthesis and recognition), and a red light. The light on the blue box turns on. It speaks. It says "Hello. How are you?"
> What is your reaction?

Figure 1: The Intelligent BlueBox

Scenario 2: You encounter a robot, with a head, two arms and two legs (figure 2). It is similarly equipped with a camera-based vision system, the ability to talk and listen through a microphone and speaker. The robot starts to move. It turns and looks straight at you and says “Hello. How are you?”

Figure 1: The Intelligent JoeRobot

The differences between the two interactions become clear. JoeRobot has the significant added functionality of being able to move. The actions of the robot reinforce its perceived intentions. Its form also supports the notion of the robot having some degree of “intelligence”, even emotion.

While these scenarios are clearly leading, it nevertheless becomes both quite difficult, and questionable as to whether it is possible to design *out* features and functionality that biases the human participant in their interactions with the device (i.e. remove anthropomorphic references as much as possible). Anthropomorphism is both a powerful metaphor and quite an unwieldy imposition [Duffy, 2003].

1.1 Anthropomorphism

Anthropomorphism, whether intentional or not, plays a significant role in our social interaction with artificial systems [Duffy, 2003]. Krementsov and Todes [1991] comment that “*the long history of anthropomorphic metaphors, however, may testify to their inevitability*”. Figure 3 [Duffy, 2003] provides an illustrative “map” of anthropomorphic features as applied to a design of existing robotic heads in the development of social relationships between a physical robot and people. The three extremities of the diagram (human, iconic and abstract) embrace the primary categorisations for robots employing anthropomorphism to some degree. “Human” correlates to an as-close-as-possible proximity replication of the human head. “Iconic” seeks a very minimum set of features as often found in comics that still succeeds in being expressive. The “Abstract” corner refers to more mechanistic functional design of the robot with minimal attention to human-like aesthetics.

The examples demonstrate some of the strategies used to augment the humanness of a robot interface. Some utilise a visually iconic [Breazeal, 2000; Duffy, 2004] whereas others a more strongly realistic humanlike construction (i.e. with synthetic skin and hair) [Hara, 1995] for facial gestures in order to portray artificial emotional states. The more iconic head approach constrains

and attempts to manage the degree of anthropomorphism employed. On the other hand, building mannequin-like robotic heads, where the objective is to hide the "robotic" element as much as possible and blur the issue as to whether one is talking to a machine or a person, results in effectively unconstrained anthropomorphism and a fragile manipulation of robot-human social interaction. As Mori outlined with "The Uncanny Valley" [Mori, 1982]; the closer the design and functionality of the robot comes to the human, the more susceptible it is to failure unless such a high degree of resolution is achieved that its distinction from a human becomes very difficult.

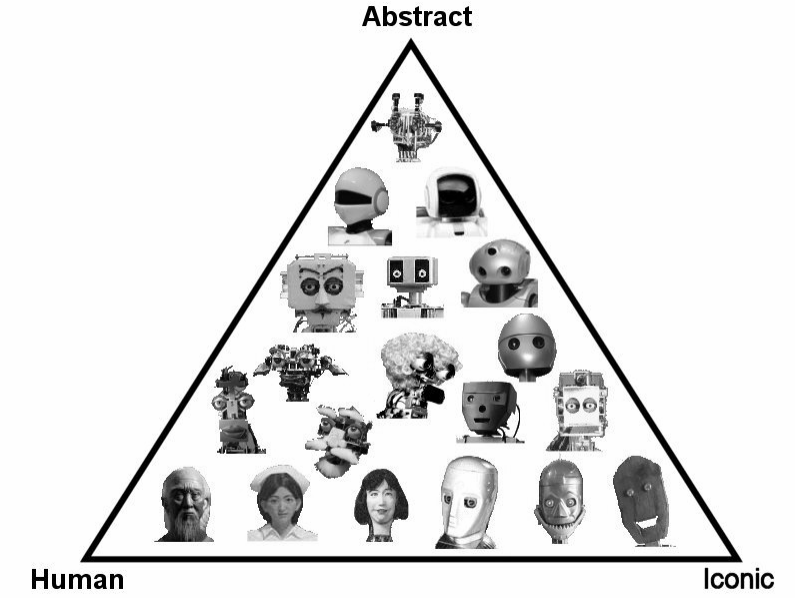

Figure 3. Anthropomorphism design space for robot heads [Duffy, 2003]. Notes: The diagram refers uniquely to the head construction and ignores body function and form. This is also by no means an exhaustive list. Examples were chosen to illustrate the proposed idea (motivated by [McCloud, 1993])

1.2 What's in a name?

What if BlueBox (Figure 1) was named JoeBox? Even something as simple as giving a machine a name can promote anthropomorphism – but not just any old name. The notion of a device having a unique identity has been actively used from the early stages of computer networking. In order for one device to transmit data from one point to another on a network, it needs to know some information about the destination machine and vice versa. The standard Internetworking Protocol (IP) addressing system, one of the core network protocols of the WWW, provides a unique identification which, through Domain Name Servers, results in the machine having some form of name. When a robot communicates data to another, whether through a network or using voice, a minimum degree of transmitter and receiver information is required. It is a fundamental tenet of communication. When robots enter our physical and social space, our references to them will extend from "the robot" to a need to differentiate between more than one. The robot then has, through our projection, at least a minimum form of identity. As it functions, we inherently build internal representations of the robot as we try to rationalise its actions. If we can already attribute qualities we would usually associate with sentient beings (even just a name) to a machine, then we can more easily develop a relationship with it.

We can also extend these associations and expectations with such anthropomorphic paradigms as stereotypes and characteristic traits to facilitate social interaction. Such paradigms can facilitate explicit social interaction between robots [Duffy, 2000] as well as robot-human social interaction by facilitating the robot's development of internal social models of people it interacts with.

Allowing the robot to portray a sense of identity makes it easier for people to treat it as a socially capable participant. In that sense, it is differentiated from the pool of social space. Because it is more a "participant", it is more seamlessly integrated into the social scene rather than simply existing physically somewhere in the social space. Consequently, the difficult philosophical and psychological issues arise of how humans then deal with the associated problems of a robot asserting itself in our social space.

2 The Machine Tool vs. Mr. Robot

If, in the most robotic monotone of voices, you are asked the direct question, "Do you want the light turned on?" You may reply "Yes". Here, the query was addressed and the robot proceeds to turn the light on. Does this represent an example of social interaction or simply the functional interface of a tool? What happens if you said "yes, please" and "thank you" when the task is completed?

The boundaries become blurred in the following two situations. The first involves the issue of being a tool or not. If the language of interaction extends beyond the purely functional, the issue of whether the system becomes more than a tool may become relevant. This raises the issue of intentionality and whether it is intelligent. The second raises the questions of from what level does a system become social. What does it mean to be social? Dictionary definitions of the term "social" are distinctly based on the human frame of reference: "the interaction of the individual and the group, or the welfare of human beings as members of society" [Merriam-Webster]. Connotations of mutual support and friendship are also present. When looking to extend the term social to the artificial systems domain, numerous problems arise, such as; can a robot become a friend? If we are willing to allow a dog be interpreted as a "man's best friend", the extension to the domain of artificial systems becomes plausible.

While Shneiderman [1988] presents strong arguments against the anthropomorphisation of artefacts and interfaces, these rely on the importance of clear, comprehensible and predictable systems that support direct manipulation, i.e. a clear classification as a tool. Should we aim to have a clean cold tool-like approach to the development of the social capacities of the robot and effectively a more anti-anthropomorphism stance? This raises the issue of the purpose of the robot in the first place. If it is clearly designed as a functional tool, then implementing communication mechanisms is only justifiable if they efficiently contribute to the robot's accomplishment of the tasks required of it. Additionally, the most useful robot tools to date are far removed from the human frame of reference, i.e. assembly robotics in manufacturing.

On the other hand, if it is designed for social purposes, then it is intrinsic that people be able to establish a relationship with it. The more it is perceived as personable (which can sometimes be measured by how much it resembles human-like behaviour), the more personable the relationship.

In using the human frame of reference, the key question becomes: Is the robot required to socially engage with people? If yes, the effective communication of the robots' intentions is required. Similarly, if the robot can utilise mechanisms that facilitate our social engagement with the robot such that the robot can interpret our intentions more easily, then these mechanisms could be implemented. Once an artificial system is explicitly designed to engage people socially, we begin to employ the standard mechanisms for understanding and behavioural rationalisation while communicating in our day to day social encounters.

3 The Role of Imperfections

Perhaps the most obvious and effective means of communication is speech-based interaction. In human-to-human interaction, disfluencies in conversation have so much information packaged in them that such "imperfections" contribute greatly to the communication process. They become such an important element of the way we naturally converse and the way we make inferences about the speaker's mental states and intentions [Smith & Clarke, 1993] (see also [Bailey &

Ferreira, 2003]) for an overview) that we usually hardly notice our reliance on these cues [Clark and Fox Tree, 2002]. However, we definitely notice when these cues are missing. Subsequently, it may be difficult to establish a rapport with a social robot when its conversation is more stilted and “clean” (tool-like?).

Along with prosodic information, disfluencies also offer emotional expression. Although emotional expression can lead to misunderstandings in human interaction where they are misread, when unambiguous, they are crucial to interpreting what someone has said (pragmatics). They are often what allow one to identify sarcasm or jokes from truthful statements. The degree of fluency and articulation of the robot’s speech mechanisms can consequently facilitate speech-based interaction. However, if implemented inappropriately, it may load a person’s expectations of the capabilities of the robot in a similar way as humanlike facial features. It may be beneficial to constrain the robot’s speech mechanisms in order to manage expectations [Jacobus and Duffy, 2003].

What is the best paradigm for socially engaging in human-level social interaction for a social robot? Should (artificial) flaws be introduced into the social mechanisms for a robot to be more socially acceptable? Emotional expression can lead to misunderstandings in human interaction where they are misinterpreted, but they can also promote clarification and augment the social interaction overall. Are “flaws” fundamental social features for successful interaction? However, if you re-create this in a machine, do you create unnecessary complicating chronic annoyances?

Arguably, research into the development of artificial emotions [Hara, 1995] in machines can lead to confusing social interfaces where complicated human facial emotional expressions are recreated on a robot head using servomotors and silicone skin. The quality of the artificial expressions varies greatly, as can be seen in the examples shown in figure 3. The justification for creating effectively fake emotional expressions in order to aide a social situation between the robot and a person is difficult when the resolution of expression is so low as to even lead to difficulties in differentiating between happy and angry. On the other hand, implementing small random motion behaviours of the head, the principles of Perlin Noise [Perlin, 1985], could increase the impression that the robot is “alive”, a technique often used in the animation of virtual characters.

4 The Quality of the Fake

A machine is obviously not human, so all endeavours to try and develop a machine to be more socially acceptable in our environments can only try to replicate with varying degrees of success what key human traits support its social integration. It will always be a fake. However, increasing the quality of the synthetic through increasing the complexity resolution of the robot may ultimately make it difficult to know it is a fake. High complexity resolution is therefore where the sophistication of the artefact surpasses our ability to explain and easily predict it. Too high a complexity resolution will lead to a robot so complex that we are unable, through observation, to completely rationalise its behaviour. If the complexity of a robot’s control mechanism is managed adequately, our propensity to over-lay meaning and interpretation on the device may result in our perceiving the device as being intelligent, emotional, and even sentient. This facilitates the establishment of a relationship with such a robot. With the increasing behavioural complexity, our willingness to abstract, generalise, and infer meaning becomes inevitable as we try to understand.

Robots built with little reference to the human occupants that may share its physical space may or may not display what can be justifiably interpreted as *artificially* intelligent behaviour. This necessitates some form of observer-based assessment of the robot (which is not without its complications). If, on the other hand, the robot explicitly targeted the human observer and effectively aimed to include that person in some way both into their physical *and* social space, then there is a social robot paradox. The robot is trying to be something that it may inherently not be ever capable of, i.e. being seamlessly integrated into our social space but, on the other hand, in the very attempt to engage people socially, has become a social robot. A machine trying to be

human is still a machine. This highlights an important point in artificial systems development which has been discussed at length in the literature, for example, the distinction between synthetic and natural systems as presented in discussions about embodiment [Sharkey and Ziemke, 2000; Duffy and Joue, 2001].

In determining whether a machine could appear to think, Alan Turing came up in 1950 with what has become known as the Turing Test [1950]. The test is based on whether a machine could trick a person into believing they were chatting with another person via computer or at least not be sure that it was "only" a machine they were conversing with. Weizenbaum's 1960 conversational computer program Eliza [1966], employed standard tricks and sets of scripts to cover up its lack of understanding to questions it was not pre-programmed for. It has proved successful to the degree that people have been known to form enough attachment to the program that they have shared personal experiences.

But, if there was a balance between the development of highly sophisticated robots with the most powerful artificial intelligence strategies and a physical and behavioural aesthetic which augments its perceived capabilities and "intelligence", could this succeed in realising a socially capable robot? Could technological sophistication in conjunction with theatrical faking realise the holy grail of robotics, an artificial human? The most important criteria for such a system to succeed are to implement mechanisms within the design boundaries of the robot to specifically manage its perception by people who encounter it. Its perceived capabilities should remain within its actual capabilities.

5 Conclusion

This paper has raised numerous issues relating to the social robot paradox. Can the system justifiably become more than a tool? If it employs social conventions generally reserved for human interactions, can it become part of our social circle? If it socially "cheats" in a similar vein to how a system can appear "intelligent", then the consequences may ultimately be more similar to human-human social interactions than we may be comfortable with.

The ideal is a balance between developing a system that can realistically meet the expectations of the people it interacts with while embracing its machine-like qualities that make it ultimately useful. As to whether a machine can be *too* human-like, if it can successfully maintain the sophistication and its perceived intelligence, then the limits become extended. The resolution of the system as a whole (behavioural, aesthetic, and computational complexity) can develop in accordance with advancing technologies. Will a machine become intelligent, social, even emotional? Effectively yes. It will be able to achieve an illusion which, based on our interactions with it, we won't know that fundamentally it is not. Will it be able to maintain this illusion over time? This is the very difficult problem. Given our abilities to often tell when someone is lying, the possibility that a robot could succeed presents hard research questions.

When we experience robots, we draw on our previous experiences, from films and stories and life in general, to compare this new complex device with something with which we are familiar. When the context becomes social interaction with people, the familiar object is inherently our fellow man with all its complications. In the social robot paradox, the robot is basically a machine which can only ever be a tool; however, if it employs our social references, we will neglect to see it as a tool. In managing the complexity resolution and increasing the quality of the fake through, for example, incorporating human imperfections, we will therefore facilitate the development of a human-robot relationship. The machine tool can then become Mr. Robot.

References

Bailey, K.G.D., & Ferreira, F., (2003), "Disfluencies affect the parsing of garden-gate sentences" Journal of Memory and Language (in press)

Breazeal, C. Sociable Machines: Expressive Social Exchange Between Humans and Robots, Sc.D. dissertation, Department of Electrical Engineering and Computer Science, MIT. 2000

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, 2003

Duffy, B.R The Social Robot Ph.D Thesis, November 2000, Department of Computer Science, University College Dublin

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

Duffy, B.R., The Anthropos Project: http://anthropos.medialabeurope.org

Clark, H. H. & Fox Tree, J. E., “Using uh and um in spontaneous speech”, Cognition, 84, p73-111, 2002

Hara, F. Kobayashi, H., Use of Face Robot for Human-Computer Communication, Proceedings of International Conference on System, Man and Cybernetics, pp10. 1995

Jacobus, E., Duffy, B.R., "The Language of Machines", Digital Interaction, International Symposium on Information & Communication Technologies, September, Trinity College Dublin, Ireland, 2003

Krementsov, N. L., & Todes, D. P. On Metaphors, Animals, and Us, Journal of Social Issues, 47(3), pp67-81. 1991

McCloud, Scott, Understanding Comics: The Invisible Art, Kitchen Sink Press, 1993

Merriam-Webster Online Dictionary, http://www.m-w.com

Mori, M., The Buddha in the Robot. Charles E. Tuttle Co., 1982; ISBN 4333010020

Perlin, K. An image synthesizer. Computer Graphics (SIGGRAPH '85 Proceedings), volume 19, pages pp287-296, July 1985.

Sharkey N., Zeimke, T. “Life, mind and robots: The ins and outs of embodied cognition”, In S. Wermter & R. Sun (eds), Symbolic and Neural Net Hybrids, MIT Press, 2000

Shneiderman, B. A nonanthropomorphic style guide: Overcoming the humpty-dumpty syndrome. The Computing Teacher, October, 9-10. 1988

Smith, V.L., & Clarke, H.H., (1993) “On the course of answering questions”, Journal of Memory and Language, 32, 25-38

Turing, A. M. Computing machinery and intelligence, Mind Vol.59, pp433-460, 1950

Weizenbaum, J., ELIZA - a computer program for the study of natural language communication be-tween man and machine. Communications of the ACM 9. 1966.

Figure 3: Robot Heads (from top to bottom-right): COG:MIT-AI Lab, www.ai.mit.edu; SIG: Kitano Symbiotic Systems Project, www.symbio.jst.go.jp/sigE.htm; ASIMO: Honda www.honda.co.jp/ASIMO/; The Humanoid Cranium Robot: Waseda University, www.humanoid.waseda.ac.jp; H6: JSK Laboratory, www.jsk.t.u-tokyo.ac.jp/research/h6/; SDR4X: Sony, www.sony.com.au/aibo/; Kismet: MIT AI Lab www.ai.mit.edu/projects/sociable/; JoeRobot: Media Lab Europe, anthropos.mle.ie; Isamu: Kawada Industries Inc., www.kawada.co.jp/ams/isamu/index_e.html; Inkha: King's College London, www.inkha.net; Doc Beardsley: CMU, www.etc.cmu.edu/projects/iai/; Elvis: Chalmers University of Technology, human-oid.fy.chalmers.se; Hadalay-2: Waseda University, www.humanoid.waseda.ac.jp/; Master Lee: YFX Studio, www.yfxstudio.com/human.htm; Saya: Kobayashi Lab, koba0005.me.kagu.sut.ac.jp/newsinfo.html; Roberta: Science University of Tokyo, hafu0103.me.kagu.sut.ac.jp/haralab/; R.Max: www.howtoandroid.com; Maxwell: Medonis Engineering, www.medonis.com; Woody: Media Lab Europe, anthropos.mle.ie