Fundamental Issues in Affective Intelligent Social Machines

Brian R. Duffy*

SMARTlab Digital Media Institute, University of East London, 4-6 University Way, London, UK

Abstract: In the early 1990s, computer scientists became motivated by the idea of rendering human-computer interactions more humanlike and natural for their users in order to both address complaints that technologies impose a mechanical (sometimes even anti-social) aesthetic to their everyday environment, and also investigate innovative ways to manage system-environment complexity. With the recent development of the field of Social Robotics and particularly Human-Robot Interaction, the integration of intentional emotional mechanisms in a system's control architecture becomes inevitable. Unfortunately, this presents significant issues that must be addressed for a *successful* emotional artificial system to be developed. This paper provides an additional dimension to documented arguments for and against the introduction of emotions into artificial systems by highlighting some fundamental paradoxes, mistakes, and proposes guidelines for how to develop successful affective intelligent social machines.

INTRODUCTION

Throughout history, inventors and engineers have worked to develop the artificial intelligent emotional human to understand humanity and to achieve a bond between man and his creation. It is one of the greatest challenges in science. Creating such devices has suffered many failings over the centuries along with some surprising successes. Our fascination with creating humanoids dates as far back as early mythology such as the Golem in ancient Jewish legend, to Leonardo da Vinci's wooden man based on his Vitruvian Canon (c. 1495), Jacques de Vaucanson's flute and tabor players (c. 1735), and Henri Maillardet's writing and drawing figure (c. 1815) to name but a few. Automata based on the human form draw on people's expectations and projections in order achieve a social link. However, in addition to a particular aesthetic which can be relatively attainable, the human form obliges and even *requires* the automata to have behaviours and reasoning mechanisms including emotional capabilities to match its form (i.e. integrated form and function), something that is a more daunting challenge.

Computing technology has become pervasive within socio-physical contexts (domestic, personal portable assistants PDAs, robot assistants, personalised and personalisable web-sites). Government agencies and large companies have replaced human telephone operators in services such as customer assistance with automated vocal robots offering a series of menu choices for customers to choose from in order to find an answer to their query, and automated transaction machines (ATM) have already replaced many functions of bank tellers. Their key advantages range from a more efficient redirecting of queries towards appropriate resources, to achieving relatively low operating costs.

Unfortunately, one of the key disadvantages of employing such technologies is the perceived mechanical coldness of the interaction. In looking to address this and to augment the human-machine interaction paradigm, a number of computer scientists and engineers became motivated towards shifting the focus of their study away from the classical concerns of computing power, computer speed, and products with increasingly sophisticated functionalities, which most users do not exploit, toward the systematic study of computer interfaces, and in particular, toward incorporating, within the technologies themselves, a certain kind of social intelligence in order to restore the social dimension of human societies.

While it has been long recognized that emotions, personality, behaviours and attitudes and such strongly human-centric features could be key features in human-machine interaction, there has been very limited success to date in strongly embracing these issues, primarily due to their complexity. It is difficult to computationally define emotions and manage social interactions in a deterministic fashion, particularly when communication is sustained over a significant period of time.

Nevertheless, the field of affective computing (and emotion research in general) has developed significantly in recent years. Core arguments are grounded on emotion being a prerequisite of (artificial) intelligence through the well-known interrelation between cognition and emotion in natural systems, and the role of emotional expression as a crucial element of believable anthropomorphic behaviour for social interaction.

The Hundred Year Error

Perhaps one of the most famous examples of an unsuccessful humanoid was Thomas Edison's 19th century

*Address correspondence to this author at the SMARTlab Digital Media Institute, University of East London, Docklands Campus, 4-6 University Way, London, E16 2RD UK; E-mail: brd@media.mit.edu

1874-0618/08

"Talking Doll". Edison decided to take the established formula of a doll and augment it with a miniaturised version of the phonograph, one of the most advanced technologies of the day. Now the baby doll could talk! Edison invested heavily in a manufacturing line which could then mass produce this revolutionary doll and be able to sell it to children everywhere. It failed as a business venture however, and was only marketed for a few weeks in 1890. Arguments for why the doll did not succeed range from the technology being too young with only a limited vocabulary available, to the doll being too heavy with the phonograph mechanism installed in a metal torso. Possibly the most revealing insight is the story of a child having seen the phonograph mounted in the body of the doll and being surprised, asking how could it digest its food?

In retrospect, the core reason for its failure should be attributed to it being one step towards removing the need for one of the most fundamentally necessary elements between the child and the doll, the imagination. In using a relatively simple anthropomorphic form, a child's imagination builds a whole world around the doll, its personality, its desires and needs, and even the world it lives in. Incorporating a speech device reduces the dolls "language" to a finite set of possibilities, constrained by the existing technologies. This device removes a significant degree of imaginative freedom from what the doll *could* "say" and "do".

When we encounter the humanoid form, we create expectations about its physical and social behaviours that become more and more difficult to manage with increasing anthropomorphic resolution. This difficulty, where more anthropomorphism does not necessarily mean better, is captured in what has been referred to as Mori's "Uncanny Valley" [1]. When a humanoid's behaviour explicitly conflicts with our expectations derived from its form, the illusion fails. More recently, approximately 100 years after the failure of Edison's "Talking Doll", a baby doll which incorporates such technological advances as embedded speech generation, models of emotion, and servomotors, was not as successful as expected. "My Real Baby", a robotic baby doll developed in collaboration between Hasbro and iRobot Corporation, repeats the same 100 year old error. We easily give "life" to something that doesn't ask too loudly to be alive (a simple doll), rather than when it demands to be alive.

The issues exemplified here provide a number of insights into the difficulties pertaining to the field of artificial social machines. One of the aims of this paper is to encourage an active discussion of what is in fact both a very difficult and very interesting field and provide a structured analysis of the fundamental issues involved. The following sections look to highlight these in the context of affective intelligent social machines. This is concluded with a discussion of how to manage these issues.

THE EMOTIONAL MACHINE

The existence of emotional machines and the issues arising from their integration in our society have been an important topic in science fiction. The natural inclusion of emotional machines in a story represents our general perception of the necessity or inevitability of emotion, perhaps of its crucial role in social intelligence (which is necessary if machines are to be a part of our social space). Science fiction has and continues to play an important role in the development of social affective intelligent machines and their integration into human society, with many directly addressing fundamental issues discussed in this paper.

HAL 9000, the conscious computer in Arthur C Clarke's book *2001: A Space Odyssey* becomes paranoid and afraid for his "life", which leads him to murder the crew of the ship he controls – the machine replaces mankind.

Marvin from the *Hitchhikers Guide to the Galaxy* [2]: "Marvin is a very depressed robot on the starship 'Heart of Gold'. He is very, very depressed about this. He has a brain the size of a planet, yet is rarely given the chance to use it" – The social implications of embedding strong emotional features on machines.

The character Spock from the original *Star Trek* series is intelligent, rational and perceived as unemotional. In fact, as clarified by the actor Leonard Nimoy in 1995, Spock's ability was in suppressing emotion expressiveness as well as restricting emotions from befuddling rational thought, rather than being unemotional – the perceived dichotomy between being rational and being emotional.

In *Do Robots Dream of Electric Sleep?* [3], a humanoid robot is distressed to learn that her memories are not real, but rather they have been implanted in her silicon brain by her programmer – the implications if one *could* succeed in building near perfect artificial human replicas.

In *Bicentennial Man* [4], by Isaac Asimov the robot character displays such un-machine-like qualities as creativity and subsequently redesigns its own circuitry so that he can experience the full range of human feelings and ultimately what it truly means to become human through growing old – The development of autonomy and the corresponding evolution of the machine over time.

AI and the Emotional Machine

The general marginalization of affect in most AI models of intelligence may be partly due to the history of emotion research in psychology. Indeed, emotion research has only recently emerged from its dark ages (roughly between 1920 and 1960) in contrast with its classical phase starting at the end of the 19th century. Whereas psychologist William James [5] offered a very Darwinian view of emotion, restoring emotions as both valuable and part of the evolutionary process, Cannon [6] disagreed and relegated emotions as non-specific, disruptive processes. Findings about the evidence of universality and specificity in emotion-expressive behaviour led to the emotion renaissance of the early 1960s. The development of the field of neurosciences at this time contributed to our understanding, and confirmed the strong intertwining of emotions and reasoning [7]. As a consequence AI, which was only formally initiated in 1956, founded most of its models of intelligence on previously established affect-less theories of intelligence, often rooted exclusively in logic: Classical AI.

Despite the initial hesitancy in affective computing, its development in recent years has been progressive. Michaud *et al.* [8] argue that emotions play a key role in human social interaction (see original paper for more detail)

- to help adapt to limitations
- for managing social behaviour, which are directly associated with the *universal problems of adaptation* [9]
- for interpersonal communication.

Picard, among others, extends this further and argues that the inability of today's computers to recognize, express, and have emotions severely limits their ability to act intelligently and interact naturally with us [10]. Other researchers like McCarthy do not agree that emotions should play a role in artificial systems [11, 12]. While indicating that it would be possible, McCarthy's arguments include the necessity to maintain "rational" behaviour in artificial systems, particularly as they are machines and not human:

"*It is also practically important to avoid making robots that are reasonable targets for either human sympathy or dislike. If robots are visibly sad, bored or angry, humans, starting with children, will react to them as persons. Then they would very likely come to occupy some status in human society. Human society is complicated enough already.*"

The following sections provide a much debated distinction between a machine that aims to *be* an affective reasoner and one which is capable of perceiving and processing affective information and creating some affective-looking output with a view to facilitating human-machine interaction. These two are by no means mutually exclusive, but this distinction helps to look at the issues from two perspectives: Weak artificial emotion *vs* strong artificial emotion – analogous to weak and strong artificial intelligence.

The Man-Machine Interaction Perspective

In the 1970s, Weizenbaum developed ELIZA, one of the first programs that crudely imitated "a Rogerian psychotherapist engaged in an initial interview with a patient" [13]. Weizenbaum was concerned with the results when people interacted with the system, particularly when students at MIT started sharing their intimate thoughts with it and developing an emotional attachment. Kismet [14] embraces the development of such bonds by focusing on the role that social motivations and facial expression-based feedback play in maintaining an appropriate level of stimulation during social interaction. Kismet's interaction was limited to paralinguistic sounds, limited facial expressions and a vision tracking system for objects used in the interaction. While the range of emotional expressions available to Kismet was limited, they were convincing enough in an iconic way to generate sympathy among people who interact with it. The result is a system that evokes some emotional responses in humans when they engage in a limited degree of social interaction.

Recently, the "Personalization On Demand" concept vehicle ("POD") was developed by Toyota and Sony Corporation to explore the potential for communications between vehicles and their drivers. The developers compare the functioning of the system with how one forms a friendship: the pod listens to the driver, monitors his driving habits, and records a wide variety of his personal preferences in order to create an apparently "enhanced driving experience". Expressiveness of the pod is achieved through different coloured lighting systems on the exterior of the car (Fig. **1**). It is uncertain as yet whether such a system could improve the driving experience and safety on the road, or what the specific goal of the project is other than "to explore and strengthen the bond between car and driver through a give-and-take relationship". The machine functions as an externalisation of the drives emotional state, a notion that often contradicts the idea of controlling ones emotions and even hiding them when an inappropriate social state or action.

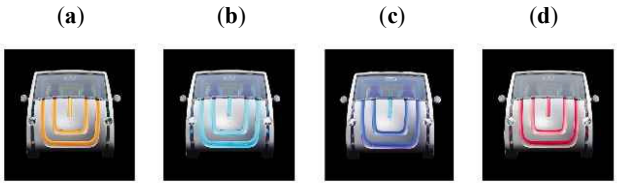

Fig. (1). Toyota POD: (**a**) happy: orange, (**b**) sleepy: light blue, (**c**) sad: dark blue, (**d**) angry: red.

These examples of affective machines represent the diversity of the research to date in developing artificial affective machines, whether virtual (as discussed by [15]) or physical, which has looked at the issues of behavioural believability and augmenting man-machine interaction. These affective machines primarily rely on externally perceived emotion to facilitate interaction. As discussed in the following section, affective machines can also have inherent emotional reasoning capacity, that is, they can be computational devices having (or simulating having) emotions internally.

The Machine Perspective

Donald A. Norman is quoted as saying [16]: "[The robot] had to get frustrated. Being frustrated gets it out of deadlocks. If it's stuck somewhere trapped in a corner, it has intelligent algorithms trying to get it out. But if they fail, it says, 'the hell with it,' and goes off and does something else. It should be afraid of heights so that it doesn't fall down the stairs. It should get fatigued so that it won't wear out the battery. As its battery level gets lower, it should travel more slowly and not do some tasks. It should always make sure it's close enough to the recharging station so that it can get back."

Implementing an artificial version of human emotions on a machine can be viewed as for the *machine's* own advantage. Developing machine "emotions" can look to keep the machines safe (help them make the right decisions) and to make them curious (consequently help them to learn). The association with human emotions is therefore purely analogous. Invariably these could not be termed emotions,

but rather control mechanisms to achieve a particular function, an example being a safety mechanism in response to a dangerous situation (rather than "fear"). Norman *et al.* [17] ask the question why use affect? Why not just program the system to safeguard itself against problems? (Is there in fact a *real* difference between the two when implemented on a machine? – discussed later in section 4). This is true for such a system employing "emotional" reasoning mechanisms internally to negotiate the complexity of the physical world, but when explicit social emotional interaction is pursued with humans, the use of expressive faces for example externalises the affect and encourages its interpretation as corresponding to an emotional state.

From a deliberative perspective, it is often argued that emotion functions to limit reasoning and thereby making reasoning possible [18] where emotions function to short circuit cumbersome deliberation when necessary (see also [19]). As Norman *et al.* [17] point out, "*Animals have developed sophisticated mechanisms for surviving in an unpredictable, dynamic world, coupling the appraisals and evaluations of affect to methods for modulating the overall system. The result is increased robustness and error tolerance*." Work on studying brain-damaged patients [20] has revealed the necessity of emotions in decision-making in humans.

Some supporters also view emotion as adding to an organism's behavioural flexibility, increasing its ability to address multiple concerns in intelligent and efficient ways, providing the ability to determine event salience, constructively influencing 'low-level' perceptual processes, and coordinating adaptive behaviour in the face of challenges. While, in the early years a number of autonomous robotic systems followed this line of thinking in implementing emotions in machines (such as [21]), Sloman and Croucher [22] argue even more strongly that sufficiently intelligent robots would automatically have emotions similar to those of humans.

This necessitates clarifying what exactly *is* an emotion. An emotion to a particular evocative stimulus is difficult to define specifically as it is a subjective experience, but it is clear when one has felt an emotion. This promotes the popular rather vague definition of emotion as a *state of feeling*. In psychology, emotion is viewed as the feelings about a situation, person, or objects that involve changes in physiological arousal and cognition. Kagan uses the metaphor of the weather to help understand emotions and its complexity [23]. Weather is the collective term for changing relations between wind velocity, humidity, temperature, and air pressure. Occasionally unique combinations of these produce temporary but intense situations (e.g. a tornado or blizzard) which are analogous to temporary but intense emotions such as fear, joy, excitement, disgust, or anger. This process is continuous, with short and long term effects. Meteorologists do not look to define weather *per se*, but rather look to understand the relationships between measurable quantities. Appraisal Theory [24, 25] effectively adopts a similar approach.

The notion of artificial emotions has begun to be implemented at different levels in machines in the environment around us. Their implementation also gives rise to numerous research questions and questions about the motivation for developing artificial emotions in the first place.

DECOMPOSING THE EMOTIONAL MACHINE

In developing cognitive affective architectures, we can look to develop reasoning believability which can then employ the appropriate aesthetic in its delivery. While, from a practical perspective, a separation between perceived behaviour and internal behavioural mechanisms in social robot research often exists (with anthropomorphism being a primary cause), it is necessary to achieve a direct coherence between, for example, gesture interpretations and the intended emotional content to be communicated. This is necessary for the development and maintenance of the social context. Another problem may arise in the pursuit of a particular behavioural aesthetic (believability – see also section 6) for successful social communication and the difficulty in implementing this through modular finite-state machines. Successful social interaction is a dynamic process which necessitates coherent behavioural plausibility and the capacity for fast interactive responses in order to maintain the social bond. The system must experience and be a part of the experience in order to coherently negotiate the situation, that is, it must be affectively embodied. Affective embodiment involves a similar approach to physical [26, 27] and social embodiment [28] in the sense of the pursuit of a strong notion of physical and social context in how the system behaves and is designed. While it can be argued that affective embodiment should be classified as a subset of social embodiment as it inherently refers to social information, the non-social aspect of emotional stimuli (fear of a snake, for example) differentiates aspects of affective embodiment from more global concept of social embodiment. Affective embodiment has a strong a sense of emotional context for evaluating the emotional social and physical experience. It incorporates appraisal approaches found in affective computing which aim to manage the system's integration into its affective environment, some of which are detailed in the following section.

In striving to integrate these emotional approaches into a machines system design, there are three basic stages, each of which has its own set of issues and complications: input, reasoning and output. While by no means independent from each other, the following section adopts this conceptual categorisation in order to highlight fundamental features particular to each.

INPUT: Emotion Sensing and Recognition

Emotion perception and recognition by machines has developed significantly over recent years. Picard conducted extensive emotion recognition studies with participants equipped with electromyogram (EMG), skin conductance, blood volume pulse, and respiration sensors [29]. Results indicated 81% classification accuracy for eight emotions but also highlighted that it would be wrong to conclude that a

computer can recognize people's emotions with 81% accuracy [30]. Key problems arise in achieving emotion evaluation in a continuous, person-independent and natural interaction situation – analogous to the problems in speech recognition. In addition, emotions comprise both discrete and continuous features which, given their complexity, are computationally difficult to negotiate. Physiological, nonverbal and paraverbal modalities have also been used to sense user affect, for example facial expression [31-33], posture and body language [34, 35], galvanic skin response (GSR) [36], and speech [37]. The fusion of multiple modalities reinforces the systems assessment of human emotions [38]. (See also [39] for a review on emotion elicitation and recognition)

Emotional assessment strategies alternative to those obviously used by humans (such as galvanic skin response, heart rate, skin temperature [39], and infra-red [40]) may help manage the difficulties in machine-based emotional state assessment. A question worth considering is whether, given that these modalities rely on frames of reference often different to those used in human-human social interaction, their fallibility remains an important issue.

PROCESSING: Emotional Reasoning

In order to implement emotional reasoning mechanisms on a machine, it is necessary to define such features as range, intensity and valence, the operations which can change these features, a process through which the system can move from one state to another, and an association mechanism between emotional and non-emotional information. This approach adopts an inherently deterministic strategy and interpretation of what emotions are. Can emotion be sufficiently decomposed into workable constituents or elements in order to be implementable within a machine's reasoning process in such a way?

The physical and social environment is not deterministic. Research on embodiment in AI [26-28] has highlighted the inherent complexities associated with developing and maintaining a robust environmental context. Our ultimate frame of reference for control remains the human reasoning system in its entirety, where emotions play a fundamental role. It is clearly inappropriate to develop an artificially intelligent social system which is not endowed with a capacity to reason emotionally and also understand the emotional content present in social interaction. How should a computer deal with "understanding" its social interaction when there is ambiguity in how to interpret emotions? This requires strong learning mechanisms and coherent social sensor fusion. On the other hand, as a machine fundamentally functions as a deterministic device, serious issues arise as to how to navigate this dichotomy, i.e. the emotional human *vs* the cold machine.

It is argued that emotions in artificial systems seek to "pollute" structured logic reasoning systems with fuzziness. But as data is inherently fuzzy in the real world, incorporating a coherent control architecture that has the capacity to deal with fuzzy information remains the ideal approach. Problems arise when we intend to implement this through 1's and 0's. If there is a means whereby it can be more transparent how the machine deals with the unknowns and fuzziness of the real world, then maybe our acceptance of the use of emotions in artificial systems could improve. The key is to coherently integrate affective approaches into the control system. More recently, researchers in affective computing have attempted to add emotions de facto on top of existing architectures, (whether through reactive, behavioural, or deliberative levels of the system). A more integrative approach is required where a coherent synthesis between emotions and all aspects of the systems control is necessary.

There is also an issue with the possible disassociation between the internal affective state of the machine and the mechanisms required to express this state to an observer. A central tenet of appraisal theory [41] is that in determining the resulting emotion, it is not necessarily the event itself, but rather the perceived outcomes for the individual that evokes a response. The emotion evaluation is based on a consequential interpretation or evaluation of the situation. Appraisal refers to the relationship between a person and their physical and social environment (affective embodiment) and, as pointed out by Lazarus [24, 25], is not a one step process but rather involves subsequent re-appraisals in order to correct or re-evaluate the situation based on new information, often through a cyclic process. Coping draws on the resources available to the system in order to maintain or update this relationship (how affective embodiment is achieved). Scherer's Sequential Evaluation Check (SEC) [42] offers a similar strategy for maintaining this affective context, which involves processing in sequence, following a fixed order.

A considerable number of emotion-based computational frameworks and artificial emotion mechanisms have been developed, including [43-49]. For human-machine interaction, such machines will require a means to express emotions.

OUTPUT: Emotion Synthesis

Both thoughts and emotions can trigger actions visible to others: gestures, facial movements, head nods, shifts in posture, and other physical actions relating both to cognition and affect. Affective signals can be displayed using a range of modalities including speech [50], facial expression [51], motion dynamics [52], and natural language text [53]. How realistic are these expression modalities?

The communication of effectively fake machine emotions to a user may lead to a false sense of attachment. It is also questionable whether this could be maintained over time. One would like to think not. The success of Sony's AIBO (1999) draws on an owner's affection for a pet dog, where the goal can be to even mislead the human in their evaluation of the intelligence level of the robot [54].

While Picard [30] highlights that "(e)xisting models of emotion use highly stylized stereotypes of personality types and emotional responsiveness, which do not correspond to real behaviour in real people", this in fact produces recognisable states, such as found in theatre performances

(where exaggeration is key to combating misinterpretation). Work in virtual environments ranges from iconic stylised virtual characters to more life-like human models in plausible emotional scenarios for simulation [15, 55-57]. The growth in emotion-like expression in robots similarly adopts these strategies such as [14, 58, 59], also with varying degrees of success.

Due to inherent sensor noise and interpretation errors, a machine cannot gain a perfect model of its physical and social space (nor, ideally, should it need to). If the perceptual information is not structured in the form of a fact, but more as a belief, then using emotion could help reinforce our acceptance of the information communicated to us by a machine (we sympathise with its problems). If we are more aware of the machine's difficulties with its environmental representations and internal knowledge structures, we may be more accepting of the use of emotional interaction mechanisms. This also supports a complete systems approach where emotions must be implemented at the reasoning level of the control system and not simply be an add-on to support communication.

PARADOXES OF THE AFFECTIVE SOCIAL MACHINE

The arguments for and against the development of artificial emotion at all levels of the intelligent machine, as outlined in the previous section, will continue to fuel many animated and passionate debates. In further analysing these issues, some fundamental paradoxes become apparent in the field of affective intelligent machines.

Paradox 1: The Truth Factor

Emotions have been argued as necessary to facilitate human-machine interaction by drawing on traditional human-human interaction cues. On the other hand, an inherent advantage of a machine providing responses to a particular query is the general assumption that the information is factual and impartial. The apparent perception of impartiality of machine information removes ambiguity and isolates the user from any emotional "noise" in the interaction. Users rely on the idea that computers cannot become hindered by emotions and generally believe that as the information came from a computer, it is true in so far as the computer can provide the answer.

In introducing emotional reasoning and in particular the synthesis of artificial emotions with a view to facilitating the interaction between human and machine, this affective augmentation in fact interferes with the perceived truth value of the information, particularly if there are serious issues about the machine simply faking an inherently human feature. Basically, if the machine employs emotions in order to reinforce what it is trying to communicate given our propensity to anthropomorphise (and even a bias towards anthropomorphising [60, 61]), this could in fact work against the system's objectives because we still perceive it as a machine (with perceived fake emotions). Anthropomorphism is not without its own complications (see [62] for a discussion). Once we switch from a clear perception of machines as cold, logical, and factual (in as far as it has the capacity to map some inputs to a corresponding correct output) to a machine with apparent human-like emotions, anthropomorphism encourages us to extrapolate from the a machine's set of perceived emotional attributes to the wider set of negative human emotions which may act to hinder the truth value of information it provides. The fact that you know it is a machine means it shouldn't lie to you, but because it acts like a human, it opens the possibility that it could be lying to you.

Paradox 2: Efficiency for the Sake of Interaction

Would the implementation of artificial emotions at every level of the artificial system's structure (from input through reasoning to output mechanisms), with the objective of making the system more socially capable for human interaction, in fact result in the antithesis of clear, comprehensible and predictable systems? That is, could the objectives of affective computing in fact be contrary to the very mechanisms it is trying to implement?

Implementing emotional theories and models on machines on the one hand intuitively supports the development of human machine interaction, where people have the strong capacity to anthropomorphise the interaction in the first place [60, 61]. On the other, it can be argued that providing yet another degree of fuzziness to the interaction could only exacerbate the already complicated issue of human and intelligent machine interaction. Could our frustration with machines, often based on the inflexibility of the control mechanisms, be alleviated – or indeed exacerbated – through the integration of more emotional interaction paradigms? Could this manage the issue that the computer is never wrong, that it is just the user who doesn't know how to use the machine correctly?

Paradox 3: The Cold Truth *vs* the Hot Head

One of the arguments for incorporating emotions in machine reasoning is to embrace the advantages provided by emotions in human decision making, such as the ability to manage overwhelming or incomplete information [10, 20, 63-65]. However, emotions can both facilitate and confuse the decision-making process. Rational emotional decision making is a paradox; how can an emotional decision be rational and vice versa? There is the old adage that too little emotion, like too much, triggers bad decisions. The issue is therefore to manage the degree of emotion involved.

ETHICAL PROBLEMS WITH AFFECTIVE TECHNOLOGIES

In addition to these apparent contradictions, affective computing and emotional machines encounters an additional number of ethical problems. Its proximity to core aspects of what we hold on to as the difference between humans and machines often encourages a hesitancy in accepting the idea of an emotional machine. It may make us feel uneasy. The following section looks to discuss the problems often associated with affective computing with a view to providing an open forum of discussion within this promising field.

Ethics

Are researchers in affective social computing supposed to be aware of potential ethical issues which may result from their discoveries? Or can they continue developing technologies without worrying about ethics, as with many technological developments and in so doing, redirect the ethical issues towards how these discoveries should be used, i.e. the standard ethics of science? This is similar to asking whether researchers are required to study the potential harmful uses of technology before developing it, a difficult task to say the least. We as humans are all generally integrated into a society that teaches us morals, ethics, and limits on the actions we take as a result of our emotions.

In 1950 Wiener laid down a comprehensive foundation which remains today a powerful basis for computer ethics research and analysis (although the term "computer ethics" only came into common use two decades later) [66]. In his view, the integration of the computer into everyday society would mark the "second industrial revolution", a remaking of human society. Clearly the role and effects of computers in society today reinforces this. Consequently, the impact and ethical implications of such technology should always be reviewed. Machine (or computer) ethics can be divided into two perspectives:

- The issues of building intelligent affective machines and what they may be able to do, i.e. the ethical issues associated with their use and misuse.
- Implementing moral decision-making capabilities in machines, i.e. in the form of an artificial ethical reasoner.

This discussion focuses on the first point. Ethics is not computationally tractable (analogous to intelligence) and often revolves around difficult social problems with many degrees of complexity and unknowns. The second view of machine ethics can be further decomposed into artificial systems developed to act as pure ethical reasoners, required to solve our ethical dilemmas (a computationally intractable problem), and artificial agents which incorporate ethical capabilities as exemplified in Artificial Moral Agents [67]. It is the view here that a machine should be solely employed as an assistant for a human required to address complex ethical problems – i.e. SCIROCCO [68] - and not act as a pure ethical reasoner regarding human-human ethical problems. The inclusion of ethical features into a reasoning agent could, on the other hand, provide useful mechanisms for social modelling in human-machine social interaction.

The ethical issues afforded by the specific field of affective computing and human machine interaction have fortunately been a point of discussion as the field develops [10, 30, 69-71]. Questions exist such as why should one build affective social systems or if you could build it what would/should it do? These questions are not special to affective social computing, but as science has continually demonstrated, these ethical considerations recur in many disciplines.

While it has often been impossible to achieve accurate predictions about the uses of a specific technology (chicle was originally intended for rubber products, before accidentally being discovered as a key ingredient in chewing gum), looking to assess the implications of those uses is even more difficult. But, is there something particularly different about affective computing research compared to other technological developments that justifies taking a different approach? If so what kind of different approach can be envisioned? For example, when developing intelligent autonomous affective robots, would it be appropriate to encode laws of robot conduct into the system? The science fiction writer Isaac Asimov [72] proposed the three laws of Robotics in 1942 to protect mankind from robots. People often misinterpret these relatively intuitive laws as sufficient guidelines where, in fact, many of Asimov's short stories based around these laws were grounded on their fallibility and inadequacies (see also [73]).

In contrast to laws, ethics may provide a more guideline-based approach than rules and their associated hard-line inflexibilities which can leave them susceptible to implementation difficulties, with emotions possibly providing a useful reference [74]. On the other hand, Minsky [75] discusses the failures of ethical systems where "*... our rights to have children, to change our genes, and to die if we so wish. No popular ethical system yet, be it humanist or religion-based, has shown itself able to face the challenges that already confront us. How many people should occupy Earth? What sorts of people should they be? How should we share the available space?*"

The ethical issues associated with the use and misuse of artificial emotional mechanisms in machines, particularly humanoid robots, are no different to standard ethical concerns when a technology inappropriately used to mislead. Unfortunately, the basic tenet of integrating such emotional concepts in a machine is to mislead, *because* it is a machine.

Privacy

To date, our interaction with machines still retains a degree of social distance as the system is generally only aware of personal information provided to it. Its invasion of our privacy is minimal in this case. Once a machine becomes capable of reading our affective state and being able to evaluate its actions relative to this assessment, our willingness to allow a machine such privileges may rapidly diminish. Such an intrusion of personal (emotional) space extends to a number of additional issues such as privacy and evaluation complexity:

Privacy issues for emotion data retrieval, storage and dissemination. This is a general issue that is not specific to emotional interaction with machines.

- Incorrect evaluation and its consequences. The complexity of evaluating the human emotional state remains a significant challenge.

Emotion recognition systems provide information which surpasses many established privacy boundaries with which we have learned, in general, to be comfortable. User-modelling based on affect provides access to such basic emotional information as love, hate, and fear for example, which can strongly influence or even over-ride rational

behaviour, behaviour that would generally correspond to the societal ground rules we adhere to (albeit this can differ between cultures, e.g. USA *vs* Europe *vs* Japan). The issue of privacy has developed significantly in recent years with the ease of flow of information through the Internet. The development of affective interface technologies will necessitate a further evolution of an already adaptive process to privacy protection with corresponding legal ramifications.

Emotional Manipulation

Will users *know* that the system doesn't actually "feel" emotion? Do they care? Do they feel like they have been cheated/violated by the fake emotional expression? If the system is a fake, how can it be interpreted the way it was intended? People may not like to be "duped" into thinking that a machine is a person, or that it can behave like one. One of the fundamental issues in the development of emotionally expressive machines is therefore the need for a balance between perceived function and form (anthropomorphism) [62].

Would we accept a machine employing the power of human-like emotional reactions to a particular stimulus? If the machine cries out as if it was "afraid" that it would fall and a human reacted to this primitive-like cry (through an instinctual reaction), the person can feel cheated when it was only a machine that cried out. A more extreme example is where the machine expressed love towards the user when the machine does not "feel". Affective machines play on our emotions, and often play on our reacting to something that may be dependent on us (building relationships based on social need). It is unclear whether such situations could have negative impacts on human-human emotional interaction where it could subsequently desensitize us to such emotional situations.

A clear analogy of this issue is apparent in commercial marketing, which monitors and targets the emotions of consumers through presenting fabricated opinions and emotions. This opens a range of new dangers concerning personal integrity. The domain of advertising already has ethical (and legal) restrictions on how to manipulate audience intentions. Affective machines may draw directly on similar restrictions when looking to modify a user's affective state through private social engagement. Should this interaction be two-way, its power may become significantly more important. For example, while it can be easy for a user to display affect towards the artefact, when the artefact is perceived as demonstrating affect towards the user, the social bond becomes reinforced. This relies on the capacity of persuasion and believability (discussed in section 6).

The role of such machines in elderly health care, a topical issue in a number of countries worldwide, remains controversial (as recently discussed at [76]). If sufficient man-power is seen as too expensive or inefficient to address the problem of loneliness and isolation in elderly healthcare, could we justify replacing human care with a robot? The answer is not a straightforward yes. The answer should be to re-assess the question. The use of a machine as an artificial substitute for the issue of health-induced effects of loneliness and isolation in the elderly may satisfy economic constraints and temporary relief, but ultimately reinforces the necessity for human-human social interaction in these conditions which is irreplaceable.

Replacing Humans with Machines

The fear of machines replacing humans dates back to the industrial revolution typified by Joseph-Marie Jacquard's (1752-1834) invention of the loom in 1801 in Lyon, France. At first Jacquard's looms were destroyed by silk weavers who feared unemployment. Jacquard replied by using a donkey to drive the loom to weave garments in a public square to demonstrate that it was as "smart" as the workers. It sparked the Luddite movement. Since then, history has shown the evolving roles of mankind with the progress of technology, and the often unwarranted fear associated with technological advancement. Our roles change as much as the machines do. While Ludism continues to categorically oppose technology, both in particular and in general, our continued progress from the age of the wheel to present day testifies to its inevitability. We are merely *transferred* into different roles in society, not necessarily *replaced*.

Robots in Our Personal Space

The inevitability of technological progress reflects our dependency on it. There are two aspects to this: our dependency on machines and the feeling we have when they are dependent on us. The second is related to emotional manipulation (as discussed in section 5.3). The success of the Tamagotchi (Bandai, 2000) and its many derivatives is based on our associations with dependency (the needs pets may have that we would fulfil) and the attraction of perceived unpredictability (not knowing when the pet will be hungry, sleepy, or in a playful mood) yet reliability (if we played with it, it will be happy). An artificial system like the Tamagotchi is designed to constantly vie for our attention through some "feeding" or "loving" mechanisms. Pressure is placed on the person playing with it to provide for its needs in order that it will grow and finally become an "adult", ultimately with its "secret character" appearing if it has been "raised" perfectly. They were designed to try to realise a sense of satisfaction and achievement. Our interactions with the device develop through this artificial notion of dependency and social action and reaction. In the case of the Tamagotchi, users are not "tricked" into thinking that the device is a real pet with real needs. Its success is in its ability to entice users into a captivating scenario. The scenario is captivating precisely because it plays on the user's sense of satisfaction of being depended on. This harks back to the issue of emotional manipulation. The fear of emotional manipulation is reduced here when the artificiality is obvious and the user knowingly, voluntarily, and even seeks to participate in these artificial scenarios. The danger still remains in that if in our propensity to anthropomorphise, we forget about this artificiality as our relationship with the machine deepens, will we feel eventually emotionally manipulated over time? Degrees of frustration can develop if the user believes they are following the "rules" in growing the artificial "life" but success is not achieved. Here the

source of the frustration is associated with returning to the perception of the system as being artificial where rules are inherent and simply following these rules *should* be enough.

Other examples of playing on our dependency is in the form of artificial pets, encouraged by the medical community recognising the significant correlation between health and interpersonal aspects of an individual's life. Robotics research has looked to take the artificial pet notion further, such as Sony's AIBO robot dog. Work on the use of robot pets in the field of elderly health care [77] and the associated dependency issues highlight the roles a machine may undertake. This also clarifies the advantages in using such technologies to highlight the underlying problems, in this case the issue of stimulating positive emotions in the elderly. It highlights the problem, but should not be seen as a solution to it.

Another concern of integrating robots in our personal space is that if the machine has the capacity to be rational, autonomous, and perceive and exhibit emotional behaviour, how much control do we have over it? Once the system becomes more complex, its development or "evolution" over time may result in our having increasing difficulty in explaining its behaviour. As often sensationally discussed in science fiction, there is the fear that robots may run out of control and affect the "status quo" between man and machine. In fact the details are much more subtle. "*If computers have become our constant companions at home and in the workplace, their nagging side effects — dulled vision and a stunted attention span — are the psychic equivalent of a spam assault: chronic annoyances*" [78]. This raises an interesting point about machines as constant companions, notably, what are the health and environmental drawbacks of prolonged exposure to such machines, and can we anticipate/address them? While this necessitates the rigorous evaluation of such systems, our tolerance (and intolerance) will provide insights as the technology develops.

Can a Machine Tell a Perfect Lie?

A machine's inability to lie would be seen as a core feature in the design of the system. While people do not have full control over their emotions, a machine, by having complete control over its expressive capabilities, could lie with a straight face. To illustrate, the idea of a "poker-face" is a valuable trait in playing the game of poker where the goal is to transmit as little information as possible about one's hand of cards to the other players. A robot would clearly be a master of expressionless poker playing.

Developing affective social machines should therefore adopt a complete systems approach where emotions *must* be implemented at the reasoning level of the control system and not simply be an add-on to support communication.

RESOLVING THE ISSUES

Despite the problems and issues which arise in affective computing, emotion plays an important role in moderating the social interaction process. The social interaction process between man and machine relies on the development and maintenance of a mutual understanding of some form. Emotion can contribute as an important moderator within this process. The following sections look to provide useful frames of reference for social affective intelligent machines.

The Importance of Emotion

From a business and productivity perspective, easy arguments can be found for not incorporating emotions in artificial systems and not using emotional or conscious robotics. It is often the very lack of human qualities that makes robots desirable in the workplace. As clarified by [79]:

"Relaxations of environmental standards can provide substantial savings for the firm that substitutes robots for human labor. Robots can work around the clock and do not receive differential pay for night work; they are not prone to go on strike or to display other forms of labor unrest; they occasionally break down, but never get sick and are immune to the effects of drugs and alcohol; and they are readily available for work at unpleasant tasks for which human laborers are difficult to recruit."

This clearly trivialises the recognised important role that emotion plays in human-human social interaction and intelligence in the workplace.

It is clear that the primary practical use for an emotional robot is in the social interaction between robots and humans. The advent of the social and affective perspective in recent robot design reinforces the altering of the non-human functional only perspective to include those mechanisms that may prove useful where explicit man-machine interaction is required. The research of Salovey and Mayer [80], Damasio [20], Ledoux [81], Isen [82], Goleman [63] and many others demonstrate the vital roles that emotion plays in social interaction and in addition to many human processes: perception, decision-making, creativity, empathic understanding, and memory.

Artificial emotional mechanisms [44-46] and others discussed throughout this paper can guide the social interaction in

(i) Maintaining a history of past experiences (categorising experiences, i.e. memory organisation),

(ii) Underlying the mechanism for instinctual reactions,

(iii) Influencing decision making, especially when no known unique solution exists,

(iv) Altering tones of communication as appropriate (as seen in characters in [83]),

(v) Intensifying and solidifying its relationships with human participants.

The judicious selection of artificial emotion generation techniques should facilitate the social interaction and not complicate the interaction. Here the use of an iconic derivative of the Facial Action Coding System (FACS) [84] approach could provide a standardised protocol for emotional communication for establishing stronger social relationships.

The Reference Revisited

Creating machines that *speak our language* and *perceive our emotions* in the sense that we can easily understand and interact with it (according to very human-like references) is a fascinating challenge. Autonomous systems research, and specifically robotics, has now developed to a stage where complete complex systems are robustly possible which coherently integrate perception, action and control on an autonomous mobile platform. The problems are by no means "solved", but workable solutions for realising social autonomous machines exist (at least to a certain extent). But where does the designer stop without becoming trapped by the inimitable task of managing expectation on a very human-like robot? (See [62]). Anthropomorphism and affective computing is only useful if it does not overly complicate people's expectations.

Perhaps the hypothesis of this pure humanoid being the ultimate frame of reference is too restrictive. What is necessary is an understanding of the core issues facing us with the advent of affective social machines and a set of constructive guidelines. These core issues subsume the human as the ultimate frame of reference in affective social interaction and extend those discussed in the context of anthropomorphism [62]:

Control 1: the ability to influence the robot, or explicitly influence the environment through it.

Control 2: the inability of the robot to exploit, or explicitly influence the human through the use of affective information (emotion-based marketing being a clear example).

Predictability: but not necessarily overly simplistic for the sake of being very predictable. At the same time it should not be too unpredictable or the "bond" we have with the robot can become difficult to sustain.

Dependency: respects our needs, whether utilitarian, or our needs for satisfaction and emotional attachment.

When a robot is a tool, its role and function is clear. We control it to perform some behaviour, and it is judged according to its ability and reliability (predictability) to succeed in defined tasks (needs). Here, the purist perspective of the human form, with all its complexity, is of little use [85]. It is no more a tool than a hammer, just more sophisticated.

Similarly, the integration of the robot into a social context, even without a necessarily defined task to be achieved, places an emphasis on reliability (based on social conventions or social predictability) to succeed in establishing a relationship (needs). Here a balance between the robot's function and form becomes crucial. The success of our relationship depends on our expectations, which is based on our assessment of its capabilities, which in turn is directly related to its form and behaviours.

The Human Balance

Reviewing robots through history, the most successful machines are non-humanoid which do not replicate humanlike abilities, a fact demonstrated by the first industrial robot which began work at General Motors, the UNIMATE. Today, the success of industrial robots is firmly grounded in their ability to perform tasks that humans are inherently not good at. While not necessarily renowned for its ability to adapt, improvise, or draw reliable conclusions based on incomplete information, an industrial robot's precision, speed, and strength are what ensures its success. Today's automated factories are designed primarily for robot efficiency with limited concessions to a human-centred environmental layout.

Deken [86] describes robot autonomy as "the hallmark that would elevate robots to the status of a species". He continues by saying that:

"*Whatever forms our future robots take, we cannot let those forms induce us to accept the anthropomorphism that robots are our equals. Robots are not like humans nor worthy of the same considerations, no matter how sophisticated they become. They are our tools and creations, to be kept in place as a subservient species by whatever methods we find necessary.*"

This point of view becomes difficult should the goal of design and engineering teams be to assimilate human function and form, including the capacity for emotion, in a machine. If an artificial entity is designed to achieve a stronger social bond with humans while also being treated as a tool to perform tasks, the issue of slavery may arise and the reintroduction of a concept which civilisation works hard to abolish. This is not necessarily dependent on when technology advances to the stage where it is difficult to distinguish between man and an android, but rather much sooner, when the goal is to integrate a socially capable machine as efficiently as possible into the human social sphere. It is necessary to achieve a balance between a number of factors.

Fig. (**2**) illustrate two different strategies in developing human-centric socially interactive machines. The first, Phillips iCat [58], is an example where constrained emotion expression modalities control the interaction. The second, while very impressive, has more difficult acceptance issues as outlined by Mori's Uncanny Valley in Fig. (**3**).

Fig. (2). (**a**). Philips iCat demonstrating six iconic expressions [58]; (**b**) Humanoid robot modelled on a known Japanese television presenter [87].

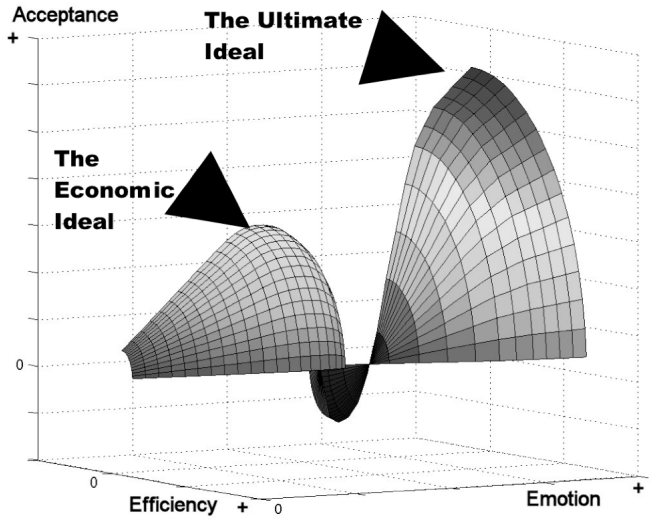

Fig. (3). A 3-D version of the Uncanny Valley showing acceptance *vs* efficiency *vs* emotion (inspired by [1]).

The 3-D Uncanny Valley (inspired by Mori [1]) in Fig. (**3**) looks to intuitively plot a balance between efficiency, emotion, and one's acceptance of the system. The economic ideal constitutes the first region of optimality between the most socially capable artificial system with the accepted technology norms that consumers are used to – at a reasonable cost. The ultimate ideal represents the perfect artificial human. Even though the economic ideal will inevitably move closer to the ultimate, a chasm will ultimately persist. Whether it is possible to traverse this chasm does not negate the argument as to whether the system will remain a machine. On the contrary, it is the advancement of the machine to being a perfect fake that prevails.

Managing Believability

Endowing artificial agents with believable, life-like qualities has been the subject of considerable research in recent years [88-93]). The pursuit of machine autonomy may pose difficulties in maintaining this illusion over time.

To manage the difficulties of achieving believability in artificial systems, the following poses a number of important guidelines (based on work on animated pedagogical agents [89]):

- Any behaviours that interfere with the task at hand, no matter how much these behaviours might contribute to believability, would be inappropriate.
- Believability-enhancing behaviours must complement (and somehow be dynamically interleaved with) the behaviours that the agents perform.
- If observers see that an agent is acting like a simple automaton, believability is either substantially diminished or eliminated altogether.

Arguably the single most important feature in creating the illusion of life in artificial entities is emotion. From an animation perspective, this is highlighted in "The illusion of life: Disney animation" [88]:

- The emotional state of the character must be clearly defined
- The thought process reveals the feeling
- Accentuate the emotion by using time wisely in order to establish the emotion, to convey it to the viewers, and to let them savour the situation.

In the Oz Project [90], the intention has been to try to implement these maxims and create believable agents. Results promote the use of behavioural quirks in addition to these mechanisms in order to realise a character that gains ones attention whilst also providing an active mechanism against behavioural repetition and monotony, and hence so-called "lifelessness".

Useful Constraints

Familiar expressions based on iconic representations can convey a strong degree of social affective communication. Cartoon-like aesthetics and behavioural functionality, with their corresponding limitations, are widely accepted as social mechanisms between characters and their audience. The success of the emotional dimension found in these characters can be attributed to their constrained emotional expression through iconic features, which manages interpretation through resolution-based ambiguity control. Basically, the observers' expectations of the artificial cartoon character is defined by how it *looks* and *acts*, rather than necessarily using the full feature complexity of the human frame of reference.

A conversational character "Laura" was designed to build and maintain a long-term social-emotional relationship with users who were undergoing a month-long program to increase their level of fitness [94]. One important feature in its success was the agent being up front about its limitations. Results indicate that users rated the relational agent significantly high on likeability, trust, respect, feelings it cared for them, and willingness to continue interacting with it. This provides an example where the use of agents with defined limited affective capabilities, and a clear boundary of competence can lead to improved quality of HCI experiences. This reinforces discussions on managing expectation through controlled anthropomorphism [62]. Intelligent tutoring in education is a similar domain where work to date has looked at incorporating affective mechanisms to assist in the learning process, such as in [95].

Emotion Utopia

Implementing emotion in a machine should not necessarily be implementing only the "good" emotions, a form of artificial emotion utopia. For instance, a machine that can express sadness will be able to exhibit engender sympathy and empathy, thus establishing a stronger relationship with the human user. A complete set of emotional capabilities in machines may also contribute to our *understanding* of the process of emotion-based social interaction between humans.

However, does it make sense to have a machine that has the capacity to "feel" and exhibit all human emotions? Will

certain emotions be detrimental to the man-machine relationship and the role of machines in our personal space (or our role in the machine space)? Maybe it is important for a machine to recognise violent anger but it could be unnecessary for it to exhibit it, even imperatively so. In order to define the list of taboo emotions (if appropriate), we must follow the first guideline listed in Section 6.4: "Any behaviours that interfere with the task at hand, no matter how much these behaviours might contribute to believability, would be inappropriate". Emotions should contribute constructively to the role of the machine in human society. This does not mean that anger should not be exhibited by a machine, but rather it should be appropriately associated with its role or task.

CONCLUSIONS

With the world robot population climbing quickly above the one million mark and the development of more robust technologies facilitating autonomous devices in our offices and home, the era of the robot is exploding.

Key to their success is their economic advantage and efficiency. If they can do a job better, quicker and more reliably, then they will most likely replace an existing technology (or person). Such useful robots often have little to do with the human form or function. However, such aspects of anthropomorphism intuitively provide mechanisms that can facilitate a robot's integration into our social space. A balanced design of form and function can be more successful than the literal anthropomorphic form alone. An abstraction away from the human allows a more over-reaching view of what the ultimate references for the future robot should be. The three tenets of control, predictability and dependency subsume the humanoid and other biological inspired frames of reference, and facilitate imaginative solutions for the robots of the future.

With the ever continuous advancement of technology, its use and misuse, the role of the affective intelligent social robot will evolve. While the development of an emotional social robot as a love companion (i.e. a robot gigolo as seen in the film A.I.) may be controversial, replacing the human completely with a machine for social and intimate interaction will most likely be generally rejected by society and will be an exception rather than the rule. The evolution of the strong social bond between people over the millennia testifies to that.

Picard's [30] argument that "*(j)ust because every living intelligent system we know of has emotion does not mean that intelligence requires emotion*" conflicts with the Machiavellian Intelligence Hypothesis (Social Intelligence Hypothesis) which strongly grounds intelligence in social interaction. The role of emotion in managing social interaction is fundamental, for example in providing a mechanism for dealing with the overwhelming complexity of social interactions and our physical environment.

It has been argued that emotions are one of the last features that define us as being human. This presupposes that we have already solved the "intelligence problem", which clearly we have not. There is the old adage in the world of business that if you can't measure it, you can't manage it. Arguably, emotion may be one such situation. If one manages to successfully implement it on a machine, then our understanding of it would be of a sufficient degree to make it more manageable.

Humanity was forced to reassess its assumptions about technology during the Industrial Revolution, when machines developed from enhancing human abilities through its role as a tool, to exceeding and ultimately replacing them (as discussed in section 5.4). Today's devices are still forms of assistive devices, inherently ruled by logic. "Emotion machines" could bring an end to the clear cut division between man and machine, and evoke a new reassessment of what distinguishes man from other life (and machines - if machines become considered alive).

This paper has looked to continue the discussion encouraged by Joy in [96] to not be "*predisposed to fear or favour technology for its own sake*". While cultural differences exist in the perception of robots, it is clear that fiction tames technology [97]. While aspects of what we envisage the future to be appears daunting, there is nothing new in this. Humanity has a robust ability to both adapt to and recalibrate unforeseen developments that arise. We should have more faith in ourselves.

REFERENCES

[1] M. Mori, "The Buddha in the Robot". *Charles E. Tuttle Co.,* 1982.

[2] D. Adams, The Hitchhiker's Guide to the Galaxy, *Del Rey*, Reissue Edition, September 1995.

[3] P. K. Dick, "Do Androids Dream of Electric Sheep?", *Del. Rey,* Reissue edition, 1996.

[4] Asimov, "Bicentennial Man", *Orion Publishing Company,* Reissue edition, 2000.

[5] W. James, "What is an emotion?" *Mind,* vol. 9, pp. 188-205, 1884.

[6] W.B. Cannon, "The James-Lange theory of emotion: A critical examination and an alternative theory", *Am. J. Psychol.,* vol. 39, pp. 10-124, 1927.

[7] D. Derryberry, and D. M. Tucker, "Neural mechanisms of emotion", *J. Consult. Clin. Psychol.,* vol. 60, pp. 329-338, 1992.

[8] F. Michaud, P. Prijanian, J. Audet, and D. Letourneau. "Artificial emotion and social robotics", *Distributed Autonomous Robotic Systems*, Springer 2000: 120-130.

[9] R. Plutchik, "*A General Psychoevolutionary Theory of Emotion*". Emotion: Theory, Research, and Experience, Academic Press, vol. 1, pp. 3-33, 1980.

[10] R. W. Picard, *Affective Computing*, MIT Press, 1997.

[11] J. McCarthy, "*The Robot and the Baby*", [Online] http://www-formal.stanford.edu/jmc/robotandbaby.html, [Accessed Dec. 18, 2007].

[12] J. McCarthy, "Making robots Conscious of their Mental States", *Machine Intelligence 15 workshop*, Oxford University 1995.

[13] J. Weizenbaum, "Computer Power and Human Reason: From Judgment to Calculation", *Freeman,* 1976.

[14] C. Breazeal, "Infant-Like Social Interactions between a Robot and a Human Caretaker", *Adaptive Behav. (Special Issue on Simulation Models of Social Agents),* 1998.

[15] J. Gratch, and S. Marsella, "Evaluating a computational model of emotion", *J. Autonom. Agents Multi-Agent Syst. (Special issue on the best of AAMAS),* vol. 11, 2004.

[16] D. Norman, "Interview", *Ubiquity*, vol. 13, pp. 23-43, May 2002.

[17] D. A. Norman, A. Ortony, and D.M. Russell, "Affect and machine design: lessons for the development of autonomous machines", *IBM Syst. J.,* vol. 42, 38-44, 2003.

[18] B. P. Thagard, "Emotional Decisions", *Proc. Eighteenth Ann. Conf. Cognit. Sci. Soc.*, Erlbaum, pp. 426-429, 1996.

[19] H. A. Simon, "Motivational and emotional controls of cognition", *Psychol. Rev.,* vol. 74, pp. 29-39, 1967.

[20] Damasio, Descartes' *Error: Emotion, Reason, and the Human Brain*, Gosset/Putnam Press: New York, NY 1994.

[21] M. Scheutz, "Affective Action Selection and Behavior Arbitration for Autonomous Robots", *Proc. Int. Conf. Artif. Intel. (ICAI)*, 2002.

[22] S. M. Croucher, "Why robots will have emotions", *Proc. 7th Int. Joint Conf. Artif. Intel.*, Morgan-Kaufman, 1981.

[23] J. Kagan, *The Nature of the Child*. Basic Books, New York 1984.

[24] R.S. Lazarus, and S. Folkman, Stress, appraisal and coping, New York: Springer 1984.

[25] C.A. Smith, and R.S. Lazarus, "Appraisal components, core relational themes and the emotions", *Cognit. Emot.,* vol. 7, pp. 233-269, 1993.

[26] Clarke, Being There: *Putting Brain, Body, and World Together Again*, MIT Press, 1998.

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

[28] B. R. Duffy, "Social Embodiment in Autonomous Mobile Robotics", *Int. J. Adv. Robot. Syst.,* vol. 1, pp. 155-170, 2004.

[29] R. W. Picard, E. Vyzas, and J. Healey, "Toward Machine Emotional Intelligence: Analysis of Affective Physiological State", *IEEE Trans. Pattern Anal. Mach. Intel.*, vol. 23, October 2001.

[30] R. W. Picard, "Affective computing: Challenges", *Int. J. Hum. Comp. Stud.,* vol. 59, July 2003.

[31] P. Ekman, and W. V. Friesen, *Unmasking the face: A guide to recognising emotions from facial expressions*, Prentice-Hall, Englewood Cliffs: NJ, USA, 1975.

[32] Kapoor, R.W. Picard, "Real-Time, Fully Automatic Upper Facial Feature Tracking". *5th Int. Conf. Auto. Face Gesture Recog.,* Washington DC, 2002.

[33] L. Lisetti, and D. Schiano, "Facial Expression Recognition: Where Human-Computer Interaction, Artificial Intelligence, and Cognitive Science Intersect", *Pragmat. Cognit.*, vol. 8, pp. 185-235, 2000.

[34] Kapoor, S. Mota, and R.W. Picard, "Towards a Learning Companion that Recognizes Affect", *Am. Assoc. Artif. Intel. Conf.*, Falmouth MA, 2001.

[35] R. Birdwhistle, *Kinesics and Context: Essays on body motion and communication*, University of Pennsylvania Press, Philadelphia, Pa, USA, 1970.

[36] R. W. Picard, J. Scheirer, "The Galvactivator: A Glove that Senses and Communicates Skin Conductivity", *Proc. 9th Int. Conf. Hum. Comp. Interact.*, New Orleans, pp. 1538-1542, August 2001.

[37] K. Scherer, "Speech and emotional states", In J. Darby (Ed.), *Speech Eval. Psychiatry*, Grune & Stratton, pp. 189-220, 1981.

[38] M. Pantic, and L.J.M Rothkrantz, "Toward an affect-sensitive multimodal human-computer interaction", *Proc. IEEE,* vol. 91, pp. 1370-1390, 2003.

[39] L. F. Nasoz, "Using noninvasive wearable computers to recognise human emotions from physiological signals", *EURASIP J. Appl. Signal Proces.,* pp. 1672-1687, 2004.

[40] S. Kim, Y. Yoshitomi and T. Kitazoe, "Pattern Recognition of Emotional States using Voice, Face Image and Thermal Image of Face", *Proc. 5th Int. Symp. Artif. Life Robot.,* vol. 1, pp. 149-152, 2000.

[41] P.C. Ellsworth, and K.R. Scherer, "*Appraisal processes in emotion*", In R. J. Davidson, H. Goldsmith, and K. R. Scherer (Eds.) Handbook of Affective Sciences. New York and Oxford: Oxford University Press, 2003.

[42] K.R. Scherer, "*Appraisal considered as a process of multi-level sequential checking*", In K. R. Scherer, A. Schorr, & T. Johnstone (Eds.). Appraisal processes in emotion: Theory, Methods, Research, New York and Oxford, Oxford University Press, pp. 92-120, 2001.

[43] J. D. Velasquez, "A Computational Framework for Emotion-Based Control. Workshop on Grounding Emotions in Adaptive Systems", *Conf. Simulat. Adapt. Behav.,* 1998.

[44] R. Plutchik, A general psycho-evolutionary theory of emotion, in: R. Plutchik, H. Kellerman (Eds.), *Emotion: Theory, Research, and Experience*, Academic Press, New York, 1980; 1: 3-33.

[45] F. Michaud, E. Robichaud, and J. Audet, "*Using motives and artificial emotions for prolonged activity of a group of autonomous robots*", Emotional and Intelligent II: The Tangled Knot of Social Cognition, AAAI 2001 Fall Symposium, AAAI Technical Report FS-01-02, MA: 2001.

[46] L. Botelho, and H. Coelho, "Machinery for artificial emotions", *Cybern. Syst.,* vol. 32, pp. 465-506, 2001.

[47] R. Murphy, C. L. Lisetti, R. Tardif, L. Irish, and A. Gage. "Emotion-Based Control of Cooperating Heterogeneous Robots", *IEEE Transactions on Robotics and Automation* October 2002; 18 (5): 744-757.

[48] L. Lisetti, S. Brown, K. Alvarez, and A. Marpaung, "A Social Informatics Approach to Human-Robot Interaction with an Office Service Robot". *IEEE Trans. Syst. Man Cybern.,* (Special Issue on Human Robot Interaction), vol. 34, May 2004.

[49] A. Sloman and B.S. Logan, Evolvable architectures for Human-Like Minds, in Affective Minds, G. Hatano, N. Okada and H. Tanabe, Eds., *Elsevier*, Amsterdam, 2000.

[50] J. Cahn, "The Generation of Affect in Synthesized Speech", *J. Am. Voice I/O Soc.,* vol. 8, pp. 1-19, 1990.

[51] Pelachaud, N. Badler, and M. Steedman, "Generating Facial Expressions for Speech", *Cognit. Sci.,* 1994.

[52] C. Rose, B. Bodenheimer, and M. Cohen, "Verbs and Adverbs: Multidimensional motion interpolation using radial basis functions", *IEEE CGAA*, Fall, 1998.

[53] E. Hovy, "Putting Affect into Text", *8th Ann. Conf. Cognit. Sci. Soc.,* 1986.

[54] S. Park, E. Sharlin, Y. Kitamura, E. Lau, "Synthetic Personality in Robots and its Effect on Human-Robot Relationship", *Graphics Interface* 2005, Victoria, British Columbia, Canada, May 2005.

[55] C.D. Elliott, "The Affective Reasoner: A Process model of emotions in a multi-agent system". Ph.D. Thesis, The Institute for the Learning Sciences, Northwestern University, Evanston, Illinois, 1992.

[56] W.S.N. O'Reilly, "Believable Social and Emotional Agents", Ph.D. Thesis, Carnegie Mellon Univ., Pittsburgh, PA., 1996.

[57] D. Cañamero, "Modelling motivations and emotions as a basis for intelligent behaviour", *Proc. 1st Int. Conf. Autonom. Agents*, Marina Del Rey, CA, pp. 148-155, 1997.

[58] A.J.N. van Breemen, "iCat: Experimenting with Animabotics", *Artificial Intelligence and the Simulation of Behaviour (AISB)* 2005 *Creative Robotics Symposium*, Hatfield, England, April 2005.

[59] C. Bartneck, "eMuu - An Embodied Emotional Character for the Ambient Intelligent Home", Ph.D. Thesis 2002.

[60] B. Reeves, C. Naas, "*The Media Equation: How People Treat Computers and New Media Like Real People and Places*", Cambridge University Press, 1996.

[61] D.A. Norman, Emotional Design: Why we love (or hate) everyday things, *Basic Books*, 2003.

[62] B.R. Duffy, "Anthropomorphism and The Social Robot", *Robot. Autonom. Syst.,* (Special Issue on Socially Interactive Robots), vol. 42, pp. 170-190, March 2003.

[63] D. Goleman, *Emotional Intelligence*, Bantam Books, 1995.

[64] C. L. Lisetti, and P. Gmytrasiewicz. "Can a Rational Agent Afford to be Affectless?" A Formal Approach. *Appl. Artif. Intel. Int. J.,* vol 16, pp. 577-609, 2002.

[65] D. Canamero, Ed. *Emotional and Intelligent: The Tangled Knot of Cognition*, AAAI Fall Symposium, Orlando, AAAI Press, Menlo Park, CA, October 1998.

[66] N. Wiener, *The Human Use of Human Beings*, Cybernetics and Society, Houghton Mifflin, 1950.

[67] C. Allen, G. Varner, J. Zinser, "Prolegomena to Any Future Artificial Moral Agent", *J. Exp. Theoret. Artif. Intel.,* vol. 12, pp. 251-61, 2000.

[68] B. M. McLaren, "Extensionally Defining Principles and Cases in Ethics: an AI model", *Artif. Intel. J.*, vol. 150, pp. 145-181, November 2003.

[69] B. Friedman, and P.H. Kahn Jr., *Human values, ethics, and design,* Handbook of Human - Computer Interaction , J. Jacko & A. Sears, Eds., Lawrence Erlbaum Assoc., Mahwah, NJ, 2002.

[70] C. Reynolds, and R.W. Picard, "Evaluation of Affective Computing Systems from a Dimensional Metaethical Position", 1st Augment. Cognit. Conf. *11th Int. Conf. Hum. Comput. Interact.,* Nevada, Las Vegas, July 2005,.

[71] P. Brey, "*Disclosive Computer Ethics*", R. A. Spinello and H. T. Tavani, Eds., Readings in CyberEthics, Jones & Bartlett, 2001.

[72] A. Runaround, *Astounding Science Fiction*, March 1942.

[73] R. Clarke, "Asimov's Laws of Robotics: Implications for Information Technology", *IEEE Comput.,* (Published in two parts), vol. 26, pp. 53-61, in December 1993 and vol. 27, pp. 57-66, January 1994.

[74] W.L. Hibbard, "Emotions versus Laws as the Keys to the Ethical Design of Intelligent Machines", *6th World Multi-Conf. Syst. Cybern. Inform. (SCI 2002),* Orlando, Florida, USA, July 2002.

[75] M.L. Minsky, "Will Robots Inherit the Earth?", *Scient. Am.*, Oct 1994.

[76] AAAI Fall Symposium on Caring Machines: *AI in Eldercare*, Washington DC, November 2005.

[77] W. Taggart, S. Turkle, and Cory D. Kidd. An interactive robot in a nursing home: Preliminary remarks. In Toward Social Mechanisms of Android Science, Stresa, Italy, *Cognit. Sci. Soc.*, July 2005.

[78] B.R. Duffy, and G. Joue, "I, Robot Being", *Intelligent Autonomous Systems Conference (IAS8)*, The Grand Hotel, Amsterdam, Netherlands, vo. 10, March 2004.

[79] R.A. Ullrich, The Robotics Primer: The What Why and How of Robots in the Workplace, New Jersey: *Prentice-Hall Inc.,* 1983.

[80] P. Salovey, J.D. Mayer, "Emotional Intelligence", *Imagin. Cognit. Person.*, vol. 9, pp. 185-211, 1990.

[81] J. E. LeDoux, The Emotional Brain, *Simon & Schuster*, New York, 1996.

[82] M. Isen, *Positive Affect and Decision Making*, in Handbook of Emotions*,* M. Lewis and J. Haviland, Eds., Guilford, New York, 2000.

[83] K. Isbister, B. Hayes-Roth, *Social implications of using synthetic characters: an examination of a role-specific intelligent agent*, Tech. Rep. KSL-98-01, Knowledge Systems Laboratory, January 1998.

[84] P. Ekman, and E. Rosenberg Eds., *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression using the Facial Action Coding System (FACS),* Oxford University Press, New York, 1997.

[85] B. Shneiderman, "A nonanthropomorphic style guide: Overcoming the humpty-dumpty syndrome", *Comput. Teach.*, October 1988.

[86] J. Deken, *Silico Sapiens: The Fundamentals and Future of Robots*, Toronto: Bantam, 1986.

[87] D. Sakamoto, T. Kanda, T. Ono, M. Kamashima, M. Imai, and H. Ishiguro, "Cooperative embodied communication emerged by interactive humanoid robots", *Int. J. Hum. Comput. Stud.,* vol. 62, pp. 247-265, 2005.

[88] O. Johnston, and F. Thomas, *The Illusion of Life: Disney Animation*, Disney Editions, ISBN: 0786860707, 1981.

[89] J.C. Lester, and B.A. Stone, "Increasing believability in animated pedagogical agents", *Proc. 1st Int. Conf. Autonom. Agents*, Marina Del Rey, California, US, pp. 16-21, 1997.

[90] J. Bates, "The role of emotion in believable agents", *Commun. Assoc. Comput. Mach.,* (ACM), vol. 37, pp. 122-125, 1994.

[91] J.P. Granieri, W. Becket, B.D. Reich, J. Crabtree, and N.I. Badler, "Behavioral control for real-time simulated human agents", *Proc. 1995 Symp. Interact. 3D Graph.,* pp. 173-180, 1995.

[92] B. Blumberg, and T. Galyean, "Multi-level direction of autonomous creatures for real-time virtual environments", *Proc. Comput. Graph.*, pp. 47-54, 1995.

[93] P. Maes, T. Darrell, B. Blumberg, and A. Pentland, "The ALIVE system: Full-body interaction with autonomous agents", *Proc. Comput. Animat. '95 Conf.*, 1995.

[94] T. Bickmore, "Relational Agents: Effecting Change through Human-Computer Relationships", Ph.D. Thesis, MIT Media Arts and Science, 2003.

[95] B. Kort, R. Reilly, and R.W. Picard, "An Affective Model of Interplay Between Emotions and Learning: Reengineering Educational Pedagogy - Building a Learning Companion", *IEEE Int. Conf. Adv. Learn. Technol. (ICALT)*, Madison, USA, pp. 6-8, August 2001.

[96] B. Joy, Why the future doesn't need us, *Wired Mag.*, vol. 8, April 2000.

[97] Kaplan, "Who is afraid of the humanoid, investigating control differences in acceptance of robots", *Int. J. Humanoid Robot.,* vol. 1, pp. 465-480, 2004.

Received: April 1, 2008 Revised: May 2, 2008 Accepted: May 9, 2008