My Computer, My Friend, My Therapist:
Exploring and questioning the theories and methodologies that guide the design of empathic agents

Sean Voisen
svoisen@uci.edu
Arts Computation Engineering (ACE)
University of California, Irvine

December 9, 2009

Abstract

Affective computing is an approach to computer systems design that attempts to understand, sense, influence or evoke human emotions during interaction with intelligent systems. In recent years, many researchers and practitioners of affective computing have focused their efforts on the design and development of empathic agents – virtually embodied artificial intelligence systems that serve as "emotionally intelligent" help systems, teachers, coaches, therapists or even "friends" to human users. This paper takes a critical look at the fundamental role of the human tendency to unconsciously over-anthropomorphize the behaviors of computer technology – a phenomenon known commonly as the "ELIZA effect" – in empathic agent systems. Using the ELIZA effect as a springboard, I first examine and then question the theories and methodologies behind empathic agent design, ultimately suggesting alternative theories that may be both more fruitful and less ethically consequential than those currently in existence.

Introduction

> *Men are all alike.*
> *IN WHAT WAY*
> *They're always bugging us about something or other.*
> *CAN YOU THINK OF A SPECIFIC EXAMPLE*
> *Well, my boyfriend made me come here.*
> *YOUR BOYFRIEND MADE YOU COME HERE*
> *He says I'm depressed much of the time.*
> *I AM SORRY TO HEAR YOU ARE DEPRESSED*
> *It's true. I am unhappy.*
> *DO YOU THINK COMING HERE WILL HELP YOU NOT TO BE UNHAPPY*
> *I need some help, that much seems certain.*

In 1976, Joseph Weizenbaum, a professor of computer science at MIT, published *Computer Power and Human Reason*, a humanist treatise on mankind, intelligent machines, and their respective places within the world. The above transcript, extracted from the introduction to this book, represents a small portion of a conversation between a young

woman and ELIZA, a primitive artificial intelligence program designed by Weizenbaum partially as an early experiment in natural language processing and partially as a kind of parody of Rogerian psychoanalysis.[1] Shortly after its inception, ELIZA gained fame throughout both artificial intelligence circles and popular culture as a beacon of hope and promise for the future of human-machine relationships. Though ELIZA's linguistic capabilities were limited at best, constrained to very particular sub-domains of interaction for the purposes of "fooling" humans into believing it was capable of meaningful conversation, the publicity surrounding ELIZA often ignored this fact. Some computer scientists believed that Weizenbaum had stumbled upon a general purpose solution to the problem of natural language processing. Prominent psychiatrists – like Dr. Kenneth Colby, at the time a professor at Stanford – lauded Weizenbaum's efforts at automating their practice. Even the renowned astrophysicist, Carl Sagan, had something to say about ELIZA: "No such computer program is adequate for psychiatric use today, but the same can be remarked about some human psychotherapists. In a period when more and more people in our society seem to be in need of psychiatric counseling, and when time sharing of computers is widespread, I can imagine the development of a network of computer psychotherapeutic terminals, something like arrays of large telephone booths, in which, for a few dollars a session, we would be able to talk with an attentive, tested, and largely non-directive psychotherapist" (Weizenbaum 1976).

Weizenbaum penned *Computer Power and Human Reason* in hopes of tempering the public's almost unbridled enthusiasm towards his creation and its perceived implications for the future of computing. The book's 300 pages attempt to explain, in layperson's terms, the inner workings of numerous artificial intelligence systems, warning readers not to confound the simulation of human intelligence and emotion with the real things. "I was startled to see how quickly and how very deeply people conversing with [ELIZA] became emotionally involved with the computer and how unequivocally they anthropomorphized it," wrote Weizenbaum. "What I had not realized is that extremely short exposures to a relatively simple computer program could induce powerful delusional thinking in quite normal people. This insight led me to attach new importance to questions of the relationship between the individual and the computer, and hence to resolve to think about them" (1976).

Today, psychologists and artificial intelligence researchers call this "powerful delusional thinking" the *ELIZA effect*, described as the unconscious tendency to over-anthropomorphize the behaviors of intelligent systems.[2] Critics of

1 Rogerian psychotherapy, sometimes called person-centered therapy, was developed by the psychologist Carl Rogers in the 1940s and 1950s. In this form of therapy, therapists frequently repeat emotionally significant statements back to their patients, often in the form of a question.

2 Some researchers in psychology and communications studies prefer the classical Greek term *ethopoeia* over both anthropomorphism and the more colloquial "ELIZA effect," citing the fact that the direct attribution of a "self" most directly characterizes the phenomenon (Nass, *et. al.* 1993). Nevertheless, the three terms are largely synonymous, and for the sake of clarity, combined with the fact that "anthropomorphism" and "ELIZA effect" appear more frequently in the literature, I have chosen to use these latter terms exclusively throughout this paper. I use "anthropomorphism" when discussing the broad psychological phenomenon characterized by attribution of human-like agency to nonhumans, and "ELIZA effect" when nonhumans connotes computational technology specifically.

artificial intelligence often cite the ELIZA effect as a primary reason for AI researchers to both overstate the capabilities and underestimate the shortcomings of their systems (Ekbia 2009). They claim that it invites a lack of rigor in the language used to describe the operations and behaviors AI systems, where words that researchers should envelope in scare quotes – words like "thought," "creativity," "empathy," and "emotion" – are frequently left bare. Weizenbaum himself became a vocal critic of AI, arguing that the ELIZA effect (though it had no such moniker at the time) put the entire enterprise of artificial intelligence into ethically questionable territory. And though Weizenbaum spent the latter half of his life crusading against the ELIZA effect and its exploitation in artificial intelligence research, today such exploitation has only intensified.

One area where AI researchers continue to take advantage of the ELIZA effect is in the design of empathic agents – virtually embodied artificial intelligence systems that serve as "emotionally intelligent" help systems, teachers, therapists, advisors, storytellers, coaches and even "friends." In the literature, researchers use the terms *empathic agent, synthetic agent* and *animated agent* interchangeably, but the descriptive intent is the same: to convey the idea of a virtual, animated character that lives on a user's computer screen, capable of interacting with and conversing with the user in a variety of specific or general domains (Paiva, *et. al.* 2004). Empathic agent design is a sub-domain of a larger field called *affective computing*. Practitioners of affective computing attempt to design and build interactive computer systems that enchant users with the skillful *sense of* and *response to* human emotions. Put more broadly, the goal of affective computing is to develop a practice in the computer sciences that lobbies for the consideration of human emotion and other affective phenomena during the design of interactive systems (Sengers, *et. al.* 2008). Empathic agent design embodies this goal in its most extreme interpretation, seeking not only to sense and respond to human emotions, but also to evoke emotional responses through "empathic relations between the user and the [agent]" (Hall, *et. al.* 2006). In short, empathic agents are ELIZAs for the 21st century.

Though a relatively recent phenomenon, made possible largely by the rapid hardware improvements described by Moore's law, and by new advances in artificial intelligence and computer graphics, empathic agents can already be found in a wide variety of applications and contexts. Software designers employ these agents in activities as diverse as dispensing healthcare advice (Farzanfar 2006), working as digital dossiers in art museums (Hoorn, *et. al.* 2007), counseling children in appropriate methods for handling schoolyard bullying (Hall, *et. al.* 2006), and providing real-time interactive help to frustrated computer users. Sometimes these agents succeed in their designated roles, and sometimes, like the infamous "Clippit," the cartoon paper clip help agent from Microsoft Office, they fail.

My goal in this paper is not to understand why they succeed or fail necessarily, but rather to question the theories

and assumptions that lie behind these successes and failures, and in the end, suggest alternative theories and methodologies that may prove more fruitful.

I start with the theoretical foundation. The success of an empathic agent hinges in large part on how easy the agent is to anthropomorphize – that is, how readily it exploits the ELIZA effect. As such, a fundamental understanding of anthropomorphism is paramount to understanding why and how empathic agents work. In the pages that follow I first summarize the psychological theories that attempt to explain why a phenomenon like the ELIZA effect exists. I search for answers to the questions: What, exactly, is anthropomorphism? And, more importantly, why do human beings so readily attribute human-like agency to obviously nonhuman objects and mechanisms? Next, I narrow down the discussion, focusing specifically on the predominant theories of empathic agent design. I explore the many assumptions and design frameworks that researchers commonly employ in their quest to create the ultimate in anthropomorphic software: empathic agent systems that are engaging, useful, and most importantly, well-liked by the humans that interact with them. Finally, I examine the primary assumptions and hypotheses that proponents of empathic agents often cite as arguments for the need to develop and deploy such systems. In the spirit of Joseph Weizenbaum's original critique, I propose a few first steps towards a research program that moves beyond simply asking how to make better empathic agents, and begins to ask exactly where and when the use of empathic agents is appropriate and acceptable.

Theories of Anthropomorphism

The human tendency to treat machines as though they are people is a well-documented phenomenon (Turkle 1984, Nass 1996). Precisely how and why this phenomenon occurs and persists, however, remains a mystery. Part of the reason for this may lie in the fact that while much psychological research has been spent studying the extent to which people anthropomorphize technology, little has been dedicated to providing a psychological account of anthropomorphism itself (Epley, *et. al.* 2007). Nevertheless, many researchers – particularly when it comes to the anthropomorphism of machines – fall into one of two camps of opinion on the matter. The first camp argues that "sustained, quasi-social or para-social behaviors," like those described by the ELIZA effect, have their roots in a fundamental ignorance of the inner workings of computers and their mechanical kin (Nass *et. al.* 1994). In short: the only reason people respond socially towards machines is because they don't know how they work. The argument of the first camp predicts, therefore, that the ELIZA effect will lose its power as soon as this ignorance is dispelled – that humans will no longer respond socially to a machine once the veil has been removed and the magic of the machine has been revealed to be nothing more than a clever manipulation of bits.

Researchers in the second camp reject this idea. They argue that humans respond socially to machines *not* because of ignorance or a deficiency in their mental models, but because of an implicit awareness that machines act as a mediating force between their users and their creators. In this view, machines are considered *proxies* that "reflect the attitudes, conceptions and intentions of their [creators]," and through which the users of such machines engage in sustained social relationships with the creators (Nass, *et. al.* 1994). Put more succinctly, people respond socially to machines because they imagine machines as intermediaries between themselves and other humans.

Research by Nass and colleagues, however, contradicts the theories of both camps. In a 1994 study, Nass *et. al.* provided evidence that individual behavior towards computers (and other machines that exhibit even a modicum of human-like characteristics) often *inconsistently* reflects espoused beliefs about such machines. Put more directly, according to Nass, even those who are most certainly *not* ignorant of the inner workings of machines, and even those who do *not* believe machines to be proxies for their creators, nonetheless still tend to anthropomorphize and behave socially towards them (Nass, *et. al.* 1994). Nass' later research on speech-enabled and voice-activated computers suggests instead that humans employ the same small set of rules, cues and heuristics to determine behavior towards "human-like" machines that they employ to determine their behavior towards other humans. For further insight on this matter, Nass looks to evolutionary psychology. In the case of speech-enabled computers, for instance, he argues that because the ability to speak most definitively separates *homo sapiens* from the rest of the animal kingdom, and because the physiological response to speech is largely automatic and unconscious, responding to talking computers as if they were talking humans is simply the most natural response. That is, we are biased, as a result of evolutionary adaptation, to always assume that when we hear speech, it must have originated from a human (Nass and Gong 2000).

Unfortunately, Nass' appeal to evolutionary psychology is largely speculative. As mentioned previously, there is a surprising paucity of research that attempts to provide a detailed account of the constituent psychological mechanisms of anthropomorphism in general, and the ELIZA effect more specifically. Like Nass, traditional psychology has long considered anthropomorphism to be simply an automatic and largely involuntary aspect of human nature – an unconscious by-product of our innate faculties of judgment (Guthrie 1997). Such explanations are vague at best. Fortunately, a more recent theory, entitled "SEEK," dispenses with this bit of hand-waving. SEEK is a three-factor explanation of anthropomorphism based on a set of cognitive and motivational determinants, namely S*ociality*, E*ffectance* and E*licited agent Knowledge,* first proposed by Epley *et. al.* in their paper, "On Seeing Human" (2007). Unlike the theories of Nass or others in computer science and informatics, which focus specifically on the anthropomorphism of computer technology,

SEEK derives from the literature of psychology, and attempts to explain anthropomorphism more generally. I describe it here because it provides perhaps the most detailed and comprehensive picture of the phenomenon available. I do so partially to offer some background, and partially as an attempt to extrapolate its relevance to technology, the ELIZA effect, and empathic agent interaction specifically.

Effectance, the second determinant in the SEEK acronym, describes the human motivation to interact effectively with the environment, and more particularly, with nonhuman agents. Through this determinant, the theory posits that by "attributing human characteristics and motivations to nonhuman agents" humans can use a familiar framework (human sociality) to make predictions about a nonhuman agent's otherwise unpredictable behavior. In so doing, they can reduce the uncertainty and anxiety they might otherwise feel when encountering such an agent, and improve the effectiveness of the interaction. As a result, the theory predicts that anthropomorphic attribution should occur more readily during encounters with more unfamiliar – and hence more anxiety-inducing – nonhuman agents.

Momentarily setting aside the discussion about empathic agents and other technologically manifested nonhumans, this prediction makes sense. In the unpredictable *natural* world, where even the most rigorous scientific explanations may fail us in our attempt to understand, for instance, the peculiar behavior of wild animals or to cope with the extreme devastation of a natural disaster, anthropomorphism may offer some solid ground. Expanding on this example, some scientists may assume that wild animals – say, chimpanzees – have many of the same needs and desires that we do. They may proceed with the working assumption that these animals think and operate with some semblance of our own rationality, and then use this assumption as a basis for effectively interacting with such animals. In the case of natural disasters, consider also that many of us – the most reverent in particular – may understand the appeal of submitting to the unknowable plans and desires of human-like deities as an attempt to assuage our pain during the loss of a home or the loss of a loved one to an earthquake, fire or flood.

The extent to which this aspect of effectance is relevant to ELIZA and research in empathic agents is unclear, however. Its applicable scope may be limited. A large part of the reason for this lies in the fact that though empathic agents may be nonhuman, they are nonetheless designed to emulate human behavior *as closely as possible*, rendering their behavior far more predictable and less anxiety-inducing than, say, the behavior of an imaginary god or a troupe of chimpanzees. Empathic agents make extensive use of existing social models intentionally; they have been designed in many cases to be as easy to anthropomorphize as possible *because of*, rather than *albeit*, their predictability. As such, it is likely that we anthropomorphize these agents not so much out of a need to interact effectively with them when no other

methodology is available, but because this is the *only way* to do so.

Of greater relevance may be *sociality*. This first determinant in the SEEK acronym describes the human need to engage in social behavior and establish connections with other humans, implying that anthropomorphism arises at least partially as a result of this need. Here the theory predicts that anthropomorphic attribution of nonhuman agents should increase during periods of increased social depravity, when developing human-like relations with nonhumans may offer the only means for fulfilling the social necessity. We need look only to the most common nonhuman companions to get a sense of socially determined anthropomorphism in action – namely, pets. Dogs, cats, birds and other pets may be the most commonly anthropomorphized nonhuman agents that offer social companionship, particularly among reclusive individuals and the elderly. But it's possible to imagine empathic agents one day supplanting them. Consider, for instance, "Milo," a next-generation empathic agent developed by Microsoft as part of project "Natal," Microsoft's latest effort to develop a controller-free gaming environment for the popular Xbox gaming platform (Microsoft 2009, Kohler 2009). Milo is the name of a simulated pre-teen boy; his female counterpart is aptly named Millie. Milo lives entirely on-screen, rendered in highly-realistic 3D graphics, and has the capability to engage in meaningful dialogue using audible speech. "He" possesses the ability to "sense" the emotional state of the human with which he interacts, tailoring his reactions and behavior based on his own "observations" and "intuition." And unlike other empathic agents, whose roles tend to fall within areas of limited scope technical assistance or pedagogy, Milo has been designed specifically to fulfill the more general role of a friend or social companion.

Figure 1: Milo, a prototype empathic agent developed by Microsoft

Do we anthropomorphize Milo and other empathic agents as the result of an innate need for social engagement? Unlike effectance, this is muddier territory with no available studies that may provide an answer in the affirmative or negative. Part of the reason for this may be simply that such studies would require of their participants the undue hardship of extended social isolation. Nevertheless, anecdotal evidence does exist. In *Computer Power and Human Reason*, for instance, Weizenbaum describes the startling experience of his secretary shutting him out of his lab so that she could

continued to engage in "therapy sessions" with ELIZA. In addition, several toy makers have produced successful "empathic" and "intelligent" robotic toys – Teddy Ruxpin, Furby, and Pleo to name but a few – in hopes to offer surrogate companionship to bored or lonely children. Sociality plays a key role for human-machine engagement in both contexts, where humans anthropomorphize the machine out of the need to have "someone to relate to" when no flesh-and-blood human is either physically or psychologically available.

The final determinant in SEEK, elicited agent knowledge, describes the notion that humans will first rely on their knowledge about the behavior of other humans, as well as their knowledge about their own behavior, as a foundational basis for understanding and predicting the behavior of unfamiliar nonhuman agents. Unlike effectance and sociality, both of which are considered *motivational* determinants of anthropomorphism, elicited agent knowledge is a *cognitive* determinant. Here the emphasis is on how anthropomorphism arises as a result of cognitive necessity. The theory predicts that, while humans may initially rely on their knowledge of other human behavior when encountering novel agents, as they gain familiarity with an agent they will begin to rely more and more on newly developed knowledge structures over the course of extended interaction.

Like sociality, few studies exist that attempt to affirm or deny this assertion in the context of empathic agents. Though empathic agents attempt to emulate human behavior as close as possible, such emulation is still far from perfect. Therefore, it is likely that elicited agent knowledge does play a role in the ELIZA effect and machine anthropomorphism; while users may initially rely on their knowledge of human behavior to interact with unfamiliar agents, as they become accustomed to an agent's imperfections and behavioral nuances, they may soon develop a new cognitive model vastly different from the one they started with.

As a result, elicited agent knowledge may actually lead to a *reduction* of the efficacy of the ELIZA effect. Indeed, as Wardrip-Fruin notes in his book *Expressive Processing* (2009), the ELIZA effect readily breaks down once a user has gleaned an understanding of the basic processes governing the agent system. The user can then employ that knowledge to either "help maintain or further compromise the illusion."

Theories of Empathic Agent Design

Though psychological models of anthropomorphism such as SEEK offer researchers the opportunity for strong theoretical grounding of empathic agent design, most tend to ground their work not in theory, but in empirical studies. Of particular interest in these studies are users' attitudes and feelings towards empathic agents – that is, users' subjective experience of such systems. To this end, many researchers develop questionnaires in which users are asked to rate their

experience of empathic agent systems on a 5- or 7-point scale, with questions covering a wide variety of characteristics believed to affect anthropomorphic attribution. Dehn and Van Mulken (2000) provide the most comprehensive list of these characteristics, a list that includes such notions as *intelligence, believability, likeability, agency, entertainment value, comfortability, smoothness of interaction, utility,* and *reported attention*. In general, researchers strive to maximize user ratings of one or many of these characteristics. They work under the assumption that the better a user's subjective experience of an empathic agent system, the more likely he or she is to use the system and deem it effective. Here, and throughout the literature, efficacy is often contingent on how easy an agent is to anthropomorphize; the better the ELIZA effect – that is, the more a person responds to an agent socially – the better the system (Gong 2008).

Throughout much of this empirical work, there is a common assumption that the best method for rendering an agent easy to anthropomorphize is through realism and believability; the more human-like an agent appears, the better a user will relate to it. Intuitively, this seems reasonable. Nevertheless, Groom *et. al.* (2009) question this assumption, citing that "literature on agent realism does not offer conclusive support for this claim." Instead, they describe three competing theories relevant to agent realism – the R*ealism Maximization, Uncanny Valley* and C*onsistency* Theories.

The Realism Maximization Theory most closely mirrors the commonly accepted assumption mentioned above – that the more human-like and life-like an agent is, the better. Researchers who most readily subscribe to this theory often turn to the now-burgeoning video game industry for technology that will allow them to render highly photorealistic human-like characters. They seek to develop characters that can both speak with human voices, and communicate "emotion" through the varied and myriad nuances of human body language. Some evidence does support the Realism Maximization Theory, including a study by Gong in which study participants were presented with a wide variety of animated agent faces, from the most abstract and cartoonish to the most photorealistic (2008). Gong found that study participants tended to give more realistic-looking agents higher ratings in questionnaires that measured perceived social judgment, competency and trustworthiness. Unfortunately, detrimental to Gong's study is its confounding of anthropomorphism and realism. The study assumes a false synonymity between the two terms that hampers it with an unacknowledged bias towards Realism Maximization. As such, Gong's evidence for Realism Maximization seems tenuous at best, and while other studies have shown that, for instance, people perceive conversations with realistic agents as "natural" (McBreen and Jack 2001), or prefer agents with faces to those without (Takeuchi and Nagao 1993), the Realism Maximization Theory is far from an accepted fact. Indeed, Groom *et. al* provide significant evidence refuting Realism Maximization (2009).

Contrary to the Realism Maximization Theory, the Uncanny Valley Theory posits that realism is only effective up

to a certain point, after which the more realistic an agent is, the worse the user's subjective experience of that agent. The name of the theory derives from the work of Masahiro Mori and his late 1960s study of human-android interaction, in which he discovered that people tend to find highly-realistic looking androids to be "disturbing" and too "disconcertingly lifelike." He called the middle ground between abstractness and realism in android appearance, where humans are most comfortable interacting with electromechanical imitations of themselves, the "uncanny valley" (Groom 2009, Mori 1970). Though few researchers have attempted to study empathic agents from the perspective of the Uncanny Valley Theory, one need only look to pop culture or personal anecdotal experience with movies, video games or amusement park animatronics to get a sense of the uncanny valley in action. It's not at all uncommon for people to claim that highly-realistic robots like, for instance, the now infamous "Terminator" from the series of movies of the same name, give them the "chills" simply because they seem far *too real*.

Finally, unlike both the Uncanny Valley Theory and the Realism Maximization Theory, the Consistency Theory avoids measuring the overall realism of an agent, and instead focuses on the *consistency* of various indicators of agent realism. Proponents of the Consistency Theory presume that people prefer to interact with agents that exhibit consistent behavior and appearance, regardless of realism. Some evidence supports this: Gong and Nass (2007) found that users respond more socially to agents that use consistently realistic or consistently synthetic voices and faces than to those that do not. Another study found that agents which exhibit "consistent verbal and nonverbal personality cues are more liked, persuasive, and deemed more useful and fun to interact with" than those with inconsistent verbal and nonverbal cues (Nass *et. al.* 2000, Groom 2009).

The Case for Empathic Agents

Designers of empathic agent systems proceed in their practice because they believe that, in the end, empathic agents can lead to a better interaction with a software system over traditional means. As mentioned previously, they believe that empathic agents can improve a user's subjective experience of a system by rendering the system more engaging or entertaining. They worry that in certain contexts, systems without agents may be perceived as dull or not worthy of use, and may not receive the quality of attention they need to accomplish their goals. Proponents also argue that empathic agents can positively affect a user's behavior, for instance by heightening their level of trust or by prolonging their interaction with the system. They claim that empathic agents can give machines the persuasive human presence they need to grab attention and improve the "flow of communication" between machine and user. Others argue that empathic agents can improve

efficiency and performance, either by "intuitively" providing help as needed in realtime, or by improving skill development and knowledge retention through "social" learning (Dehn and Van Mulken 2000). Finally, some have gone as far as to argue that empathic agents are simply outright necessary – computer systems have become too complex and computer users too "naive" that agents offer the only way to effectively interact with these systems (Shneiderman and Maes 1997).

Certainly *some* evidence does exist to support these claims. Maldonado *et. al.* (2005) found that in the particular case of pedagogical software designed for children, systems with empathic agents employed as "co-learners" do improve the learning experience and lead to better retention compared to pedagogical software without agents. And Koda and Maes (1996) found that a system with an animated, empathic agent was more well-liked than an equivalent system without an agent. But these studies are rarities. Most other studies that do provide some shred of support for the notion that systems with agents provide better interaction experiences do so only circuitously. In particular, many studies pit the design of one empathic agent over another, rather than compare the performance of a system with an empathic agent over the performance of a system without an agent. For instance, a study by Bickmore and Picard on "caring machines" found that, in general, users preferred to continue working with behavioral therapy software that included a "caring" empathic agent, over software that had an "uncaring" agent. Yet, the study did not compare the "caring" agent's performance with the performance of software that did not have any agent (2004). Similarly, Nguyen and Masthoff discovered that people prefer empathic agents that are animated and embodied to those that are not, and that people also *expect* agents which assume a human form to exhibit empathic behavior (2009). Yet, again the study offers no "control group" – no system without an empathic agent – that can be used to judge agent efficacy as a whole. Other studies continue this trend, offering rigorous comparisons of the various design facets that affect agent efficacy, but offering little to support the claims that systems with empathic agents offer better user experiences overall (Sproull et. al 1996, Cassell and Thorisson 1999).

The Case Against Empathic Agents

ELIZA notwithstanding, empathic agents are a relatively recent technological development, and as such they do not yet enjoy widespread deployment. But of the few that the public have encountered, some have proven to be embarrassing failures. Microsoft Bob was released in 1995 as an attempt to give a "user friendly" interface to the popular Windows operating system, and included a variety of cartoonish "assistants" that offered the user support and guidance as he or she navigated the the system's environment. Microsoft discontinued the product shortly after its release, citing it as a failure, and earning it a place on PC World Magazine's list of the twenty-five worst technology products of all time (Tynan 2006).

Similarly, Microsoft discontinued "Clippit," the agent from the Microsoft Office software suite, in 2007 after it was met with strong negative responses by many Office users, and ultimately became the brunt of much criticism and parody in the popular media. Some researchers claim that the reason Clippit, Bob and other similar agents evoked irritation and negative reactions in users is because these agents often offered untimely and irrelevant "help," or were deployed in systems focused on information processing only to impede the user's ability to quickly complete his or her task (van Vugt *et. al.* 2006).

Other failures can be found in the research literature. A study of agents as health behavior advisors found that many people prefer interacting with real human beings over empathic agents (Farzanfar 2006). And in a study by Sproull *et. al*. (1996), participants rated a software system with an empathic agent as *less likable* compared to a similar system without an agent.

In 1976, Joseph Weizenbaum was arguably one of the first critics to speak out against the idea of intentionally anthropomorphized machines. But he is not alone. Despite the flurry of research activity surrounding empathic agents, some modern researchers echo his sentiments. Similar to the previously cited theory proposing that machine anthropomorphism arises as a result of ignorance, some argue that empathic agents may disrupt human-computer interaction by inducing false mental models computer systems (Shneiderman and Maes 1997). It is important to note that this is essentially Weizenbaum's original critique – that the anthropomorphism may proceed too far, leading users to ascribe capabilities to the system that it simply does not possess, capabilities like empathy, intuition, and even human-like intelligence. Others propose that agents are simply distracting and require excessive amounts of attention, sentiments of which former users of Clippit or Microsoft Bob may well relate. Finally, some claim that there is simply no need for empathic agents as animated characters; users already anthropomorphize their machines even when they are not represented as embodied human beings (Dehn and van Mulken 2000).

Discussion

What emerges most clearly from this long list of successes, failures, and competing theories on machine anthropomorphism and empathic agent design is a picture of a practice in its infancy. The general lack of consensus among researchers on a variety topics, from the role that realism plays in creating an effective user experience to the merits of such abstract notions as "likeability," gives the impression that the field still does not fully understand the psychological mechanisms or power behind the technology it is so actively developing. An ignorance of this magnitude should sway the field towards a far more cautious type of optimism than it currently exhibits. Consider the following:

> "Building computers that can provide emotional support to their users to help them deal with experienced emotions and offer them encouragement and comfort during difficult times has important implications in many areas." (Nguyen and Mastoff 2009)

> "Animated pedagogical agents offer great promise for knowledge-based learning environments … The extent to which they exhibit life-like behaviors strongly increases their motivational impact ..." (Lester and Stone 1997, Dehn and van Mulken 2000)

My argument for more caution and trepidation in empathic agent research is three-fold. First, as described in the previous pages, the evidence that supports empathic agents as a significant improvement in user interface design is controversial. While some argue that empathic agents "offer great promise," and can provide much-needed "emotional support" to users, others describe them as a distracting nuisance, citing the abysmal track-record of the few agents that have received mass public deployment. Certainly, more work needs to be done. And while some researchers approach their work on agents as if they are a kind of user interface panacea (Shneiderman and Maes 1997), I would begin to question whether empathic agent design theories are universally applicable in all domains. Presently, many studies approach agent design in a particular domain with the pretext that the observations – for instance, that real human voices are better than synthesized voices – can then be extrapolated more generally. This may not hold true. Furthermore, as Dehn and van Mulken (2000) note, "it may be the case that animated agents are only advantageous with regard to the user's attitudes towards the system if they are used in particular domains (such as game playing)."

The second part of my argument is one of ethical considerations. Here I must ascribe myself as in solidarity with Weizenbaum. At its heart, the issue stems from language. Using terms such as "empathic" and "empathy" to describe the behavior of a machine may not be indicative of a false mental model about the machine, but it can easily induce false mental models, even among the experts. Ekbia (2009) writes that "it is surprising how much AI researchers, wittingly or unwittingly, both fall for and take advantage of this common fallacy." Indeed, many AI researchers employ these types of anthropomorphic terms to induce mental models *on purpose*, primarily as a means of making their work appear more groundbreaking and important than it actually is. The danger of this practice lies in the fact that humans, particularly laypeople, may begin to confound the *simulation* of something like emotion for the real thing. When we call an agent

"empathic," the power of language makes it all too easy to believe that the agent is even somewhat capable of feeling and expressing empathy. There exist at least two ethical challenges with this practice: First, as we have seen previously in the study by Farzanfar (2006), researchers have already commenced experiments in which empathic agents serve as healthcare advisors, a vital role traditionally filled by highly trained and experienced humans. It's possible that false mental models of an agent's inner workings can lead to disastrous results in this type of circumstance if the agent is given even a modicum of autonomy. Second, this type of linguistic practice evokes the larger philosophical debate about the nature of human conscious experience and whether the algorithmic simulation of a conscious human quality suffices in these types of critical roles. In the case of ELIZA, is a therapist who has never experienced pain, love, sadness or hope still qualified to work as a therapist? Is a pedagogical agent that does not understand the meaning of learning still qualified to teach? These are the types of questions that the field must address if it wishes to progress in pushing the technology of empathic agents to the general public.

Finally, as we have seen, a vast chasm lies between the theories of anthropomorphism in the field of psychology and the theories of anthropomorphic design in informatics. While one field focuses on notions of sociality and base cognitive knowledge as anthropomorphic determinants, the other focuses on how to evoke a social response to technology almost exclusively through realism and the mimicry of human behavior. The latter approach assumes that the best way to get users of technology to relate to it socially is to make it as human-like as possible. To this end, researchers equip empathic agents with human faces, human voices, human names and human behaviors all in an attempt to get users to relate, socialize and "make friends" with their software. I question the wisdom of this approach.

The SEEK model of anthropomorphism suggests that a machine need not look or talk like a human for users to socially engage with the machine as if it were a human. Users will engage socially purely based on social need, existing cognitive models, and the need to interact effectively. Other research supports this notion, clearly demonstrating that human users will attribute personality to a machine even without pictorial representations, natural language processing, artificial intelligence or other agent-like qualities (Nass *et. al.* 1995). Perhaps, then, the field needs to re-evaluate its methods and branch out from simply trying to emulate human appearance and behavior in software and hardware.

To this end I suggest a look back at the theoretical roots of affective computing. While some claim that the goal of affective computing is to give computers the ability to identify and respond to distinct human emotional states, Sengers *et. al.* (2008) offer an alternative vision in which affective computing "provides opportunities for users to experience, interpret, and reflect on their own emotions." In this alternative vision, "affective presence systems focus on affective *experience*,

rather than affective *computing*; they support reflection on rich, enigmatic experiences of affect." While empathic agents may certainly still play a role in this alternative approach to affective computing, they may also look and behave vastly different from their current incarnations. Rather than attempt to directly interact with the user through text or natural language, they may instead act as cinematic characters, acting out or proposing affective scenarios that provoke us to reflect on our own humanity. Rather than exist as depictions of human-like forms, they may instead possess a far more abstract embodiment, evoking emotional response primarily through behavior rather than appearance.

Some practitioners in the new media arts use the term "expressive AI" to denote the use of artificial intelligence as an expressive medium (Mateas 2001, Penny 2009). In many ways, the practice of expressive AI is similar to the alternative depiction of empathic agent design described above. As researchers in affective computing search for new ways to develop empathic agents without appealing to human appearance or behavior, it is possible that a marriage between the scientific practice of empathic agent design and the artistic practice of expressive AI could prove fruitful. At the very least, it would open up new avenues for emotion-based computing that are not simply limited to building artificial characters, but instead expand into unexplored realms of human-machine interaction, better enhancing our understanding of anthropomorphism, as well as what it means to be human.

Conclusion

Though the ELIZA effect is a well-documented phenomenon, as we have seen there is little consensus and substantial ignorance as to how or why it works. And while the pursuit of realism in appearance and behavior may offer some promise in inducing this effect, it is not the only available strategy. Practices of affective computing such as empathic agent design should take this into greater consideration, especially considering the fact that the exploitation of the ELIZA effect has real ethical and philosophical ramifications. Researchers must take care to remind themselves and the users of their creations that the simulation of human behavior is not the same as the real thing, and in many contexts, may not be a viable substitute either. If we want future computer users to engage socially with technology, we may not need to try to fool them into believing that machines can think, act, and feel as they do. Instead we may need only to design the machines so that they can create an *experience* where human actions, thoughts and feelings take center stage.

Works Cited

Bickmore, T. & R. Picard (2004) "Towards Caring Machines" Proc. of CHI 2004, Vienna, Austria.

Cassell, J. & K. R. Thorisson (1999) “The power of a nod and a glance: envelope vs. emotional feedback in animated conversational agents.” *Applied Artificial Intelligence*, Vol. 13:519-538.

Dehn, D. & S. van Mulken (2000) “The impact of animated interface agents: a review of empirical research.” *Int. J. Human-Computer Studies*, Vol. 52:1-22.

Ekbia, H.R. (2009) *Artificial Dreams: The Quest for Non-Biological Intelligence.* Cambridge University Press, New York, NY.

Epley, N., Waytz, A. & J. T. Cacioppo (2007) “On Seeing Human: A Three-Factor Theory of Anthropomorphism.” *Psychological Review*, Vol. 114, No. 4:864-886.

Farzanfar, R. (2006) “When computers should remain computers: a qualitative look at the humanization of health care technology.” *Health Informatics Journal*, Vol. 12, No. 3:239-254.

Gong, L. (2008) “How social is social responses to computers? The function of the degree of anthropomorphism in computer representations.” *Computers in Human Behavior*, Vol. 24:1494-1509.

Gong, L. & C. Nass (2007) “When a talking-face computer agent is half-human and half-humanoid: human identity and consistency preference.” *Human Communication Research*, Vol. 3, No. 2:163-193.

Guthrie, S. (1997) “Anthropomorphism: A Definition and Theory” in *Anthropomorphism, anecdotes, and animals* (Mitchell, R. W., Thompson, N. S., & H. L. Miles. eds) State University of New York Press, Albany, NY.

Koda, T. and P. Maes (1996) “Agents with faces: the effect of personification.” Proc. of 5th IEEE International Workshop on Robot and Human Communication.

Kohler, C. (2009) “Hands On: Milo and Kate, and Other Project Natal Games.” *Wired Magazine*, June 3, 2009. Retreived from http://www.wired.com/gamelife/2009/06/project-natal/ on November 6, 2009.

Lester, J. C. & B. A. Stone (1997) “Increasing believability in animated pedagogical agents.” *Proc. of 1st Int. Conf. on Autonomous Agents.* Marina del Rey, CA.

Maldonado, H., Lee, J.-E., Brave, S., Nass, C., Nakajima, H., Yamada, R., Iwamura, K. and Y. Morishima (2005) “We learn better together: enhancing eLearning with emotional characters.” Proc. of CSCL, Taipei, Taiwan.

Mateas, M. (2001) “Expressive AI: A Hybrid Art and Science Practice.” *Leonardo*, Vol. 34, Part 2:147-154.

Nass, C. & L. Gong. (2000) “Speech Interfaces from an Evolutionary Perspective.” *Communications of the ACM*, Vol. 43, No. 9:36-43.

Nass, C., Ibister, K., Lee, E.-J. (2000) “Truth is beauty: researching conversational agents” in *Embodied Conversational Agents* (Cassell, J., Sullivan, J., Prevost, S., & E. Churchill. eds.) MIT Press, Cambridge, MA.

Nass, C., Moon, Y., Fogg, B., Reeves, B. & D. C. Dryer (1995) “Can computer personalities be human personalities?” *Int. J. of Human-Computer Studies*, Vol. 43:223-239.

Nass, C., Steuer, J., Henriksen, L. & D. C. Dryer (1994) “Machines, social attributions and ethopoeia: performance assessments of computers subsequent to 'self-' or 'other-' evaluations.” *Int. J. Human-Computer Studies*, Vol. 40:543-559.

Nguyen, H. & J. Masthoff (2009) “Designing Empathic Computers: The Effect of Multimodal Empathic Feedback Using Animated Agent.” Proc. of Persuasive 2009. Claremont, CA.

Paiva, A., Dias, J., Sobral, D., Aylett, R., Sobreperez, P., Woods, S., Zoll, C. & L. Hall (2004) “Caring for Agents and Agents that Care: Building Empathic Relations with Synthetic Agents.” Proc. of AAMAS '04, July 19-23, New York, NY.

Penny, S. (2009) “Rigorous Interdisciplinary Pedagogy.” *Convergence*, Vol. 15, No. 1:31-54.

Sengers, P., Boehner, K., Mateas, M. & G. Gay (2008) “The disenchantment of affect.” *Personal and Ubiquitous Computing,* Vol. 12, No. 5:347-358.

Shneiderman, B. & P. Maes (1997) “Direct manipulations vs. interface agents: excerpts from debates at IUI'97 and CHI'97.” *Interactions*, Vol. 4:42-61.

Sproull, L., Subramani, M., Kiesler, S., Walker, J. H., and K. Waters (1996) “When the interface is a face.” *Human-Computer Interaction*, Vol. 11:97-124.

Turkle, S. (1984) *The Second Self: Computers and the Human Spirit.* Simon & Schuster, New York, NY.

Tynan, D. (2006) “The 25 Worst Tech Products of All Time.” *PC World Magazine*, May 2006.

van Vugt, H. C., Hoorn, J. F., Konijn, E. A., & A. de Bie Dimitriadou (2006) “Affective affordances: Improving interface character engagement through interaction.” *Int. J. Hum.-Comput. Stud*. Vol. 64, No. 9:874-888.

Wardrip-Fruin, N. (2009) *Expressive Processing: Digital Fictions, Computer Games and Software Studies.* MIT Press, Cambridge, MA.

Weizenbaum, J. (1976) *Computer Power and Human Reason: From Judgment to Calculation.* W.H. Freeman and Company, San Francisco, CA.

“Xbox Unveils Entertainment Experiences that Put Everyone Center Stage.” Microsoft Press Release, June 1, 2009. Retrieved from http://www.microsoft.com/Presspass/press/2009/jun09/06-01E3PR.mspx on November 6, 2009.