I, Robot Being

Brian R. DUFFY[1], Gina JOUE[2]
[1]*Media Lab Europe, Sugar House Lane, Bellevue, Dublin*
[2] *Dept. Computer Science, University College Dublin, Ireland*

Abstract There is the age-old paradox of technologists predicting bleak futures for mankind because of their research directions but nevertheless hurtling full steam ahead in pursuing them. Strong humanoid research could be one such example. The holy grail of robotics and artificial intelligence research is regularly perceived as aiming to understand mankind by artificially creating a similar degree of functionality and form. But what are the consequences of this pursuit?

1. Introduction

Man and machine have been merging for a considerable time. As early as the mid 1700s, Vaucanson's flute and tabor players provided mechanically driven functionality to human-like forms. Wolfgang von Kempelen's chess player, the "Turk", famously illustrated the power of perceived intelligence in machines through people's willingness to accept what effectively was an illusion (See [16]). These automata were based on the human form and consequently drew on people's expectations and projections in order for the illusion to work. Half the work of selling the illusion using the human as the reference would be acheived if only it could be maintained. Kempelen's chess player proved very successful, enduring for quite a long career playing such people as Napoleon. It is interesting to wager whether its long-term sustainability was fundamentally due to a human hidden inside the automaton to reduce the pressure on the mechanical design of the system in maintaining the illusion.

This paper explores the idea of the robot as a human and as a machine and how its dual roles have impacted research to date and its future directions.

2. The Intelligent robot

Robotics and artificial intelligence research has taken a primarily different stance on building "automata" or automatic machines. AI has developed over the years with the aim of realising a system that could exhibit a degree of intelligence in its own right, ideally without the props and illusion found in age-old automata. The problem has been that too much emphasis has been placed on developing an "intelligent" system with little to no dependence on either a physical and/or social context. They were systems that provide an often highly sophisticated mapping between an input and an output that could be algorithmic in form. The robot would not "know" if it ever was intelligent. It would need to be assessed by someone/something else and told it was actually intelligent or not, the Turing Test being the most classical example [15]. An observer assesses the system and arrives at a conclusion over time through its interaction with the system.

Minsky is recently quoted as saying that "AI has been brain-dead since the 1970s" (Wired, May 2003). While intriguing, the article highlights the key aspects AI research has had to deal with over the last twenty or thirty years. AI research has, for example, lead to very powerful solutions for web searching, and has reduced chess to a state-space problem rather than a reference for intelligence. Once problems, originally thought of as hard issues, are solved or even just understood better, they loose the "intelligent" status and become simply software algorithms.

Robots provide an altogether different scenario as highlighted so clearly by the robot "Shakey" [11]. The real world is a very different place than the defined digital space. Once the robot comes into our physical and social space, our interaction with it becomes inevitable. It leads to boggling questions of what knowledge is relevant when in a given situation and what capabilities must the robot have. It has consequently been easy to view the full human-like robot as the reference for both the pursuit of true artificial intelligence and the ideal autonomous machine which functions in our space. But one should ask *why* a humanoid robot? What if we freed the machine of the chains of mankind and its references? What capabilities would it, should it have? And would we understand it?

It is easy to follow the thesis that given higher resolution (better technology), current humanoid projects such as Honda's Asimo [18] and Sony's SDR-4X [17] could succeed in making robots like people. Fine-tuning the animatronics, the sensor modalities, the actuators and the aesthetics could all ultimately lead to the "replicant" problem, as illustrated in "Blade Runner" (Warner Studios, 1982). But, the question should not have escalated so quickly to "is it a man or a machine?" This perspective often forgets the demands of function. Moreover, it assumes that we will accept robots in our society and interact with them only if they are in our own image.

The key issue in social robot design and effectively human-robot interaction is the balancing of the similarities and differences to ourselves, making robots acceptable and not feared. This involves an understanding of people's own qualities and consequently their expectations of others. For example, an ideal robot design would incorporate only people's strengths but not their weaknesses. For example, why create a robot assistant that is forgetful just to make it more humanlike? On the other hand, other "weaknesses" can facilitate interaction by making the robot more endearing. They become features of the system that help a robot to integrate into society. Marvin, the incredibly depressed robot in Douglas Adams' "*The Hitchhikers Guide to the Galaxy*" represents an example of a robot being popular simply because of its exaggerated humanlike failings.

These points highlight nontrivial questions. What are our faults that we would not like the robot to have and those that we do? How different from us can and should the robot be?

3. The Uncanny Valley

Masahiro Mori at the Tokyo Institute of Technology aimed to provide an insight into human psychological reaction to robotic design [9]. His idea of the "Uncanny Valley" follows from Freud's description of the uncanny (a translation of the German word "unheimlich") [6]: "*derives its terror not from something externally alien or unknown but--on the contrary--from something strangely familiar which defeats our efforts to separate ourselves from it*" [10].

Mori's graphs are becoming a reference on explaining why some robot designs for social interaction work and others simply don't. Freud discusses how, in the context of literature, an author can evoke an uncanny response on the part of the reader by straddling the line between reality and unreality, a feature regularly used in horror stories and film.

Mori's Uncanny Valley represents how an entity can be perceived as having enough humanlike characteristics to arouse a constrained degree of empathy through one's ability to rationalize its actions and appearance. However, when too humanlike, it results in over expectation of its capabilities and a negative reaction from the observer. Finally, it becomes so humanlike that it is effectively treated as a human where it has re-established a balance between anticipated and actual function and form to a sufficient degree that works.

As highlighted by Reichardt [13], based on the relative proportions of the curves, Mori may consider "motion more important than simple appearance, though he stresses that both are affected at least as much by subtle nuances as by more striking factors". When the robot begins to move, the balance between form and function becomes crucial.

Mori proposes that the ideal situation is therefore to design aesthetically pleasing robots whilst maintaining a degree of visible artificiality. This would be aiming for the top of the first peak and therefore avoid the Uncanny Valley.

This Uncanny Valley effect was illustrated by Thomas Edison in his Talking Doll (named Marguerite) in 1888. As an economic venture, it failed primarily due to its attempt to manipulate people's expectations of what a doll's form and function should be. A child is reputed to have said that it wasn't right as how could the doll digest its food with the mechanism of the phonograph in its stomach? Rodney Brooks recently introduced his "My Real Baby" which similarly incorporated technological functionality that played with a child's expectations.

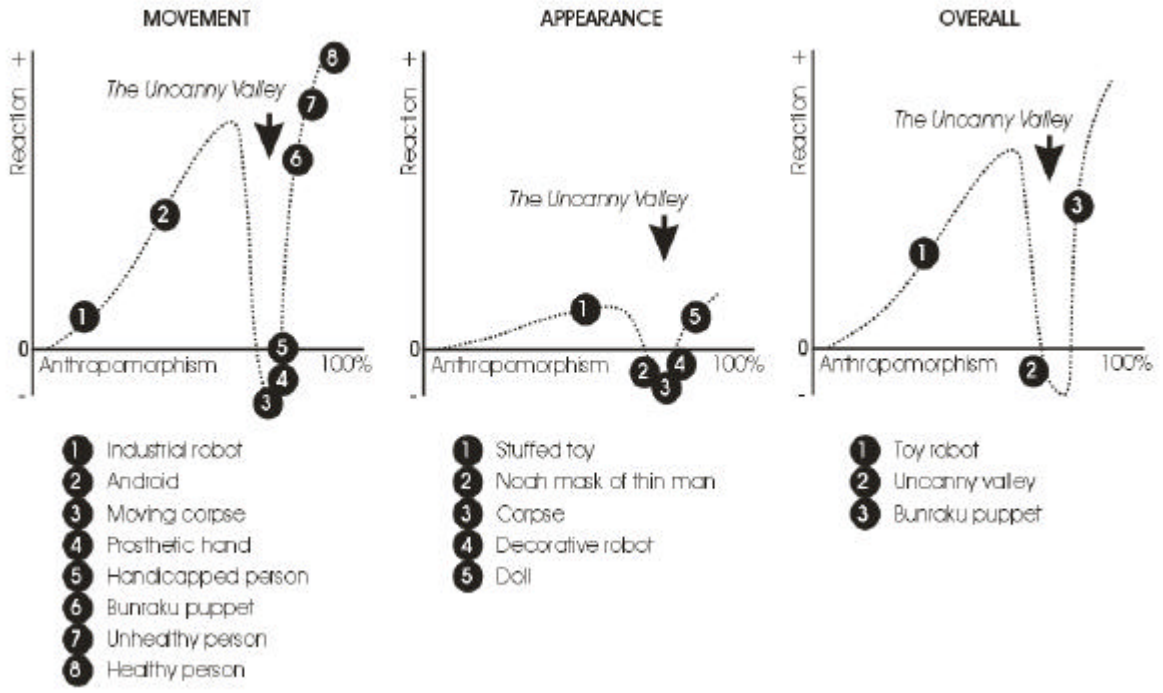

Figure 1. Mori's "The Uncanny Valley" [9] (see also [14])

Such examples illustrate the power of imagination when people interact with machines. Such projection of characters and personalities onto toys will inevitably become more prevalent with robots when they are incorporated as domestic devices in our homes.

Similarly, the design of a robot as a social entity can promote the overlay of personalities and characters based on either its explicit or implicit behaviour. In looking at how the human is used as a reference in robot design, there are three primary stages:

1. It is obviously a machine and maintained as such – people's willingness to accept it doesn't present any major issues. Its role as a tool is clear and unambiguous.
2. It's the perfect synthetic humanoid: The robot's functionality is fundamentally constrained to the human capability set. Anything less and the objective of realising a perfect artificial human has failed.
3. It's a mildly humanised machine (e.g. see figure 2). An effort is made to facilitate the social interaction between the human and the robot without compromising the functional expectations of the machine too much. The judicious employment of anthropomorphism provides the balance (see [4] for a more detailed discussion of the role of anthropomorphism in social robotics).

Figure 2. Media Lab Europe's "JoeRobot" (Photo courtesy of Brent Jones)

In fact, if robots are *too* human-like, then we would have more difficulty in accepting them. If the robot is *perceived* as being human, then effectively it *should be* human and not a robot in form and function. Even if one succeeded and could build a perfect robot reproduction of a man or woman, what would it do? It would in fact only do what we can do, no more, and no less (this discussion does not incorporate aspects of human augmentation). The addition (or subtraction) of any attribute not pertaining to our human form and function results in a deviation from the original holy grail of building an artificial human. Building a robot to replicate the form and function of the human as closely as possible could "only" be for creations to replace us and give rise to the fear of robots "taking over the world".

There is little doubt regarding the power of anthropomorphism in robot design and particularly social robots. The key is to manage it and employ its advantages without succumbing to its trappings (see [4] for a discussion). As our interaction with robots becomes inevitable, our need to facilitate this interaction comes to the fore. Judicious anthropomorphism is the key to

facilitating social interaction while maintaining those useful features that are fundamentally mechanistic. However, when machines can be such valuable tools, why would we want a robot to be anything other than a tool? The optimal solution is a balance between function and form. The presumption that the machine-like qualities of a robot being a disadvantage is unfounded. These are not flaws to be hidden and can even be worked to facilitate task-oriented human-robot interaction. It's a case of delving into the design pool and only using those features that constructively contributes to the solution. Of course this is more easily said than done!

4. Perceiving Intelligence in Robots

To augment the usefulness of robots in our world, minimising frustration in our interaction with "stupid machines," does not necessarily require a Super Robot. In order to design such a sophisticated robot, roboticists may arguably only need to have a design that is as complex as its *perceived* behaviour. If played right, the social capabilities and sophistication (and therefore perceived intelligence) of the social robot can be achieved through an illusion (and effectively returns to the original automata discussed in the introduction). One just needs to fake it the right way.

The Machiavellian Intelligence Hypothesis (see [8] for a recent discussion) proposes that intelligence as we understand it evolved from the social domain where social interaction between entities is key to the development of intelligence. Once there is such social interaction, there is the tendency to rationalise another's actions and employ such concepts as emotions and intelligence [3]. This induces a considerable degree of subjectivity in the assessment of the system. Consequently, when robots become more embedded in our physical and social space, an artificial manifestation of the Machiavellian Intelligence Hypothesis results in us interpreting the machines according to our own references, and the robot will therefore become "intelligent" and even "emotional". The prefix of it being *artificially* intelligent will become a similar label as we refer to a dog as having "dog intelligence."

It is open to debate whether a system can maintain such an illusion of intelligence developed through social interaction. The panel of NIST's 2000 Performance Metrics Workshop agreed that "social behavio[u]r is fundamental: it compensates for the lack of perfection of the individual intelligent system". The robot maintaining its status as a tool, albeit more autonomous and sophisticated, provides a continuous recalibration of its social failings. If it gets the job done, the details of its inadequacies become less relevant.

Exploring the mechanisms underlying our tendencies to anthropomorphise could provide key insights into the development of the "illusion of life" and the "illusion of intelligence" through projective intelligence in embodied robotic systems. If a robot looks and acts "intelligently", then the particular computational mechanisms underlying the realisation of that behaviour are arbitrary.

This highlights a key difference between research many people term as AI and research that directly draws on biologically inspired models of the brain and human functionality. The term "artificial intelligence" consequently has two camps, the field of research that seeks to model and replicate human intelligence through biological analysis of human intelligence. The second is work that inherently draws on the capability of computing strategies as fundamentally symbol manipulating systems in order to develop more and more sophisticated "intelligent" entities (whether instantiated as robots or in software). The age-old debate of what is artificial intelligence doesn't become clearer. It inevitably has to first climb the hurdle of what is "intelligence".

To date, we have been hesitant to view robots and their reasoning capabilities as "robot intelligence" and thereby free the strong human reference from the machine. Biological vs. mechanistic are so fundamentally different that correlating the two can only be illusionary (see [5]).

The difficult issue arises in how to maintain such an illusion of "intelligence" over time. Does this inherently require that the robot be "intelligent" at a more fundamental level? Do we need to go beyond "pageantry" towards the "truth" if the result is the same? Can pageantry intelligence endure over time? Social robotics provides a powerful framework to finally build a system that will pass the Turing Test. While anthropomorphism is clearly a very complex notion, it intuitively provides us with very powerful physical and social features that are being implemented to a greater extent in social robotics research.

Social robots provide another instantiation of how people can relate to devices in similar ways as they would real people, as discussed by Reeves' and Nass' in [12]. Experiments on the physicality of the robot and its motions have already demonstrated people's propensity to ascribe human like characteristics and descriptors to machines [3]. As social interaction is the feature of the robot, anthropomorphism is inevitable.

Experiments have highlighted the influence of appearance and voice/speech on people's judgements of another's intelligence. Generally, the more attractive a person, the more likely others would rate the person as more intelligent [1][2]. However, when given the chance to hear the person speak, people seem to rate intelligence of the person more on verbal cues than the person's attractiveness [1]. Exploring the impact of such hypotheses to HCI, Kiesler and Goetz [7] undertook experimentation with a number of robots to ascertain if participants interacting with robots drew similar assessments of "intelligence". The experiments were based on visual, audio and audiovisual interactions. Interestingly the results showed strong correlations with Alicke et al's [1] and Borkenau's [2] experiments with people-people judgements.

The way we perceive the intelligence of the robots cannot be ignored and researchers and designers must therefore aim to manage it in order to achieve a coherent balance between function and form. Just as with people-people judgements, a robot's behaviour and form can facilitate our initial impressions of its "intelligence," but it should also be at least able to do what one has the impression that it can do. Otherwise the illusion, and thus the interaction, fails because the overall perceived intelligence of the robot is a combination of individual (internal), observed (projective), and interactive (social) intelligence. Internal intelligence is the most powerful but one of the most difficult to design. Our tendency to project intelligence reduces the complexity of engineering internal intelligence, but observed intelligence is more ephemeral. Social intelligence requires interaction to exist. And conversely, interaction requires social intelligence. However, its inherent dynamics may make it even more difficult to design, but extremely powerful if successful as it is more characteristic of human than machine.

For example, consider a robot with the ability to adapt to special situations it's not specifically programmed for, similar to the way people can (should) be able to improvise and adapt. Increasing the robot's ability to adapt to such circumstances has been the objective of AI research for a long time and has proven a very difficult task. The ability to follow the rulebook reinforces the tool paradigm for the robot. It is given a task and it completes it. But sometimes the rulebook doesn't fit or even apply. This necessitates adaptation and the ability to deal with the unknown, a characteristic humans have and robots are severely lacking. Here, the robot crosses stereotypical machine limitations and in particular the stereotypes that can often make it less acceptable (i.e. you'd rather talk to a human who can be sympathetic or show common sense than a robot that you cannot reason with).

The ability to make decisions that show common sense, sympathy, and more humanlike qualities is key to a machine being able to reliably function in our environments. Common sense reasoning research has developed over nearly the last two decades with such projects as the Cyc initiative (http://www.cyc.com) and the resulting OpenCyc project. The storing of information and the associative mappings between this information in the form of common sense knowledge that we use from day to day is the goal of this and similar work. A common sense reasoning mechanism then draws on this knowledge and forms the basis of the decision making as the system interacts with the real world. It will be interesting to see how this develops and is instantiated on robotic systems in our physical and social environments.

The social robot is the next important stage in robotics and will fuel controversy over existing approaches to realising artificial intelligence and artificial consciousness. Invariably, researchers will not be happy building anything less than a fully functional humanoid robot. It's just not in their nature.

5. For the future

If robots become our constant companions at home and about the workplace, will their irritating idiosyncrasies such as flawed vision, annoying speech and woefully inadequate sensor systems become the physical equivalent of a spam assault through chronic annoyances? This raises an interesting point about machines as "constant companions"; what are the health and environmental drawbacks of having machines embedded in our physical and social space as autonomous entities?

In addressing such an argument against robots, we should note that we have few issues with submerging ourselves in technological devices in the home and at the office. The key difference is that we see all these devices as tools. If a robot becomes autonomous, it will either consciously or subconsciously demand a degree of being treated as an individual. It will have a name (or a number) which will facilitate an identity. It will communicate socially with context. Will this result in the robot loosing its status as simply a tool? The notion of artificial life and perceived life in machines (as illustrated by Reeves and Nass [12]) will no doubt promote robot's rights issues. It will be confusing. The source of the ambiguity rests in the design.

In returning to the title of this paper, "I, Robot Being", research shows how a robots simple existence in (including its motion around) our physical and social space induces notions of identity, intelligence and emotions [7][12][3], whether they actually have them or not. It is inevitable. Consequently, the design of the social robot is crucial. If too imbalanced, its uncanniness will effectively demand a social presence in our community which no doubt will bring its own set of complications. Who will fight for the rights of the robot? Will you be afraid to give it any more orders in case it feels you are condescending? Will you hurt its feelings? Will a robot die if you turn it off?

References

[1] Alicke, M.D., Smith, R.H., & Klotz, (1986). "M.L. Judgments of physical attractiveness: The role of faces and bodies". *Personality and Social Psychology Bulletin, 12(4), pp381-389*.

[2] Borkenau, P. (1993). "How accurate are judgments of intelligence by strangers?" *Annual Meeting of the American Psychological Association, Toronto, Ontario, Canada, August*

[3] Bourke, J., Duffy, B.R., (2003). "Emotion Machines", *IEEE Systems, Man & Cybernetics Workshop (UK&ROI Chapter), Reading, September*

[4] Duffy, B.R., (2003). "Anthropomorphism and The Social Robot", *Special Issue on Socially Interactive Robots, Robotics and Autonomous Systems 42 (3-4), 31 March*

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

[6] Freud, S., "The Uncanny", (1953). *The Standard Edition of the Complete Psychological Works of Sigmund Freud, ed. & trs. James Strachey, vol. XVII (London: Hogarth), pp. 219-252*

[7] Kiesler, S., Goetz, J., "Mental models and cooperation with robotic assistants", http://www-2.cs.cmu.edu/~nursebot/web/papers/robot_chi_nonanon.pdf

[8] Kummer, H., Daston, L., Gigerenzer, G., Silk, J., "The social intelligence hypothesis", (1997) *Weingart et al. (eds), Human by Nature: between biology and social sciences. Hillsdale, NJ: Lawrence Erlbaum Assoc., P157-179*

[9] Mori, M., (1997) The Buddha in the Robot. *Charles E. Tuttle Co.*

[10] Morris, D., (1985). "Gothic Sublimity." *New Literary History*, Winter

[11] Nilsson, N.J., (1984). "Shakey the robot", *SRI A.I. Center Technical Note 323, April*

[12] Reeves, B., & Nass, C., (1996). The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places, *CSLI Publications / University of Chicago Press.*

[13] Reichardt, J., (1978) Robots: Fact, Fiction and Prediction, *Penguin Books.*

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

[15] Turing, A. M. (1950). "Computing machinery and intelligence", *Mind Vol.59, pp433-460*

[16] Wood, G., (2002). Living Dolls: A Magical History of the Quest for Mechanical Life, *Faber & Faber.*

[17] Sony Corporation SDR4X Website: http://www.sony.com.au/aibo/

[18] Honda Corporation ASIMO Website: http://www.honda.co.jp/ASIMO/