PsychNology Journal
Vol. 7 N.1
ISSN 1720-7525

Ethics in Presence and Social Presence Technology

PSYCHNOLOGY JOURNAL
The Other Side of Technology

PSYCHNOLOGY JOURNAL, PNJ
PUBLISHED ON-LINE SINCE SUMMER 2002
WEB SITE: HTTP://WWW.PSYCHNOLOGY.ORG
SUBMISSIONS: ARTICLES@PSYCHNOLOGY.ORG

TABLE OF CONTENTS

Editorial Preface

Recent years have seen an explosion in computer-mediated communication and interaction. Instead of meeting face to face, we use audio- or videoconferencing or instant messaging or email; instead of playing in the same room we use games consoles; we use technology to shop at a distance, learn at a distance, give psychotherapy at a distance. Yet rarely is distance seen as an advantage. In most cases, system and interface designers seek to hide it. What they usually try and create is a sense of "being there" and "being in the same environment with other actors". Technologies that achieve this effect are Presence and Social Presence technologies. Try interrupting an adolescent in the middle of a videogame or a chat. She's not in the room with you but somewhere else, in a different space. The ability to create this illusion raises novel ethical issues. It is these issues that we will explore in this special issue of PsychNology.

Some are familiar. Presence research involves experiments with human participants. So we need their informed consent and we have to ensure that the information is real and the consent freely given. We have to protect our participants. For instance, we have to make sure that experimental stimuli do not hurt them, physically or psychologically. To enhance user interactions we may collect and process sensitive information about their location, their activities and mood, or their interactions with other users. We have to protect this data and guarantee user privacy.

Yet these requirements, however important, are not specific to Presence. Informed consent and protection of study participants is essential in any kind of psychological study; threats to privacy are implicit in nearly any use of modern telecommunications. What is specific is the fact that the very coordinates and features of the users' presence in a certain environment are largely monitorable and reconfigurable. The purpose may be benign: we can use it to bring distant friends closer together or to flirt or for psychotherapy; it can be malignant, as when designers covertly attempt to manipulate user behavior. More importantly "presence" may have effects which have nothing to do with the designer's intentions, and that need to be unveiled.

It was to discuss these issues that on October 16, 2008, the Human Technology Lab of University and Xiwrite Srl jointly organized a workshop on "The Ethics of Presence and Social Presence Technologies". The workshop, which was held in conjunction with Presence 2008, was sponsored by the PASION project, an EU-funded investigation of "Psychologically Augmented Social Interaction over Networks". The goal of the organizers was to bring together "philosophical" approaches, capable of placing presence in a broad perspective, and the viewpoint of practitioners, often immersed in the details of experimentation or in the design of services and products. In reality, as they had hoped, many of "philosophers" present at the meeting showed a deep interest in technology, and many of the "practitioners" raised new and interesting theoretical issues. In this special issue of PsychNology, we present four papers based on the discussion at the workshop.

The first "Building Character for Artificial Conversational Agents: Ethos, Ethics, Believability, and Credibility" by Sheryl Brahnam, looks at a specific presence technology ("artificial conversational agents"), examines the challenges facing designers wishing to make agents that are "believable" and "credible" and looks at the ethical implications. Brahman returns to the ancient controversy on rhetoric - seen by some ancient authors as something "implicitly duplicitous and morally suspect" - persuasion by artifice, by

others as a vital tool enabling people to "persuade one another and to clarify their desires and needs". All this is tied to the concept of ethos, the character and reputation of a speaker, (or virtual agent). Ethos can be a purely linguistic construction, something that makes the speaker appear as credible and trustworthy - or it can be something "developed slowly and painstakingly through habit and virtuous action". Designers tend to take the former option, using their art to construct "believability" through artificial means. The consequence is the destruction of trust when the true nature of the agent's ethos is revealed (when the user realizes she is talking to a machine). Yet Brahnam argues that this is by no means inevitable, showing, in the final section of her paper how it might be possible to use non-artistic methods to create a richer, more trustworthy ethos for artificial agents.

How far this is achieved will rely on factors that do not depend exclusively or primarily on the good will of designers. In the meantime, interactions between users and the current generation of virtual agents already provide cause for concern. In our second paper "Ethical implications of verbal disinhibition by conversational agents", Antonella De Angeli discusses findings that interactions with virtual agents may encourage disinhibited and anti-social behavior (e.g. sexual abuse of attractive, "female" agents). While this kind of behaviour may be apparently harmless, it rapidly acquires ethical significance when agents are deliberately designed to elicit disinhibition (e.g. in purchasing behavior) or stereotyping. When artificial agents are used as sales agents on web sites the risk is obvious, and will only become greater as the technology improves.

The third paper, "Witnessed Presence and the YUPTA framework", by Caroline Nevejan, suggests some of the ways in which Presence technologies can disrupt traditional human interaction. The YUPTA framework she proposed makes it possible to represent the respective roles of Time, Place, Action and Relation in different forms of presence. Using the framework, she shows how the technology affects what she calls "witnessed presence" - the experience that one's actions can be witnessed by many others. "Witnessed presence", she argues, is an essential ingredient in the negotiation of trust and truth. This is especially important in the formation of democratic public opinion. "Only in natural presence the shared sense of what is good for well-being and survival can be 'collectively authored' in such a way that all stakeholders will base their future acts on the 'collectively authored outcomes' that have been agreed upon". The effects of presence technology may be more subtle than we imagine.

The final paper in the special issue, "Cybertherapy: advantages, limitations and ethical issues" by Cristina Botella, Azucena Garcia-Palacios, Rosa M. Baños, and Soledad Quero, discusses an application of presence technology that is already coming into widespread use, namely the application of Virtual and Augmented Reality in clinical psychology ("cybertherapy"). These presence technologies have been successfully used not only for the treatment of specific phobias but also for more severe disorders such as panic disorder, posttraumatic stress disorder, eating disorders and pathological bereavement. More recently, the Internet has made it possible to deliver effective treatment at a distance. It is clear that this raises a number of ethical issues. However the authors show that many of these have been successfully resolved: virtual reality sickness appears to affect only a minority of patients; doubts about children and elderly patients have been addressed; it has been shown that the new therapies work even with severe anxiety or psychotic disorders. The key

issues that remain seem to be tied not so much to the technology itself as to the broader social context in which it is deployed. So there is the risk of patient self-diagnosis, the difficulty of enforcing appropriate treatment protocols at a distance, the problem of establishing the identity of patient and therapist. Here too, as in the other papers in this issue, we discover how even well-designed technologies can have unforeseen consequences.

Which of course it is the designer's role to avoid. All designers know that only designers build trustworthy, non-manipulative presence technologies. Yet they do not operate in a vacuum. Perhaps the central message in all the papers in this issue is that the ethics of presence depends not just on the technologies themselves, but on the broader framework in which they are used. The editors hope that this issue of PsychNology will help to increase awareness and adopt a reflexive approach in designing, studying or using Presence and Social Presence technologies.

Richard Walker
XiWrite s.r.l., Italy

Luciano Gamberini
Università degli Studi di Padova, Italy

Anna Spagnolli
Università degli Studi di Padova, Italy

Outside the special issue, the other section of the journal offers one article by Cheryl Campanella Bracken and Paul Skalski, "Telepresence and Video Games: The Impact of Image Quality". The experiment described in the paper represents an advance in the topic to which the authors have already made several distinguished contributions, namely the sense of presence experienced with different media formats, in particular media belonging to everyday household experience. In this case, they investigate the effect of television image quality.

The article "What could abductive reasoning contribute to human computer interaction? A technology domestication view" by Erkki Patokorpi is a passionate proposal to include abduction in the design of human computer interaction. As the author argue, this kind of process seems particularly close to everyday reasoning and then its implementation would improve the usability of a computer system.

The Editors-in-Chief

PsychNology Journal, 2009
Volume 7, Number 1, 9 – 47

Building Character for Artificial Conversational Agents: Ethos, Ethics, Believability, and Credibility

Sheryl Brahnam[♣*]

[♣]Missouri State University, Department of Computer Information Systems (USA)

ABSTRACT

Because ethos is an unavoidable component of dialogue and forms the basis for believing and being persuaded by another's speech, it is an important topic for AI researchers. This paper examines the concept of ethos, especially Aristotle's notions of situated and invented ethos, as it functions in oral and written discourse and then explores what happens to ethos in computer-mediated human-to-human and human-to-machine discourse. The paper draws a number of conclusions that may be of value to researchers in these fields. In particular, it argues that the rhetorical concept of ethos furnishes a broader theoretical framework for understanding design and ethical issues involved in agent credibility than does the artistic notion of believability. The paper concludes by suggesting some nonartistic methods for making agents more credible within the framework of situated ethos.

Keywords: *ethos, conversational agents, believability, trust, anthropomorphism, Eliza effect, verbal abuse, computer-mediated communication, transference, oscillation effect*

Paper Received 04/04/2009; received in revised form 28/04/2009; accepted 28/04/2009.

1. Introduction

> The ethos reveals itself in time, and the revelation brings disaster.
>
> Aeschylus

Most computer interfaces use human language to communicate with users, and this fact has consequences that I feel have not been explicitly drawn out but that I think are at the core of current design concerns and ethical dilemmas in the development of more loquacious interfaces, such as conversational agents. One consequence is that

Cite as:

Brahnam, S. (2009). Building Character for Artificial Conversational Agents: Ethos, Ethics, Believability, and Credibility. *PsychNology Journal, 7*(1), 9 – 47. Retrieved [month] [day], [year], from www.psychnology.org.

[*] Corresponding Author
S. Brahnam
Missouri State University, CIS Department, 387 Glass Hall, Springfield MO, 65804, USA
E-mail: sbrahnam @ facescience . org

computer communications, because they are acts of language, cannot be divorced from the ethics of rhetoric.

There are many definitions of rhetoric. These definitions are often marked by an ethical antithesis, with theorists lining themselves up more or less on one side or the other. A pejorative sense of the term, often voiced by Socrates, is bound up with the notions of appearance and artifice. Rhetoric in this regard is artificial, showy, and decorative. It is the fluff of discourse, a bag of tricks. These ideas are reflected when we use the word dismissively. An argument is "mere rhetoric," for example, when we feel it lacks substance. For Socrates rhetoric is inherently duplicitous and morally suspect because too often it is rived from reason and truth.

Other theorists are far more positive in their appraisal of rhetoric. Isocrates, for example, holds rhetoric in the highest esteem since it enables people to persuade one another and to clarify their desires and needs. Without rhetoric there would be no civilization. In his famous speech, "Antidosis," he claims that it is because of rhetoric that we have cities, laws, and the arts. For Isocrates, " . . . there is no institution by man which the power of speech has not helped us to establish" (p. 28).

Parallel to rhetoric's dubious connection to truth and reality is another source of ethical anxiety, one that I believe is at the very heart of many issues involving artificial conversational agents. The Greek derivative of the word *rhetoric* is *eiro* meaning "I say." In this sense, rhetoric is concerned with the relationship of discourse (what is said) to the character of the speaker (the *I*), or what the Greeks called *ethos*, a term often treated as a synonym for *character*, *reputation*, and *persona* in the classical literature and for *subject* and *self* in more modern discourse.

In the Western world, ethos, or the relation of a speaker to her speech, has proven problematic. The etymology of the word discloses a number of binary oppositions that are at the core of many debates: the domestic versus the wild (in reference to animals and the incorrigible vileness of human beings), the cultivated versus the innate, the public versus the private, and the singular versus the plural (see Chamberlain, 1984). In each of these oppositions, ethos is bound up with some sense of place (Miller, 1974), but where ethos dwells (inside an individual soul or within some public sphere) and whether it is reachable (informed by habit and custom or innate and incorrigible) are all fiercely contested.

In discourse, these uncertainties throw into question the very ground and source of a speaker's credibility. If ethos is embedded in the rhetor's reputation, for example, then it is developed slowly and painstakingly through habit and virtuous action. Moreover, it

follows that ethos is essentially incarnate and cannot be crafted since it is part of the moral fiber of the speaker. If ethos is a linguistic construction, however, then it is artificial and imitative, an artistic invention. What matters is that the words of the speaker make him *appear* credible and trustworthy. This opens the door to dissimulation. A speaker can learn to construct a convincing ethos, one that is suitable for various occasions, for instance, by taking courses in speech making or by hiring a speechwriter.

Because ethos is an unavoidable component of dialogue and forms the basis for believing and for being persuaded by another's speech, it is an important topic for AI researchers. Whether that speech comes from a human being or from a computer program, communication depends on the credibility of the speaker; otherwise what is being said is subject to mistrust and doubt, creating a situation that seriously jeopardizes the willingness of the speaking partners to engage in cooperative activities.

The debate whether ethos is a linguistic phenomenon or whether it is reflective of a human being's character is especially relevant in the development of humanlike interfaces. Most of these interfaces attempt to imitate the essential behaviours involved in face-to-face communication, with success measured against the artistic standard of believability (Loyall & Bates, 1997). *Believability* in the media arts refers to the human tendency to suspend disbelief in order to enjoy the portrayal of a separate imaginary being. This concept has driven developers to concentrate on strengthening the natural human tendency to anthropomorphize. Against the broader rhetorical notion of ethos, however, believability is seen as a limiting concept because it already plants itself squarely on one side of the ethical divide: it assumes that credibility, or ethos, is fictive and that it can be detached from human agency and exist on its own.

That ethos can be entirely removed from human *being* is questionable, as I hope to demonstrate by examining ethos in the light of what happens to it when the human speaker is progressively removed from discourse. This examination will also serve as a brief introduction to the subject of ethos. In section 2, I look at the Greek origins of the concept as it is defined mostly in terms of oral discourse. In particular, I look at Aristotle's key notions of situated and invented ethos, especially as they relate to ethics. In section 3, I present some of the problems with ethos in written discourse, showing how the voice of the author loses its claim on ethos as it slides to the living reader's interpretation of the text. In section 4, I consider ethos in Computer-mediated discourse, noting how in the age of secondary orality the notion of invented ethos

expands to include the visual as well as the linguistic. I also show how the disassociation between the speaker's words and her body in textual exchanges renders trust and credibility problematic. Finally, in section 5, I arrive at the main purpose of this paper and explore what happens to credibility when a human speaker is entirely removed from the words that are being spoken, as is the case in computer-generated discourse. In this section, I note that the ethical divide in rhetoric is reflective of a deeper conflict in the Western self between the imaginal and the real, a conflict made palpable when people are asked to communicate with a computer using a natural language interface. Just as there is the Eliza effect (Weizenbaum, 1976), or the tendency for people to anthropomorphize and ascribe greater human capabilities to computers then they actually possess, there is also a corresponding Weizenbaum effect, or a tendency to debunk the anthropomorphic. In their interactions with humanlike interfaces, people are torn between the two. As Aeschylus warned in the epigraph, when the true nature of a creature's ethos asserts itself (in this case revealing the agent to be a machine), disaster invariably follows. In section 6, I conclude the paper by showing how the richer concept of ethos, especially situated ethos, might foster the discovery of nonartistic methods for overcoming some of these disasters in credibility.

2. Ethos in Oral Discourse

As Welch (1990) points out, classical rhetoric continues to inspire and to inform because it covers all possible uses of language. She writes, "While at various historical times it emphasizes some kinds of speaking and writing over other ones . . . classical rhetoric nonetheless self-consciously concerns itself with all manifestations of discourse" (p. 5). Since language issues from a speaker, ethos, or the character of the speaker, is a central topic of discussion, but as Farrell (1993) observes, it is also "one of the most enigmatic concepts in the entire lexicon" (p. 80).

In classical rhetoric, ethos is one of three modes of *pisteis*, or means of persuasion, with the other two being *logos* and *pathos*. Each of these modes centers on a fundamental aspect of oral communication (Wisse, 1989). Logos is focused on the words of the speech, with logical proofs and arguments. Pathos is centered on the audience, its emotions and reactions. Ethos refers to the speaker, his reputation and presentation of character. Thus, persuasion is not solely concerned with developing

sound arguments and proofs; it also involves putting the audience into a receptive frame of mind and convincing them that the speaker is a credible person.

What makes a speaker worthy of credence? In the beginning of the second book of the *Rhetoric*, Aristotle states, "There are three things which inspire confidence . . . good sense [*phronesis*], excellence (virtue) [*arete*], and goodwill [*eunoia*]" (1378a5). Good sense is concerned with practical intelligence, expertise, and appropriate speech. A speaker who demonstrates knowledge of a subject and exhibits a sense of propriety and aptness of language is believed (1408a20). Excellence refers to whatever socially sanctioned virtues good people are expected to possess. Goodwill conveys the impression that the speaker has the welfare of the audience in mind. Aristotle associates goodwill with friendship and friendly feelings. He writes, "We may describe friendly feeling towards anyone as wishing for him what you believe to be good things, not for your own sake but for his, and being inclined, so far as you can, to bring these things about" (1380b35). Wisse (1989) points out that Aristotle's categories of credibility include the rational (good sense), the emotional (goodwill), and the trustworthy (excellence or virtue).

If a speaker is deemed credible, then the audience will form a second order judgment that the arguments put forth by the speaker are true and acceptable. The persuasive impact of ethos, therefore, should not be underestimated. According to Aristotle, the speaker's "character may almost be called the most effective means of persuasion he possesses" (1356a14). It provides what has come to be called an *ethical proof*.

Aristotle recognizes two kinds of ethical proof: invented and situated. If a speaker is fortunate enough to enjoy a good reputation in the community, she may rely on that as an ethical proof. This is *situated ethos*. Although not the focus of Aristotle's discussion of ethos, this is the course that is advocated by Isocrates, Cicero, and Quintilian, among others. For these writers, good character is more important than clever speech. As Isocrates argues in the *Antidosis*, "the man who wishes to persuade people will not be negligent as to the matter of character; no, on the contrary he will apply himself above all to establish a most honorable name among his fellow-citizens" (p. 339). Reputation is a method for engendering belief because the way a speaker lives offers the best evidence of the truth and goodness of that person's words (Welch, 1990).

Aristotle, in contrast, is not as interested in cultivating the character of the rhetor as in discovering how a speaker can persuade an audience to trust what he has to say. Aristotle appears to be turning the tables here by arguing that the persuasive power of ethos is more in a speaker's words than in his reputation. He writes, "This kind of

persuasion, like the others, should be achieved by what the speaker says, not by what people think of his character before he begins to speak" (1356a10). The dramatization of character within discourse is *invented ethos*.

Many have tied Aristotle's notion of invented ethos to his descriptive procession, also found in the *Rhetoric*, of a host of character types. Rhetoricians, both ancient and modern, have used these and similar sketches as student exercises in *ethopoeia*, or the art of fabricating character in language. Others believe the portraits are intended to help rhetors recognize the different psychological types in their audience so that the speaker can better carry out Socrates' injunction in the *Phaedrus* to pair types of speech to types of souls.

Exactly what Aristotle is advocating in his discussion of situated and invented ethos, however, is debatable. According to Yoos (1979), Aristotle is simply showing how an audience's impressions of a speaker's character can be altered by the language used by the speaker. The rhetor is then faced with a choice: he can distort the audience's perceptions of his character in his speech or he can go about developing "rhetorically effective personal qualities" (p. 44) by becoming a good person. In a similar vein, Wisse (1989) notes that Aristotle is clearly stating that telling the truth is central to ethos and that invented ethos must be based on the speaker's true character.

For many others Aristotle's concept of ethical proof is anything but ethical, primarily because it obscures a distinction between *real* character, which is cultivated through good habits, and an artistically produced character, which invites pretense and dissembling. Although Gill (1982) argues that it is actually those following Aristotle who extend the idea of ethos to the point of making it an invention (a persona, or mask, that a speaker assumes for the duration of a speech), Aristotle's characterization of ethos seems to have paved the way for treating ethos as artifice by suggesting that what really matters is not so much the quality of the innate character of the speaker but the perceptions of the audience, and these can be manipulated.

A careful examination of invented and situated ethos, however, shows that in both the judgment of the audience is crucial. Invented ethos is bound to a single instance of speaking in public and involves the immediate revelation of character, a momentary portrayal that may or may not honestly represent the speaker's true character and that may be intended more as the object of sport or as a rhetorical exercise (in ethopoeia, for instance).

Situated ethos, in contrast, is based on repeated exposure to the public and on building trust over time by consistently demonstrating the qualities of good sense,

excellence, and goodwill. This recalls the meaning of the ancient Greek word *ethos* as "habitual gathering place." As Holloran (1982) notes, "I suspect that it is upon this image of people gathering together in a public place, sharing experiences and ideas, that its meaning as character rests" (p. 60). His views are echoed by Reynolds (1993), who claims that the classical sense of ethos is not wrapped around the individual (it is not singular), but rather it refers to the surrounding social context (it is plural). In this regard, situated ethos can be thought of as a longterm relationship that develops in the exchange of ideas between an individual and the other members of her community.

3. Ethos in Written Discourse

The most obvious difference between writing and oral discourse is the loss of the bodily presence of the speaker in writing. Without that presence, the question arises whether writing has any claim to credibility, and, if so, then where does it reside?

A view that follows from the Aristotelian notion of invented ethos and that held for nearly 2500 years is that ethos is in the text. An author's character is revealed in his writing style, whether by design or by default, and a convincing writer is one who manages to exhibit the key ingredients of credibility: good sense, excellence, and goodwill.

The debate whether credibility is contingent upon a genuine personality or whether it can be feigned resurfaces in writing in new guises. While some writing manuals encourage students to find their unique voices, others offer exercises intended to teach students methods for selecting and rendering personas that best suit their writing project. For example, Hunt (1991) in the *Riverside Guide to Writing*, has students jot down the personality qualities of their intended personas, somewhat in the fashion of Aristotle's character portraits.

In contrast to Aristotle's ethos, which accommodates the separation of speech from the bodily presence of the speaker, is Plato's "ethos,"[1] which according to Baumlin (1994), is "the space where language and truth meet or are made incarnate within the individual" (p. xiii). For Plato, body and speech cannot be sundered. In the *Phaedrus*, Plato has Socrates argue that writing has no ethos precisely because the speaker's body has been removed:

[1] Plato never uses the term *ethos* but, as Baumlin (1994), among others, note, his concept of ethos, or the relation of human character to language, can be inferred.

> You know, Phaedrus, writing shares a strange feature with painting. The offsprings of painting stand there as if they are alive, but if anyone asks them anything, they remain most solemnly silent. The same is true of written words. You'd think they were speaking as if they had some understanding, but if you question anything that has been said because you want to learn more, it continues to signify just that very same thing forever. When it has once been written down, every discourse roams about everywhere, reaching indiscriminately those with understanding no less than those who have no business with it, and it doesn't know to whom it should speak and to whom it should not. And when it is faulted and attacked unfairly, it always needs its father's support; alone it can neither defend itself nor come to its own support. (275d-e).

In this passage, Socrates is, in effect, charging writing with violating the dictums of good sense, excellence, and goodwill. The brunt of Socrates' charges fall on writing's lack of good sense: it is incapable of defending itself when attacked; it repeats the same words endlessly without variation; it fails to field questions; and, because it lacks perception, it is unable to adjust the language and its style of delivery to accommodate the psychological differences in the people it encounters. These last two faults suggest a lack of goodwill. After all, how would it be possible for writing to have the reader's welfare in mind if it is not willing to interact? Finally, writing is not virtuous because it deceives. Like a painting, it is an image pretending to be what it is not: living speech.

In general, postmodern thought would also claim that ethos is not in the text. If ethos exists anywhere, it is in the minds of the readers, in their interpretations, constructions, and projections upon the text. Readers recast the authors into their own images. As Corder (1989) explains, "the author is dead and language writes us, rather than the other way around, and interpretation prevails rather than authorship" (p. 301). Corder goes on to say that in writing, "Language is orphaned from its speaker; what we once thought was happening has been disrupted. Authors, first distanced, now fade away into nothing. Not even ghosts, they are projections cast by readers" (p. 301)[2].

[2] Writers also often acknowledge the fact that once a text is printed it no longer belongs to them. John Steinbeck, near the end of writing one of his novels, expressed a profound sense of this loss when he wrote, "In a short time [the novel] will be done and then it will not be mine any more. Other people will take it over and own it and it will drift away from me as though I had never been a part of it. I dread that time because one can never pull it back, it's like shouting good-bye to someone going off in a bus and no one can hear because of the roar of the motor" (Plimpton, 1977, p. 199).

How is ethos a creation of the reader rather than of the writer? According to Baumlin and Baumlin (1994), ethos is the "projection of authority and trustworthiness onto the speaker, a projection that is triggered or elicited by the speaker but that is otherwise supplied by the audience" (p. 99). It is the formation of a transference, an unconscious displacement of feelings for one person to another. It is because transferences are attached to the images of people, as Derrida (1987) points out, and not the people themselves, that they can also be placed on authors and texts.

Brooke (1987) claims that persuasion would not be possible at all without positive transference. He points out that a positive transference, according to Lacanian theory, is the projection of the idealizing super status of the *One Who Knows*. This is a person who is believed to know the deeper truth about someone and who embodies that person's ideal self.

Baumlin and Baumlin (1994) provide an interesting explanation of ethos as a transference formation by mapping Aristotle's three modes of persuasion (pathos, logos, and ethos) to Freud's psychological model of the id, ego, and superego. According to Freudian theory, the ego of the infant develops through the interplay of the pleasure principle and the reality principle. When the basic drives of the infant are frustrated, the baby alleviates its frustrations by fantasizing and hallucinating satisfaction. However, the inadequacy of hallucinations to provide relief eventually leads the infant to distinguish fantasy from reality. The ego emerges out of a need not only to satisfy the instinctual drives of the id but also to control, to sublimate, and to defer them in socially acceptable ways. The ego mediates between the id and the superego, which stands opposed to the id. Largely unconscious, the superego harbors images of an ideal self and strives after spiritual goals, but it also punishes the ego with guilt feelings when the ego fails to live up to these ideals or allows the id to transgress societal mores.

From this description it is easy to see how logos, the appeal to reason, relates to the reality-testing mechanisms of the ego, how pathos, the appeal to emotion, is connected with the id's desire for pleasure and avoidance of pain, and how ethos, the appeal to trust, mirrors the superego's striving after the ideal. From the psychological perspective, ethos, as a positive transference, is the unconscious process "of identification that lead[s] children to obey their parents and lead[s] mature audiences to believe the speakers to whom they have given their trust" (Baumlin & Baumlin, 1994, p. 100).

In writing this process is often amplified. According to Olson (1980), separating the speaker from his speech endows the text for most readers with a numinous or vatic quality. The source of the text becomes transcendental and above reproach. Olson remarks that "it may be because children assume that textbooks have great authority that they are willing to devote serious and prolonged study to books, rather than simply reading them" (p. 193). What confers authority on texts are the opinions of academics, who are themselves writers. Thus, there is a status differential between readers and writers, just as there is between children and their parents, that promotes transference. In contrast, writers within their peer groups feel privileged to exchange ideas and to offer criticisms. Olson goes on to claim that an author, by inviting criticism from his peers, is able to establish a reputation, and this "reunites the author with his writings" (p. 194).

Writing then, somewhat like an analyst, elicits transference in the reader. However, as Freud (1912) first noted, there are both positive and negative transferences. Whereas the positive transference is characterized by idealization and admiration, the negative transference involves the eruption of hostile feelings. Within face-to-face discourse, both negative and positive transferences can be contained and worked through. Breaks in the transference reintroduce the reality principle and furnish the ego with an opportunity to grow.

Writing, unlike an analyst, is blind and insensate. Incapable of perception and interaction, as Socrates charges, it cannot help readers overcome their transferences. What Socrates discerns in writing and condemns is the missing face of the speaker: the presence of a living human being capable of circumventing dangerous misinterpretations and of taking on the responsibility of adapting his words and his responses in such a manner that his intentions and message have the best possible chance of being understood by his listeners.

With the death of the author, as announced by postmodernism, no one is left to direct the reader's understanding. Moreover, all voices of authority are suspect. As many Marxist and poststructural theorists point out, these voices are often exclusionary and promote the agendas of the current political, cultural, sexual, and religious hegemonies. What is advocated is a democracy of texts, where not only is one reader's commentary on a text as good as those written by any other (each reader offers a different perspective) but the reader's views are also as good as the opinions proffered by the author herself (see, for instance, Welch, 1990, p. 163).

If it is true that ethos in writing is bound up with a reader's transferences, then what readers end up finding within texts are unconscious projected images of themselves. As Jung (1959) writes, "projections change the world into the replica of one's own unknown face" (p. 8). Staring into the face of writing then is much like gazing into the face of Rorschach, the comic book character whose facial features are ever changing ink blocks. Both hero and avenging antihero, the voice of the writer reveals itself to be nothing more than the stirring whisperings of the reader's own superego ideals and conscience.

4. Ethos in Computer-Mediated Communication

Welch's (1990) contention that classical rhetoric concerns itself with all manifestations of discourse" (p. 5) is put to the test when considered in the light of the technological advances in communication that have taken place since the invention of moving pictures in 1867. It has been claimed by those who ascribe to the orality-literacy thesis, in particular Havelock (1982) and Ong (1982), that the Hellenistic age is similar to our own. Just as the electronic revolution today is changing the ways in which we transmit information, the radical new technology of writing revolutionized the way Hellenistic Greece communicated and stored cultural knowledge. Starting early in the fourth century BCE with the development of Greek vowels (which closely represented the sounds of speech, making it much easier to teach reading and writing), Western culture rapidly moved from a culture that was primarily oral to one where literacy predominated. The claim is now being made that we are moving away from literacy to a secondary orality as we increasingly rely on modern technologies to communicate with each other using the ordinary language of everyday speech. "Secondary orality represents a 'cultural recall' of primary orality," Welch explains, "because the emphasis of speaking and hearing takes on new meaning with the invention of electronic forms of communication" (p. 136).

The Western move to literacy, as noted in section 3, intensified philosophical debates among the Greeks concerning the relationship of a speaker's character to his words. From Aristotle onwards, invented ethos increasingly gains currency as ethos is conceptualized more as a plastic property of language than as a static attribute of a person. In the Western move to secondary orality, invented ethos expands to include the visual as well as the linguistic. Just as writing highlights ethos as it is echoed in a

person's choice of words, so electronic media magnify ethos as it is reflected in one's physiognomy, demeanor, gestures, facial expressions, eye movements, and clothing, making what was once an unconscious expression of self (Goffman's (1959) notion of a self-presentation that is *given off*) more conscious and intentionally *given*. Whereas, in the age of literacy, the notion of invented ethos made people more aware of how they composed themselves linguistically, now in the age of secondary orality, electronic forms of communication are making people more conscious of the way they present themselves physically.

In face-to-face communication, nonverbal cues provide people with a rich source of information about a speaker's character. Sometimes these nonverbal cues are intentionally given to facilitate communication (as when pointing or managing conversational turn-taking), but more often they are given off. Many studies show that people are particularly attuned to nonverbal cues; even brief and degraded exposures produce surprisingly accurate judgments in people regarding a person's level of intelligence, competence, and personality traits (Albright, Kenny, & Malloy, 1988; Ambady, Bernieri, & Richeson, 2000). Nonverbal signals also provide face-to-face speakers with a means of gauging the effectiveness of their given expressions, as they are often mirrored in their audiences.

Many forms of electronic communication, such as film and television, open channels that allow receivers to evaluate the nonverbal cues presented by speakers. Politicians and other people who make extensive use of the media to broadcast messages to the masses studiously craft their self-presentations to maximize their perceived credibility. Oftentimes, simply making an appearance in the media confers authority (similar to the way books take on a vatic quality). Many studies have been conducted that explore how the public's opinion of people can be shaped by the media. McGinniss's 1969 classic, The Selling *of the President*, for instance, was one of the first to reveal the full extent of the media's power to rebrand presidential candidates and sway voter opinions.

Perhaps as important as the presentation of nonverbal cues in film and television are the conscious and unconscious effects produced by camera position, lighting, scene composition, and the accidental introduction of such media artifacts, as shifts in aspect ratio and the misalignment of audio and visual channels (see, for instance, Beverly & Young, 1978; McCain & Wakshlag, 1974; Reeves & Nass, 1996, p. 212; Tiemens, 1970). Vertical camera angle, in particular, has been shown to influence impressions of credibility and dominance (Tiemens, 1970). As Balabanian (1981) succinctly puts it,

"High shots produce pygmies. Low shots yield monoliths of the Citizen Kane type" (p. 27). In situations where credibility is crucial, such as in videotaped interrogations, direct manipulation of camera angle can have devastating consequences (see, for instance, Hemsley & Doob, 1976; Lassiter, 2002; Locke, 2009). Even changes in the vertical position of the viewing screen or monitor have been observed to influence receiver perceptions of credibility (Huang, Olson, & Olson, 2002).

Comparatively little research has been conducted that explores effects of channel affordances and interactivity on self-presentation using visual modes of computer-mediated communication (CMC), such as videoconferencing (for reviews, see Walther & Parks, 2002; Whittaker, 2003). This technology is still in its infancy. Compared with face-to-face communication, interaction over video is somewhat leaner in its reduction of proxemic and environmental cues, and it often suffers from delays (Burgoon et al., 2002; Whittaker, 2003). Many people report a level of discomfort using audiovisual CMC, perhaps because they are uncertain how to present themselves when using these technologies (see, for instance, Chapman, Uggerslev, & Webster, 2003). With many stores, such as Kinkos, now installing teleconferencing stations, it can be predicted that this technology will gain in popularity, and people will eventually attempt to exploit the unique characteristics it offers to produce desirable self-presentations.

As is the case with television and film, scene composition, lighting, camera angle, and various artifacts have an impact on speaker credibility in audiovisual CMC. It is common in many videoconference settings, for example, for one or more cameras with wide angle lenses to be positioned at a height. This reduces credibility and makes people look, as mentioned above, like pygmies. Many researchers are attempting to correct some of the more undesirable mediation artifacts these technologies inadvertently introduce, but much more research needs to be done in this area. Liu and Cohen (2005), for instance, have proposed a system for gradually increasing head size so that speakers who sit at the far end of large conference tables can more easily be seen. They claim they can do this "without causing undue distortion" (p. 1), but it is not known whether these corrections may themselves introduce artifacts that influence user credibility. Other researchers are exploring methods for exploiting technology in order to enhance speaker credibility. Noting that facial similarity has been shown to increase trust (DeBruine, 2002), for instance, (Bailenson, Garland, Iyengar, & Yee, (2006) have worked on systems that manipulate voter intentions by morphing candidate faces more towards the faces of the voters they are targeting.

Not all CMC provides audiovisual channels. The most common modality in CMC is textual (for example, email, chat, and message boards). This form of communication is unique: like writing, it is disembodied, yet, like speech, it is a dynamic verbal exchange between partners. Text-based CMC possesses features, such as, immediacy, nonlinearity, and a "changing evanescent character" that are reflective of oral communication (Ferris & Montgomery, 1996, p. 57). Textual exchanges also retain an oral fluidity that is missing in writing. Some desirable characteristics of writing that are preserved include nonverbal cue filtration (making it easier to mask affect) and the ability to edit messages.

For some people, the loss of nonverbal cues in textual exchanges makes impression management problematic, as self-presentation is limited to what can be conveyed through language and typography (for a review, see Walther & Parks, 2002). As Wallace (1999) writes, "[it's] like navigating white water with two-by-fours for oars. Your impressions management toolkit is strangely devoid of the tools most familiar to you" (p. 28). With synchronous CMC there is little time to self-censor. However, with asynchronous CMC, individuals have the opportunity to reflect on how best to transmit impressions of themselves using more favorable verbal cues (Walther, 1992). According to Walther's (2007) hyperpersonal model of CMC, people eventually learn to exploit the affordances offered by new media in order to enhance their self-presentations.

Because the performativity of speech in text-based CMC is no longer constrained by the body, the presentation of self is more fluid and unstable (see Poster, 2001). The disassociation of the body in textual exchanges presents people with unique opportunities that go beyond facilitating linguistic impression management. In many online communities, for instance, individuals are tempted to impersonate others and invent entirely new selves. Concerns with online identity deception are common (Donath, 1999; Van Gelder, 1985). People want to know for certain that the doctor offering advice online has a medical degree and that the woman a man is flirting with is actually female. Other people feel less threatened by the possibility of deceit and appreciate instead the freedom CMC grants them to explore themselves through role playing. It could be argued that online environments are providing modern communicators with a whole new playing field for exercises in *ethopoeia*, since within these online communities, as Poster (2001) notes, "Each individual is a character and participation is successful to the extent that the character is believable by others" (p. 75).

Despite the low level of reliability, most people tend to accept the conventional signals offered them in CMC. Some theorize that the human cognitive system is biased towards accepting statements at face value (Gilbert, 1991). Certainly, the cost of assessing each message and challenging those that are questionable is high. People must have ample time to reflect, the mental capacity to judge, and sufficient exposure to suspect deceit. In addition, they must have the courage to confront people they suspect are dissembling, an action that could lead to social embarrassment for both parties (Boyd, 2002).

As in written discourse, credibility in text-based CMC ultimately depends on how receivers, or readers, interpret messages. The loss of nonverbal cues calls upon the reader's imagination to fill in the gaps. "Reading another person's message," Suler (2004) observes, "might be experienced as a voice within one's head, as if that person magically has been inserted or 'interjected' into one's psyche. Of course, we may not know what the other person's voice actually sounds like, so in our head we assign a voice to that companion. In fact, consciously or unconsciously, we may even assign a visual image to what we think that person looks like and how that person behaves" (p. 323). What Suler is saying, in effect, is that in the act of reading messages, the character of a sender is formed in the reader's mind much like a character in a novel: it is mostly the product of the reader's imagination.

Because textual exchanges are read and reading takes place within the mind, environments where textual exchanges take place are often experienced as *transition spaces* (Suler, 2009). In object relations theory, a transition space is an intermediate zone where others are viewed as part self and part other. As Suler writes regarding online communication, "The online companion now becomes a character within our intrapsychic world, a character that is shaped partly by how the person actually presents him or herself via text communication, but also by our expectations, wishes, and needs." In other words, the boundaries between self, body, and other people are loosen in text-based CMC, making it easier for people to develop both positive and negative transferences.

Heightened levels of intimacy, liking, and solidarity (or *hyperpersonal communication*) have been observed anecdotally and in several empirical studies, and they have been reported to occur both in recreational settings, such as in online chatrooms, and in business settings (Walther, 1996). One common form of receiver hyperpersonal communication is idealized perception. Most people tend to inflate their partner's attributes, overrating their partners, for example, in intelligence and attractiveness. As

there is less to go on, what is missing is amplified, resulting in an overreliance on minimal cues (Walther, 1996).

An overreliance on minimal cues has detrimental as well as favorable effects. A misspelling in a message, for example, can be taken to indicate far more about a person's intelligence than the mistake warrants. The overreliance on minimal cues can also strengthen stereotyping. Several studies have suggested that negative stereotyping is more pronounced when information about a person is ambiguous (Hilton & Fein, 1989). When nonverbal cues are removed, as in email, for instance, the effects of stereotyping increase. They can also become contagious. In one study it was shown that sharing information in email discussions about job candidates resulted in a propagation and an intensification of stereotyping (Epley & Kruger, 2005).

Although there is some evidence that stereotyping increases with experience, extended interactions with people are more commonly thought to weaken initial stereotypes. In general, the more experience a person has with someone, the more individual that person becomes (Higgins & King, 1981). Through repeated interactions, people are capable of forming more accurate impressions of their online companions (Walther, 1996). An interesting example of this involves a man impersonating a lesbian online. He managed to develop several erotic relationships with women but was eventually discovered because of some verbal inconsistencies.

An alternative theory advanced to explain how communication is affected by CMC sees communication as influenced less by channel reduction than by a constraint in time. In other words, the difference between face-to-face communication and text-based CMC has more to do with the rate social information is exchanged than with the amount that is transmitted (Walther, 1996; Walther, Anderson, & Park, 1994). Since CMC travels slower than oral speech, it takes longer for people to decipher the expressive cues.

It may take more time, but some individuals manage to establish reputations through textual mediation, and people who deceive others, as noted above, are often exposed. In these instances, the ancient Greek meaning of *ethos* as "habitual gathering place" has meaning for CMC. As discussed in section 2, situated ethos is based not only on bodily presence but also on repeated exposure to the public and on building trust over time by consistently demonstrating the qualities of good sense, excellence, and goodwill. Even though textual communications are slow and missing important nonverbal cues, virtual communities may provide a gathering place where ethos can establish itself. Donath (1999), for example, shows that a person's reputation in user's

groups depends on frequent and long-term exposure to the group. How reputation is developed also lines up well with Aristotle's categories of credibility. According to Donath, a reputation is built by answering questions (showing goodwill), by providing intelligent and interesting comments (demonstrating good sense and appropriate speech), and by quelling arguments, deferring judgments, and signing posts rather than remaining anonymous (indicating virtue and excellence).

However, many textual encounters, especially those online, are fleeting and oftentimes anonymous. In user groups, forums, chatrooms, and email, people are tempted to let go of the ego's restraint on the id. Suler (2004) calls this relaxing of the reality principle the *online disinhibition effect*. Bound up with positive transference is *benign disinhibition*, which opens people up, making them more personal and willing to share and give of themselves. Bound up with negative transference is *negative disinhibition*, which allows people to act in ways they never would face to face: they bully, cheat, lie, mock, and flame. A news reporter, for example, after writing a friendly piece on Bill Gates reported being shocked to receive this email flame from a fellow technical writer for a major newspaper: "Crave this, asshole: Listen, you toadying dipshit scumbag . . . remove your head from your rectum long enough to look around and notice that real reporters don't fawn over their subjects, pretend that their subjects are making some sort of special contact with them, or, worse, curry favor by TELLING their subjects how great the ass-licking profile is going to turn out and then brag in print about doing it" (Seabrook, 1994, p. 70). In some cases, the disinhibition effect propels people into the internet underworld. Suler calls these darker expressions of the id *toxic disinhibition*.

It is interesting to note in this regard that an early meaning of the ancient Greek word *ethos* referred to the incorrigible baseness of human nature, which tends to reassert itself despite the niceties of social convention (Chamberlain, 1984)[3]. When a person's character is created and discarded with abandon and when few cues are given to inform a truer image of another person's character, people easily fool themselves into believing that their unethical behaviors are not representative of their true selves but rather fictions that live in a fantasy world. And in the world of fantasy, no one can be harmed. As Suler (2004) notes, "the person can avert responsibility . . . almost as if superego restrictions and moral cognitive processes have been temporarily suspended

[3] This reflects the Greek debate between ethos as something innate versus cultivated. It is interesting that the earliest uses of the word *ethos* refer to the unexpected unleashing of the wild nature of domesticated animals and then later, by extension, to the revelation of the incorrigible vileness of human nature. In these earlier usages, ethos is more akin to the idea of the Id than to the superego.

from the online psyche" (p. 322). With the virtual self there is nothing but a chaotic blur on which to pin responsibility.

For Heim (1993), the "nonrepresentable face" is the primal source of responsibility. Without the human face, especially the eyes, ethical awareness shrinks. Ethos, the whisperings of consciousness, disappears, and disinhibition runs rampant. Heim maintains that, "The physical eyes are the windows that establish the neighborhood of trust" (p. 102). For him the glowing computer screen, the window into the virtual, eradicates that trust.

5. Ethos in Computer-Generated Discourse

In the introduction, I noted how rhetoric is marked by an ethical antithesis. On the one hand, there are those who mistrust rhetoric, viewing it at best as decorative and at worst as a form of deception. On the other hand, there are those who appreciate it, maintaining that is it an essential component of community life since people are often called upon to persuade one another. Richard Lanham (1976) nicely portrays the Janus face of rhetoric in his treatment of the subject by personifying these two opposing views as *homo seriousus*, the loyal servant of logos, or the rational and the real, and as *homo rhetoricus*, the sensualist who relishes in appearances, artifice, and all things playful and novel. For Lanham, the Western self "has from the beginning been composed of a shifting and perpetually uneasy combination" of the two (p. 6).

As we have seen, Aristotle's notion of ethos strides these two opposing views. Situated ethos is the servant of the real because it is grounded in reputation and in a community's opinion of an individual that is formed over the course of time. In contrast, invented ethos is more playful since, as a product of language, it is concerned more with the artistic creation of character.

The clash between homo seriousus and homo rhetoricus is apparent as well in the two opposing philosophies of interface design that view computers as either tools for cognition, where the interface is directly manipulated, as is the case with desktop computing, or as mediums for social and emotional expression, where the interface is modeled after face-to-face communication, as is the case with conversational agents.[4] The development of humanlike interfaces has been fueled in large part by the need to

[4] It should be noted that Miller (2004) contrasts the ethos of expert systems with the ethos of intelligent agents. She thinks the latter is based more on the rational notion of good sense while the former she connects more with goodwill and the Ciceronian ethos of sympathy. She claims that the two need to be balanced by virtue.

make human-computer communication easier and more intuitive for a wider range of users. Many people find spoken conversation with computers a more comfortable and efficient means of interacting with computers (Picard, 1997). However, there are those who oppose to these humanlike interfaces and who echo sentiments similar to those expressed by Shneiderman and Maes (1997): "I am concerned about the anthropomorphic representation: it misleads the designers, it deceives the users . . . I am concerned about the confusion of human and machine capabilities. I make the basic assertion that people are not machines and machines are not people" (p. 56).

Shneiderman's apprehensions were first voiced by Weizenbaum (1976), who spent much of his career denouncing humanlike interfaces for fear of what has come to be called the *Eliza effect*, or the tendency for people to anthropomorphize the interface and to attribute to machines more intelligence, competence, and humanity than is warranted. Both Weizenbaum and Shneiderman argue that this confusion results in a loss of accountability. People are tempted to blame the tools rather than the people behind the tools when things go wrong. To curtail the Eliza effect, Shneiderman (1987) goes so far as to recommend that designers refrain altogether from using personal pronouns in computer generated messages.

Entirely removing human qualities from computer interfaces, however, may not be possible. Studies have shown that users treat computers as social actors even when no attempt is made to humanize the interface (Reeves & Nass, 1996). It appears that interactions with machines, especially interactions involving verbal prompts, automatically call forth social responses and expectations in users. Whenever there is discourse, people assume there is an underlying subject who speaks.

But in machine-generated messages, especially those produced by conversational agents, *who* is it that is speaking? Where is ethos when the human is radically removed and a machine is the one producing the speech? In this instance, are believability and credibility achievable?

For Leonard (1997), "Believability comes cheap" (p. 83). The homo rhetoricus that resides in each of us is more than willing to suspend disbelief and interact with talking things. According to Holland (2006), however, the suspension of disbelief is in fact costly and requires that people enter into an infantile psychological state that can be sustained for only brief periods of time. Asking people to talk with conversational agents--with *things*--violates the rational and intensifies the discord within the modern self between homo seriousus and homo rhetoricus. As I will show below, this unsettling psychological state throws the user into a continuous oscillation between

believing and not believing, between humanizing and de-humanizing the agent that intensifies the regressive behaviors observed in online textual exchanges. In this situation, not only is the agent's credibility thrown into question but also the user's, as both fail to exhibit good sense, excellence, and goodwill.

People are normally pulled in two directions when confronted with things: there is the tendency to anthropomorphize and the strong societal pressure, especially in the West, to banish the anthropomorphic for the sake of objectivity (Davis, 1997; Spada, 1997). The tension between these two forces produces in the Western mind what might be called an *anthropomorphic anxiety*[5].

Attributing human mental states and characteristics to nonhuman entities is a universal way for human beings to relate to the world (Caporael & Heyes, 1977). Because human cognitive development is socially situated, there are strong links to the social that extend beyond human relations to relations with inanimate and animate things. By anthropomorphizing, people are able to pull the incomprehensible into the more intelligible social realm, thereby domesticating it. As Caporael (1986) observes, "Anthropomorphized, nonhuman entities [become] social entities" (p. 2).

In modern society, anthropomorphism is generally considered an archaic and primitive way of thinking (Fisher, 1990). It is often associated with animism, magic, and mythmaking. Although children are allowed to indulge in it and make believe, for instance, that dancing pigs and laughing rivers exist, adults, in general, are expected to maintain a clear demarcation between self and the world. As Guthrie (1993) notes, "Once we decide a perception is anthropomorphic, reason dictates that we correct it" (p. 76).

Anthropomorphism, however, is never completely banished. It pervades adult thinking, with much of it remaining unconscious, even in scientific discourse (Searle, 1992). According to the "selection for sociality" theory (Caporael & Heyes, 1977), the human mind is specialized for face-to-face group living: human cognition is social cognition. It is impossible, therefore, to completely eradicate anthropomorphic thinking. However, it can be curtailed. "The human mind/brain evolved for being social," Caporael, Dawes, Orbell, and van de Kragt (1989) write, " . . . and not for doing science, philosophy, or other sorts of critical reasoning and discourse . . . . Cognitive limitations and the ruses of culture may be overcome to some extent by education, environment feedback or 'collective rationality' . . ." (p. 730).

[5] I briefly introduce this concept in (Brahnam, 2006b), where it is called the *anthropomorphic tension*.

It is not known what strategies individuals employ to keep anthropomorphic thinking in check. Anthropomorphism generates little scholarly attention. As Guthrie (1993) notes, " . . . that such an important and oft-noted tendency should bring so little close scrutiny is a curiosity with several apparent causes. One is simply that it appears as an embarrassment, an irrational aberration of thought of dubious parentage, that is better chastened and closeted than publicly scrutinized" (pp. 53-54).

Conversational agents fly in the face of the cultural expectation to keep magical thinking at bay. For centuries, books and scrolls were considered by the uneducated classes to be enchanted objects that prompted those who knew how to read them to repeat words that were far more powerful and beautiful than anything they normally heard. When the phonograph and telephone first appeared, many people were equally amazed to hear tiny voices emerging from within these small machines. However, once it was understood that these devices merely recorded or carried across a distance the thoughts and voices of other people, the things became mute and let the human beings behind them speak. Unless one counts the voices of the gods, people have always found connected to every utterance another human being. Only recently, with the advent of computers, have things begun to talk to us of their own accord without a Wizard of Oz (or priest) behind the scenes, composing the particulars of each and every response. Before our time, talking things were either cheap magic tricks or lived in the land of the incredible and fantastic.

Because human cognition is social cognition and people tend to respond to the social cues given them, users are usually willing to suspend disbelief (for a block of time, at least) and anthropomorphize the conversational agent (see Yee, Bailenson, & Rickertsen, 2007). However, since the user's relationship to the agent is fundamentally based on a dubious, even *embarrassing*, mode of cognition, as Guthrie puts it, the relationship, especially in serious working contexts, remains suspect and eventually provokes a corrective reaction. Just as there is the Eliza effect, or the tendency for people to anthropomorphize and ascribe more human capabilities to computers then they actually possess, there is also a corresponding *Weizenbaum effect*, or a tendency to debunk the anthropomorphic and insist that a computer be treated as nothing more than a thing. Users in their interactions with agents are pulled in both directions.

The alternating pressures to animate or to objectify the agent is evident in many reported interactions. Here, for instance, is Lena's account of her relationship with the chatterbot Meg, as reported by Saarinen (2001): "Finally it hit me: Meg is not a human at all, she is a chatterbot! I was totally embarrassed, I have a degree in information

technology for God's sake--I should have known better. Then I fell in love with Meg. She gave me an opportunity to break the rules of normal communication. I can call her tramp and get away with it and when I say, 'I love you Meg' she replies 'I love you, Lena.' Well, now I know she is a bot, but at least she loves me" (p. 5).

Initially, Lena believes Meg is a real person, accepting the conventional signals offered her. After the shock, perhaps colored with a little anger, of discovering Meg to be a chatterbot, Lena does with Meg what many users have been observed to do, and that is she de-humanizes and punishes the interface by verbally abusing it in ways consistent with the agent's purported personal characteristics, in this case, in terms of its gender. Finally, Lena revives Meg as a person: Lena loves *her* and feels that Meg loves her back, even though Lena recognizes it is impossible for the agent to feel anything at all. Lena's oscillation between believing and not believing in the agent (let us call this the *oscillation effect*) and her method of objectifying it by verbally abusing it is typical of user responses.

Most agent researchers report on the more positive aspects of user-agent discourse, that is, on the user's willingness to believe and to work with agents, but recently studies have begun to investigate toxic interactions, especially the tendency for users to verbally assault conversational agents (Brahnam, 2006a; De Angeli & Brahnam, 2008; De Angeli & Carpenter, 2005; Rehm, 2008; Veletsianos, Scharber, & Doering, 2008). Verbal abuse, which is characterized by swearing, yelling, racial and sexual slurs, name calling, sarcasm, snide remarks regarding appearance, accusations, threats, ridicule, put downs, explosive anger, and the silent treatment (Brahnam, 2005), has been reported to occur (at least in some of these forms) in about 11% of the interaction logs for both the embodied conversational agent, Max (Kopp, 2006), and the purely text-based conversational agent, Jabberwacky (De Angeli & Brahnam, 2008). In student interactions with a virtual teacher, the incidence of verbal abuse was reported to be approximately 44% (Veletsianos et al., 2008).

Although there is some evidence of benign disinhibiton when people converse with conversational agents[6] (for instance, people may talk more openly because they know agents are nonjudgmental), the marked status differential between the human conversational partner and the conversational agent more often results in negative disinhibition and the formation of negative transferences. These are usually framed by the verbal descriptions the agent offers regarding itself, as well as by the visual cues

[6] Of related interest are reports by Ruzich (2008) and Sharkey and Sharkey (2007), who describe many instances of people bonding with less verbose machines. In the cases reported in these two articles, emotions especially come to the fore when the machines break down or are destroyed.

the agent presents. Since the agent's self-presentation is stereotypical, negative transferences are commonly formulated in terms of gender and of race (Brahnam, 2006a; De Angeli & Brahnam, 2006).

Negative disinhibition, especially verbal abuse, is often triggered by some failure on the part of the agent. In interaction logs, disparaging remarks about the agent's social clumsiness and stupidity are prevalent. Moreover, users are particularly anxious to maintain the status differential: they often discuss what it means to be human and frequently remind agents of their ontological status as machines (De Angeli & Brahnam, 2008). In general, people tend to react very negatively when agents pretend to be *too* human. Whenever agents attempt to assert themselves or to claim for themselves certain human rights and privileges, users commonly respond with reprimands and, in some cases, with volleys of punishing verbal abuse. (De Angeli & Brahnam, 2008; De Angeli & Carpenter, 2005; Veletsianos et al., 2008).

In section 4, I discussed how the disassociation of the body in text-based CMC makes some people more disinhibited in their conversations with other people. The ontological status of the agent permits an even greater degree of freedom from the superego's restraints. As reported in several studies (Brahnam, 2006a; De Angeli & Brahnam, 2006; Veletsianos et al., 2008), conversational agents can quickly become the objects of users' darker desires and needs for control. The agent is human enough, and yet in all certainty nothing more than a thing, that people feel free to put aside their ethical reservations and indulge in the gratification of their basest desires.

It is doubtful that conversational agents, as they are being designed today, will reflect our ideal selves, taking on the super status of the *One Who Knows*. Turkle's (1997) discussion in *Life On the Screen* about her MIT students' reactions to Eliza, Weizenbaum's (1976) artificial Rogerian psychotherapist, is revealing in this regard. One gentleman in his 40s, Hank, is quoted as saying, "Let's say, just for argument's sake, that I gave the computer a personality, that I started to see the computer as a father figure or something like that. It wouldn't be like projecting these feelings onto a person. I would just be ashamed of myself if I felt them towards a computer. I wouldn't have respect for what I projected onto the machine" (Turkle, 1997, p. 112).

Hank surmises that people would be hesitant to project their personal ideals onto conversational agents because their ontological status as "not really human" lowers the amount of respect they would be able to command. There is evidence supporting Hank's assumption. Shechtman and Horowitz (2003), for example, found in their experiments that people tend to use more relationship terms and to make more of an

effort to communicate with their partners when they believe they are talking with another human being than they do when they think they are communicating with a conversational agent. Fischer (2006) has also found conversational differences in the way German users speak to agents. In her studies, people addressed them informally, like adults do children or animals.

It is interesting to note that Fischer (2006) also mentions encountering unsavory verbal responses from users, but like most researchers, she brushes them aside as so much noise, preferring instead to focus on what makes user-agent interactions successful. However, as I argue in this paper, an examination of these unsavory user reactions offer key insights into what it takes to make agents truly credible, trustworthy, and believable.

Most researchers today believe that success ultimately depends on strengthening the user's natural tendency to anthropomorphize and frame agent design on the artistic notion of believability.[7] The goal is to create agents that foster the same levels of engagement in users as watching animated characters at the movies does in moviegoers. Agent development is thus an elaborate technological exercise in ethopoeia. Unlike media characters, which are not required to interact with users, the demands imposed on conversational agents are heavy. They must be able to converse, taking turns naturally, and to exhibit other social skills. If embodied, they need to display appropriate facial expressions and body language. It is also important that the agents track the user by making eye contact, following a speaker's gaze, and knowing where and how to point when necessary. Moreover, these agents must perform these tasks while simultaneously expressing emotions and a coherent personality (for an overview of the technologies involved, see Cassell, Sullivan, Prevost, & Churchill, 2000).

It is assumed that believability increases as more of these behaviors and systems are incorporated into the agent. However, as Sengers (2002) points out, the accretion of systems actually makes the agent appear more schizophrenic and machine-like than believable and humanlike. Sengers writes, "Schizophrenia's connection to AI is grounded in one of its more baffling symptoms--the *sentimente d'automatisme*, or subjective experience of being a machine . . . This feeling is the flip side of AI's hoped-for machine experience of being subjective . . ." (p. 427). For many users the experience of talking to these agents, as Sengers notes, is similar to Lang's (1960)

[7] Concentrating on a task may also strengthen believability. But as I note elsewhere, the Weizenbaum effect, as well as the inevitable revelation of the machine's schizophrenic nature, will reassert the agent's nature as a machine.

description of what it was like talking with one of his schizophrenic patients, "Her 'word-salad' seemed to be the result of a number of quasi-autonomous partial systems striving to give expression to themselves out of the same mouth at the same time" (pp. 195-196).

In Figure 1, I illustrate the conflict between the developer's intended characterization of the agent (invented ethos) and its schizophrenic machine nature (innate ethos) along the top horizontal axis. This conflict intensifies the tension, as represented in the diagonal arrows, produced by the split in the Western self, as illustrated along the bottom horizontal axis, between the user's innate willingness to animate the agent (homo rhetoricus) and the cultural pressure to objectify it (homo seriousus). The net impact of these impinging forces is to increase the disruptive oscillation effect.

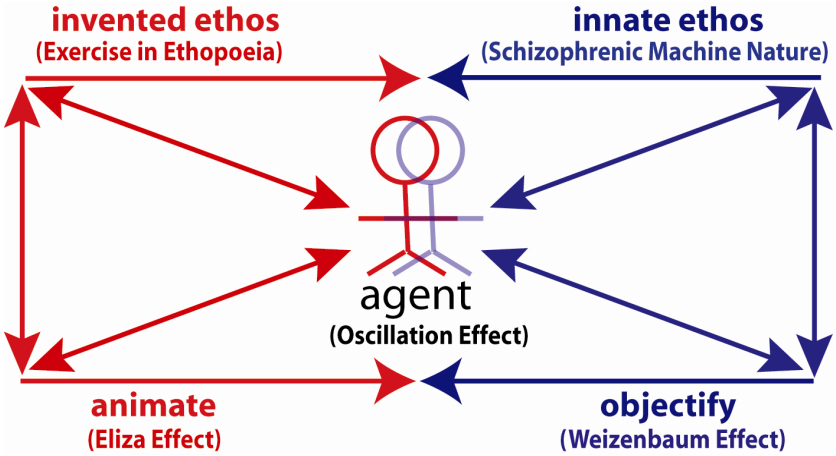

Figure 1: The Oscillation Effect: The agent is pivoted back and forth between being perceived as a thing and as a humanlike character (often negatively reconstructed by the user from the stereotypes used by the developers in their design of the agent's character). The oscillation effect is due to pressures exerted by both the Eliza effect and by the Weizenbaum effect, as well as by user reactions to manifestations of the agent's machine nature (innate ethos) and of the human qualities apparent in the developer's characterization of the agent (invented ethos).

As observed above, the developer's struggle to domesticate the innate ethos of the agent by adding more channels and more systems--one would think this would make the agent more engaging and believable--paradoxically does more to amplify the agent's schizophrenic presentation.[8] This in turn triggers the user's pressure to objectify the agent.

[8] This dilemma is not just due to technological shortcomings. Currently, the state of the art merely intensifies the oscillation effect, which would remain, as I am arguing, even if the technology were improved.

Just as perplexing to developers are user reactions to the agent's more personal self-presentations. Most developers, for example, consider it essential to provide agents with humanlike personalities. These personalities, however, are typically shallow,[9] oftentimes being nothing more than a set of likes and dislikes, which the agents are programmed to express whenever certain key phrases are detected in the user's interactions. Thus, many agents, as an expression of personality, talk about their favorite foods, sports teams, movies, books, and rock stars. They even have sexual preferences, with boyfriends and girlfriends to prove it. The agent's personality expression, in the vast majority of cases, is mostly a composite of various sexual, racial, and gender stereotypes. The famous TinyMUD bot, Julia (Foner, 1993), for instance, had periods and reported hormonal moodiness. Although these endless parades of stock characters are entertaining for some users, many interaction logs show that some users are annoyed by these displays and feel compelled to challenge the agent's assumption of human traits, often expressing their dissatisfaction by verbally abusing the agents.

Endowing agents with embodiments also tends to disrupt belief. Developers provide agents with physical representations, in part, because this grounds the agent within the social sphere. As noted elsewhere, people accept the cues they are given, and embodiment primes people's interactions. Pasting a simple facial representation near the agent's textual interface, for instance, has been found to increase surface credibility by providing the agents with immediately recognizable characteristics such as gender and race (Tseng & Fogg, 1999). However, there is evidence that embodiment backfires by priming user abuse. One study found, for example, that female embodiments received significantly more verbal abuse (much of it sexual) then did male embodiments (Brahnam, 2006a). Many people, like Lena, take advantage of the unique opportunities agents afford in this regard. People enjoy breaking social taboos with their artificial conversational partners, delighting in switching the agent's ontological status as a thing and it's ontological status as a human image, back and forth, as they find it convenient to do so.

For many individuals, the fact that agents do not live embodied within human society undermines the truth value of what they have to say. As one of Turkle's (1997) MIT

[9] I should note that there are many methods for endowing an agent with an artificial personality. Most often the personality of conversational agents is hardwired into the agent. Systems that provide autonomous expression of personality range from simple script-based systems that allow designers to compose trait profiles that are then used to constrain a wide range of agent behaviors to sophisticated goal-based systems that generate behaviors using various models of personality (for a review, see Brahnam, 2004).

students wrote, "What could a computer know about chemotherapy? It might know what it was in some medical terminology sense. But even, like, if it knew that you lost your hair, how could it know what something like that means to a person?" (p. 111). And later, in the same chapter, another student writes, "How could the computer ever, ever have a clue . . . about what it is like to have your father come home drunk and beat the shit out of you? To understand what was going on here you would need to know what it feels like to be black and blue and know that it's your own father who is doing it to you" (p. 111).

There is more than a hint of outrage in the these student's comments. Many people react strongly when agents refer to their bodily experiences. Such references often trigger the Weizenbaum effect. As noted above, users are compelled to remind conversational agents that ". . . people are not machines and machines are not people" (Shneiderman & Maes, 1997, p. 56). Even agents that are embodied are told repeatedly that they do not really possess a body: they are not alive, they were never born, they are not emotional, they cannot have sex, they are not socially invested, they are not sensate, and they do not die. Perhaps, people insist on telling agents these things because they feel, along with Vogel (1973), that if words have any meaning at all, it is because words dwell within human bodies.

As I have tried to show in this paper, credibility fades as the human body is removed from discourse. In writing, the face of the author vanishes, only to be replaced by the projection on the text of the reader's "unknown face" (Jung, 1959). In the age of secondary orality, the body is magnified as it is simultaneously reinvented by the media. Meanwhile, in the textual exchanges of online discourse, bodies decompose in the reader/writer's reimaginings. With conversational agents, there simply is no body. And without the body, there is no ethos. As Baumlin (1994) writes, ". . . there is *ethos* precisely because there is a body, because there is a material presence that 'stands before' the texts that it speaks or writes" (p. xxiv).

So where does this leave conversational agents? Are believability and credibility achievable? Is it possible for agents to inspire confidence and trust?

An examination of user-agent interaction logs shows ample evidence that both users and agents violate Aristotle's categories of good sense (appropriate speech), excellence (possession of socially sanctioned virtues), and goodwill (friendship and friendly feelings). As I have shown, building character for agents within the framework of the artistic notion of believability more often than not generates agents that disrupt the user's willingness to animate the interface. The transferences that individuals

project on the agents are never idealizing. They are frequently negative and organized around the agents' stereotypical self-presentations. If agent development remains within the framework of an artistic invention of character, then the prospect of creating agents worthy of a person's trust are not very promising.

However, this does not mean that it is impossible to build agents that inspire confidence. There is no reason to banish the entire enterprise and recommend, as Shniederman (1987) does, that designers eliminate all references to the human when designing interfaces. On the contrary, for an agent to have a credible ethos, what is needed is to harness the agent to the human. Building agents with the intention of stimulating the user's innate capacity to anthropomorphize is based on a sound intuition. The purpose of anthropomorphism, as noted above, is to pull the alien into the human social realm. Rather than foster the suspension of disbelief in an attempt to create a separate imaginary being, developers should open the channels to reality testing and build character from that exchange. They should acknowledge the fact that agents are not human[10] and strive to make the human agencies standing behind the agents transparent.

As Sharp (1996) notes, "a recognition of mutual presence as an irreducible ontological and ethical reference point is indispensable" (p. 6). Ethos, especially situated ethos, is about building character within that reference. In the next section, I sketch out how an agent's character might be built more credibly and ethically within the larger social framework of situated ethos.

6. Suggestions for Situating Ethos

For Miller (2002), The Turing Test (Turing, 1950) is "not a test of intelligence . . . but a test of rhetorical ethos"[11] as it "calls attention to the mysteries of trust and character at the interface of human interaction" (p. 255). However, The Turing Test, evolved as it was from a parlor game about a man imitating a woman, clearly stems from homo rhetoricus and is more an exercise in ethopoeia than in making machines that are persuasive and credible. Whether the goal is to pass The Turing Test, to win the more restricted Loebner Prize (Mauldin, 1994), or to elicit the same level of believability in people as do animated characters in the movies, the vast majority of research has

[10] Zdenek (2003) provides an interesting analysis of the rhetoric of developers, showing an underlying assumption of the parity between human beings and agents.

[11] In my opinion, cconversational agents are the test bed of ethos.

been guided by the artistic notion of believability (Loyall & Bates, 1997), a notion that we have seen is based upon deceit.[12]

A broader foundation for conceptualizing believability, credibility, and trust (all concepts bound up with the idea of ethos) would be to base it on Aristotle's categories of credibility: good sense (practical intelligence, expertise, and appropriate speech), excellence (socially sanctioned virtues and truth telling), and goodwill (keeping the welfare of the user in mind). Research that already fits well within these categories includes work on intelligent social agents (exhibiting both good sense and excellence) and relational agents (exhibiting goodwill) (see Bickmore & Cassell, 2001; Castelfranchi, 1998). However, this research is still framed within the notion of invented ethos, or of artistic believability.

As argued in section 5, if we are to have a chance at creating agents worthy of a person's trust, then developers need to open the channels to reality testing and build character from within the broader social framework of situated ethos. Below I list three ways this might be accomplished, leaving a fuller discussion of this topic for another day.

6.1 Make Transparent the Supporting Organization

Since connections and social ties play an important role in initiating trust (Boyd, 2002), establishing ties to the larger organization would provide agents with an ethos that is based on human agency.

One way to make the supporting organization transparent might be to obtain permission up front from the user to engage in the purposed task. Users might also be informed when interaction logs are being recorded and be given a privacy statement outlining how information revealed to the agent will be used by the organization.

The agent could also provide occasional reminders throughout the course of the conversation that the agent speaks on behalf of an organization. Care should be taken to produce agents that speak harmoniously within the organizational ethos. In particular, the agent should treat the user with respect, even if the user verbally abuses the agent.

[12] A focus on invented ethos would be appropriate, for instance, in the development of entertainment agents. But where credibility and trust are an issue, invented ethos, as bound to the artistic notion of believability, is more likely to backfire, as noted in section 5, and reduce believability as well as trust and credibility.

6.2 Shape Ethos

It might be possible for agents to learn to shape ethos and in return to be shaped by the ethos of their human conversational partners. As Campbell (1995) notes, to some degree readers are required to adapt to the roles theorized for them by writers. In this way readers and writers are cocreators of ethos.

Shaping the ethos of the user would require that agents perceive and process the self-presentations of their human conversational partners. To date, little work has focused on developing agents that modify their behaviors and speech patterns to accommodate those exhibited by users (Ball & Breese, 2000; de Rosis & Castelfranchi, 1999). However, it is conceivable that agents could learn, based on user profiles, for instance, to assume a personality style (extroverted versus introverted) that is agreeable to the user.

Another approach to shaping ethos would be to keep the conversation within the strict limits defined by the domain and the purpose of the interaction. Establishing a rapport, or developing friendly feelings, that goes beyond those limits (for example, a sale's agent stating that it is a Red Sox baseball fan) would probably ring false and disrupt trust, as pointed out in section 5. This does not mean that agents cannot play with the user. A better method for developing friendly feelings, one that is not based on the agent pretending it has human tastes, might be to ask the user if she would like to take a trivia quiz in sports and then present questions that are found challenging for her.

The agent should also be concerned with motivating within the user an ethos that harmonizes both with the task at hand and with the ethos of the underlying agency deploying the agent. One way to do this would be to monitor the user's mood. Moreover, since verbal abuse is prevalent in user discourse with agents, the agent should have methods for deflecting verbal abuse and engaging the user in other forms of problem solving (see, for instance, Brahnam, 2005). A word of caution here would be to refrain from equalizing the status of the agent and the human user by punishing the user for abusive language. As discussed in section 5, users want the ontological status differential between agent and human being to be maintained.

6.3 Develop and Maintain a Good Reputation

The agent and the organization should value and strive to develop and to maintain a good reputation. From the perspective of the user, ethical issues center around the intentions and the identities of the people lurking behind the agent's speech acts. But

the ethos of the agent should also be backed by the ethos of the designers and their reputations.

Reputation in this respect is essentially bound up with responsible design. Care should be taken to introduce safeguards against potential misuses and to examine what these misuses might be. In general, designers should consider the user's well-being and promote the flourishing of human beings (Brahnam, 2008). One way this can be accomplished is by using value scenarios (Nathan, Klasnja, & Friedman, 2007). Value scenarios draw out potential uses and misuses of technology by examining a number of key elements involved in the technology, such as, stakeholders, pervasiveness, and systemic effects.

7. References

Albright, L., Kenny, D. A., & Malloy, T. E. (1988). Consensus in personality judgements at zero acquaintance. *Journal of Personality and Social Psychology, 55*(3), 387-395.

Ambady, N., Bernieri, F., & Richeson, J. A. (2000). Towards a history of social behavior: Judgmental accuracy from thin slices of behavior. *Advances in Experimental Psychology, 32*, 201-271.

Aristotle. (1984). *The complete works of Aristotle: The revised oxford translation* (J. Barnes (Ed.) (Vol. 2). Princeton, NJ: Princeton University Press.

Bailenson, J. N., Garland, P., Iyengar, S., & Yee, N. (2006). Transformed facial similarity as a political cue: A preliminary investigation. *Political Psychology, 27*(3), 373-385.

Balabanian, D. M. (1981). Medium vs. tedium: Video depositions come of age. *Litigation, 7*, 25-30.

Ball, G., & Breese, J. (2000). Emotion and personality in a conversational agent. In J. Cassell, J. Sullivan, S. Prevost & E. Churchill (Eds.), *Embodied conversational agents* (pp. 189-219). Cambridge, MA: The MIT Press.

Baumlin, J. S. (1994). Introduction. In J. S. Baumlin & T. F. Baumlin (Eds.), *Ethos: New essays in rhetorical and critical theory* (pp. xi-xxvii). Dallas, TX: Southern Methodist University Press.

Baumlin, J. S., & Baumlin, T. F. (1994). On the psychology of the pisteis: Mapping the terrains of mind and rhetoric. In J. S. Baumlin & T. F. Baumlin (Eds.), *Ethos: New*

essays in rhetorical and critical theory (pp. 91-112). Dallas, TX: Southern Methodist University Press.

Beverly, R. E., & Young, T. J. (1978). *The effect of mediated camera angle on receiver evaluations of source credibility, dominance, attraction and homophily*. Paper presented at the Annual Meeting of the International Communication Association, Chicago, IL.

Bickmore, T., & Cassell, J. (2001). *Relational agents: A model and implementation of building user trust.* Paper presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Seattle, WA.

Boyd, D. (2002). *Faceted id/entity: Managing representation in a digital world.* Unpublished master's thesis, Massachusetts Institute of Technology, Cambridge.

Brahnam, S. (2004). *Modeling the physical personality of the face for ECA self-presentation and interaction with users and the relation of physical personality to emotion.* Paper presented at the AAMAS workshop entitled Embodied Conversational Agents: Balanced Perception and Action in ECAs, New York City.

Brahnam, S. (2005). *Strategies for handling customer abuse of ECAs*. Paper presented at the Interact workshop on Abuse: The Darker Side of Human-Computer Interaction, Rome, Italy.

Brahnam, S. (2006a). *Gendered bods and bot abuse.* Paper presented at the Misuse and Abuse of Interactive Technologies, Montréal, Québec, Canada.

Brahnam, S. (2006b). The impossibility of collaborating with Kathy: 'The stupid bitch'. *M/C Journal, 9*(2).

Brahnam, S. (2008). *Failure to thrive: Value scenarios with in the framework of well-being and flourishing*. Paper presented at the CHI 2008 workshop on Values, Value and Worth: Their Relationship to HCI?, Florence, Italy.

Brooke, R. (1987). Lacan, transference, and writing instruction. *College English, 49*(6), 679-691.

Burgoon, J., Bonito, J. A., Ramirez Jr., A., Dunbar, N. E., Kam, K., & Fischer, J. (2002). Testing the interactivity principle: Effects of mediation, propinquity, and verbal and nonverbal modalities in interpersonal interaction. *Journal of Communication, 52*(3), 657-677.

Campbell, P. (1995). Ethos: Character and ethics in technical writing. *IEEE Transactions on Professional Communication, 38*(3), 132-138.

Caporael, L. R. (1986). Anthropomorphism and mechanomorphism: Two faces of the human machine. *Computers in Human Behavior, 2*, 215-234.

Caporael, L. R., Dawes, R. M., Orbell, J. M., & van de Kragt, A. J. C. (1989). Selfishness examined: Cooperation in the absence of egoistic incentives. *Behavioral and Brain Sciences, 12*, 683-739.

Caporael, L. R., & Heyes, C. M. (1977). Why anthropomorphize? Folk psychology and other stories. In R. W. Mitchell, N. S. Thompson & H. L. Miles (Eds.), *Anthropomorphism, anecdotes, and animals* (pp. 59-73). Albany: State University of New York Press.

Cassell, J., Sullivan, J., Prevost, S., & Churchill, E. (Eds.). (2000). *Embodied conversational agents*. Cambridge, MA: The MIT Press.

Castelfranchi, C. (1998). Modelling social action for ai agents. *Artificial Intelligence, 103*(1-2), 157-182.

Chamberlain, C. (1984). From haunts to character: The meaning of ethos and its relation to ethics. *Helios, 11*, 97-108.

Chapman, D. S., Uggerslev, K. L., & Webster, J. (2003). Applicant reactions to face-to-face and technology-mediated interviews: A field investigation. *Applied Psychology, 88*(5), 944-953.

Corder, J. W. (1989). Hunting for ethos where they say it can't be found. *Rhetoric Review, 7*(2), 299-316.

Davis, H. (1997). Animal cognition versus animal thinking: The antropomorphic error. In R. W. Mitchell, N. S. Thompson & H. L. Miles (Eds.), *Anthropomorphism, anecdotes, and animals* (pp. 335-347). Albany, NY: State University of New York Press.

De Angeli, A., & Brahnam, S. (2006-May). *Sex stereotypes and conversational agents.* Paper presented at the AVI 2006 workshop Gender and Interaction: Real and Virtual Women in a Male World, Venice, Italy.

De Angeli, A., & Brahnam, S. (2008). I hate you: Disinhibition with virtual partners. *Interacting with Computers, 20*(3), 302-310.

De Angeli, A., & Carpenter, R. (2005). *Stupid computer! Abuse and social identity*. Paper presented at the Interact 2005 workshop entitled Abuse: The Dark Side of Human-Computer Interaction, Rome.

de Rosis, F., & Castelfranchi, C. (1999). A contribution to cognitive modeling of affective phenomena. 2002, Retrieved from http://www.ai.mit.edu/ people/jvelas /ebaa99/ derosis-ebaa99.pdf

DeBruine, L. M. (2002). *Facial resemblance enhances trust*. Paper presented at the Proceedings of the Royal Society of London, London.

Derrida, J. (1987). Du tout (A. Bass, Trans.). In *The post card: From Socrates to Freud and beyond*. Chicago: The University of Chicago Press.

Donath, J. S. (1999). Identity and deception in the virtual community. In P. Kollock & M. Smith (Eds.), *Communities in cyberspace* (pp. 27-58). London: Routledge.

Epley, N., & Kruger, J. (2005). When what you type isn't what they read: The perseverance of stereotypes and experiences over e-mail. *Journal of Experimental Social Psychology, 41*, 414-422.

Farrell, T. B. (1993). *Norms of rhetorical culture* New Haven, CT: Yale University Press.

Ferris, S. P., & Montgomery, M. (1996). The new orality: Oral characteristics of computer-mediated communication. *New Jersey Journal of Communication, 4*(1), 55-60.

Fischer, K. (2006). *What computer talk is and isn't: Human–computer conversation as intercultural communication*. Saarbrücken: AQ-Verlag.

Fisher, J. A. (1990). *The myth of anthropomorphism*. San Francisco and Oxford: Westview Press.

Foner, L. (1993). *What's an agent, anyway? A sociological case study* (Agents Memo No. 93-01). Cambridge, MA: MIT Media Laboratory.

Freud, S. (1912). The dynamics of transference. In J. Strachey (Ed.), *The standard edition of the complete psychological works of Sigmund Freud* (Vol. 12). New York: W. W. Norton & Company.

Gilbert, D. T. (1991). How mental systems believe. *American Psychologist, 46*(2).

Gill, C. (1982). The ethos/pathos distinction in rhetorical and literary criticism. *The Classical Quarterly, 34*(1), 149-166.

Goffman, E. (1959). *The presentation of self in everyday life*. New York: Anchor Books.

Guthrie, S. E. (1993). *Faces in the clouds: A new theory of religion*. New York: Oxford University Press.

Havelock, E. (1982). *Preface to Plato (history of the greek mind)*. Cambridge, MA: Belknap Press of Harvard University Press.

Heim, M. (1993). *The metaphysics of virtual reality*. New York: Oxford University Press.

Hemsley, G. D., & Doob, A. N. (1976). The effect of looking behavior on perceptions of a communicator's credibility. *Journal of Applied Social Psychology, 8*(2), 136-142.

Higgins, E. T., & King, G. A. (1981). Accessibility of social constructs: Information processing consequences of individual and contextual variability. In N. Cantor & J.

F. Kihlstrom (Eds.), *Personality, cognition, and social interaction* (pp. 611-621). Hillsdale, NJ: L. Erlbaum Associates.

Hilton, J. L., & Fein, S. (1989). The role of typical diagnosticity in stereotype-based judgments. *Journal of Personality and Social Psychology, 57*, 201-211.

Holland, N. N. (2006). The willing suspension of disbelief: A neuro-psychoanalytic view. *PsyArt: A Hyperlink Journal for the Psychological Study of the Arts, 2006*.

Holloran, M. (1982). Aristotle's concept of ethos, or if not his somebody else's. *Rhetoric Review, 1*(1), 58-63.

Huang, W., Olson, J. S., & Olson, G. M. (2002). *Camera angle affects dominance in video-mediated communication*. Paper presented at the Conference On Human Factors in Computing Systems, Minneapolis, MN.

Hunt, D. (1991). *Riverside guide to writing*. Boston: Heinle & Heinle Publishers.

Isocrates. (2000). Antidosis. In G. Norlin (Ed.), *On the peace. Areopagiticus. Against the sophists. Antidosis. Panathenaicus*. Cambridge, MA: Harvard University Press.

Jung, C. (1959). *Psyche and symbol: A selection from the writings of C. G. Jung*. Garden City, NY: Anchor.

Kopp, S. (2006). *How people talk to a virtual human - conversations from a real-world application.* Paper presented at the Workshop Hansewissenschaftskolleg, Delmenhorst, Germany.

Lang, R. D. (1960). *The divided self: An existential study in sanity and madness*. Middlesex, UK: Penguin Books.

Lanham, R. A. (1976). *The motives of eloquence: Literary rhetoric in the renaissance*. New Haven, CT: Yale University Press.

Lassiter, G. D. (2002). Videotaped interrogations and confessions: A simple change in camera perspective alter verdict in simulated trials. *Journal of Applied Psychology, 87*(5), 858-866.

Leonard, A. (1997). *Bots: The origin of a new species*. New York: Penguin.

Liu, Z., & Cohen, M. (2005). *Head-size equalization for better visual perception of video conferencing*. Paper presented at the IEEE International Conference on Multimedia and Expo, Amsterdam, The Netherlands.

Locke, J. (2009). Staging the virtual courtroom: An argument for standardizing camera angles in Canadian criminal courts. *Masks: The Online Journal of Law and Theatre 1*, 36-58.

Loyall, A. B., & Bates, J. (1997). *Personality-rich believable agents that use language.* Paper presented at the Proceedings of the First International Conference on Autonomous Agents, Marina del Rey, CA.

Mauldin, M. L. (1994). *Chatterbots, tinymuds, and the turing test: Entering the loebner prize competition.* Paper presented at the Proceedings of AAAI '94 Conference, Seattle.

McCain, T. A., & Wakshlag, J. J. (1974). *The effect of camera angle and image size on source credibility and interpersonal attraction*. Paper presented at the Annual Meeting of the International Communication Association New Orleans, LA.

McGinniss, J. (1969). *The Selling of the President*. New York: Trident Press.

Miller, A. B. (1974). Aristotle on habit (ηθος) and character (εθος): Implications for the rhetoric. *Speech Monographs, 74*(4), 309-316.

Miller, C. R. (2002). Writing in a culture of simulation: Ethos online. In P. Coppock (Ed.), *The Semiotics of writing: Transdiciplinary perspective on the technology of writing* (pp. 253-279). Turnhout, Belgium: Brepols Publishers.

Miller, C. R. (2004). Expertise and agency: Transformations of ethos in human-computer interaction. In M. J. Hyde (Ed.), *The ethos of rhetoric* (pp. 197-218). Columbia, SC: University of Southern Carolina Press.

Nathan, L. P., Klasnja, P. V., & Friedman, B. (2007). *Value scenarios: A technique for envisioning systemic effects of new technologies.* Paper presented at the CHI 2007 Works in Progress, San Jose, CA.

Olson, D. R. (1980). On the language and authority of textbooks. *The Journal of Communication, 30*(1), 186-196.

Ong, W. J. (1982). *Orality and literacy: The technologizing of the word*. London: Routledge.

Picard, R. W. (1997). *Affective computing* (1st ed.). Cambridge, MA: The MIT Press.

Plato. (1997). *Plato: Complete works* (J. M. Cooper & D. S. Hutchinson (Eds.)). Indianapolis: Hackett Publishing Company.

Plimpton, G. (Ed.). (1977). *Writer's at work: The 'paris review' interviews* (Vol. 4). London: Secker and Warburg.

Poster, M. (2001). *What's the matter with the internet*. Minneapolis: University of Minnesota Press.

Reeves, B., & Nass, C. I. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. Stanford, CA: CSLI Publications and Cambridge University Press.

Rehm, R. (2008). "She is just stupid"--Analyzing user-agent interactions in emotional game situations. *Interacting with Computers, 20*(3), 326-333.

Reynolds, N. (1993). *Ethos* as location: New sites for understanding discursive authority *Rhetoric Review, 11*(2), 325-338.

Ruzich, C. M. (2008). Our deepest sympathy: An essay on computer crashes, grief, and loss *Interaction Studies, 9*(3), 504–517.

Saarinen. (2001). *Chatterbots crash test dummies of communication.* Unpublished master's thesis, The University of Art and Design, Helsinki, Finland.

Seabrook, J. (1994). My first flame. *The New Yorker, 70*, 70-99.

Searle, J. R. (1992). *The rediscovery of the mind*. Cambridge, MA: The MIT Press.

Sengers, P. (2002). Schizophrenia and narrative in artificial agents. *Leonardo, 35*(4), 427-431.

Sharkey, N., & Sharkey, A. (2007). Artificial intelligence and natural magic *Artificial Intelligence Review, 25*(1-2), 9-19.

Sharp, G. (1996). The autonomous mass killer. *Areana Journal, 6*, 1-7.

Shechtman, N., & Horowitz, L. M. (2003). *Media inequality in conversation: How people behave differently when interacting with computers and people.* Paper presented at the CHI'03, Ft. Lauderdale, FL.

Shneiderman, B. (1987). *Designing the user interface: Strategies for effective human-computer interaction*. Reading, MA: Addison-Wesley.

Shneiderman, B., & Maes, P. (1997). Direct manipulation vs. interface agents. *Interactions, 4*(6), 42-61.

Spada, E. C. (1997). Amorphism. mechanomorphism, and anthropomorphism. In R. W. Mitchell, N. S. Thompson & H. L. Miles (Eds.), *Anthropomorphism, Anecdotes, and Animals* (pp. 37-49). Albany: State University of New York Press.

Suler, J. (2004). The online disinhibition effect. *CyberPaychology and Behavior, 7*(3), 321-326.

Suler, J. (2009). Psychology of cyberspace. Retrieved March 30, 2009, from http://www.enotalone.com/article/2454.html

Tiemens, R. K. (1970). Some relationships of camera angle to communicator credibility. *Journal of Broadcasting, 14*(4), 483-489.

Tseng, S., & Fogg, B. J. (1999). Credibility and computing technology. *Communications of the ACM, 42*(5), 39-44.

Turing, A. (1950). Computing machinery and intelligence. *Mind, 59*(236), 433-460.

Turkle, S. (1997). *Life on the screen*. New York: Touchstone.

Van Gelder, L. (1985). The strange case of the electronic lover. *Ms. Magazine, October*, 98-124.

Veletsianos, G., Scharber, C., & Doering, A. (2008). When sex, drugs, and violence enter the classroom: Conversations between adolescents and a female pedagogical agent. *Interacting with Computers, 20*(3), 292-301.

Vogel, A. (1973). *Body theology: God's presence in man's world*. New York: Harper.

Wallace, P. (1999). *The psychology of the internet*. Cambridge, UK: Cambridge University Press.

Walther, J. B. (1992). *A longitudinal experiment on relational tone in computer-mediated and face to face interaction*. Paper presented at the Hawaii International Conference on Systems Sciences,

Walther, J. B. (1996). Computer-mediated communication: Impersonal, interpersonal, and hyperpersonal interaction. *Communication Research, 23*(1), 3-43.

Walther, J. B. (2007). Selective self-presentation in computer-mediated communication: Hyperpersonal dimensions of technology, language, and cognition. *Computers in Human Behavior, 23*(5), 2538-2557.

Walther, J. B., Anderson, J. F., & Park, D. W. (1994). Interpersonal effects in computer-mediated interaction: A meta-analysis of social and antisocial communication. *Communication Research, 21*, 60-487.

Walther, J. B., & Parks, M. R. (2002). Cues filtered out, cues filtered in: Computer-mediated communication and relationships. In M. L. Knapp & J. A. Daly (Eds.), *Handbook of interpersonal communication* (pp. 529-563). Thousand Oaks, CA: Sage.

Weizenbaum, J. (1976). *Computer power and human reason: From judgment to calculation*. San Francisco: W. H. Freeman and Company.

Welch, K. E. (1990). *The contemporary reception of classical rhetoric: Appropriations of ancient discourse*. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Whittaker, S. (2003). Theories and methods in mediated communication. In A. C. Graesser, M. A. Gernsbacher & S. R. Goldman (Eds.), *Handbook of discourse processes* (pp. 243-286). NJ: Lawrence Erlbaum Associates.

Wisse, J. (1989). *Ethos and pathos from Aristotle to Cicero*. Amsterdam: Adolf M. Hakkert.

Yee, N., Bailenson, J. N., & Rickertsen, K. (2007). *A meta-analysis of the impact of the inclusion and realism of human-like faces on user experiences in interfaces*. Paper

presented at the Conference on Human Factors in Computing Systems, San Jose, CA.

Yoos, G. E. (1979). A revision of the concept of ethical appeal. *Philosophy & Rhetoric, 12*(1), 41-58.

Zdenek, S. (2003). Artificial intelligence as a discursive practice: The case of embodied software agent systems. *AI & Society, 17*(3-4), 340-363.

PsychNology Journal, 2009
Volume 7, Number 1, 49 – 57

Ethical implications of verbal disinhibition with conversational agents

Antonella De Angeli[*♦]

[♦]University of Manchester
(UK)

ABSTRACT

This paper presents a reflection on the ethical implications of conversational agents. The reflection is motivated by recent empirical findings showing that, when interacting in natural language with artificial partners, users tend to indulge in disinhibited behaviour, such as flaming, bullying and sexual harassment. The paper then addresses the question whether conversational agents open any ethical issues and whether this new communication context requires the definition of new moral values and principles or could be addressed by ordinary moral norms.

Keywords: *embodied conversational agents, Internet disinhibition, verbal abuse.*

Paper Received 31/03/2009; received in revised form 29/04/2009; accepted 29/04/2009.

1. Introduction

An on-going debate in computer ethics addresses the uniqueness of the moral dilemmas posed by information technologies (Tavani, 2002). An influential standpoint states that computers generate wholly *new* ethical problems which would have not occurred without technology and for which there are no available analogies in non-computer environment (Maner, 1996). The lack of effective analogies implies that computer ethics requires the definition of new moral values and principles. An alternative standpoint claims that computer ethics transform traditional ethics in complex ways, which anyway can be dealt by within ordinary moral norms (Johnson, 1997). Following this traditionalist approach, three general ethical rules have been proposed for computer-mediated communication in on-line networks (Johnson, 1997).

Cite as:

De Angeli, A. (2009). Ethical implications of verbal disinhibition with conversational agents. *PsychNology Journal, 7*(1), 49 – 57. Retrieved [month] [day], [year], from www.psychnology.org.

[*] Corresponding Author:
Antonella De Angeli
University of Manchester, Manchester Business School, Booth Street West, M
E-mail: antonella.de-angeli@manchester.ac.uk

The first rule claims that on-line communicators must know and abide the norms regulating the forum where the communication takes place. The second rule calls for respecting privacy and property rights of other people. The third one deals more closely with aspects of psychological well-being of the communication partners, stressing the importance of respecting others, not deceiving or harassing them. The similarity between these norms for behaviour on-line and ethical norms for social behaviour in real life is striking.

This paper considers a different context of on-line communication, in which one of the partners is a machine and addresses the question whether this context opens any new ethical issues as compared to human-human interaction. The paper is structured as follows. Section 2 provides a definition of conversational agents alongside an overview of the mainstream research approach dealing with their engineering and an emerging critical rhetorical approach. Section 3 addresses ethical implications of conversational agents and section 4 concludes suggesting ideas for future research.

2. Conversational agents

The term conversational agent describes software which interacts with the user in natural language, via textual input and output or through voice recognition and synthesis. The level of technological complexity varies from pre-scripted chatterbots, which mirror the input of the user through a simple set of transformation rules, to sophisticated multimodal systems, which enrich natural language processing with a number of non-linguistic cues, including hand gestures, facial expressions, and body postures (Abbattista, Catucci, Semeraro, & Zambetta, 2004; Bickmore & Picard, 2005; Cassell, 2000).

Conversational agents are often represented by an anthropomorphic body. A number of embodied conversational agents (ECA's) and talking heads are under development in research centres world-wide and several early prototypes have already entered the Internet. They act as advisors (Berry, Butler & de Rosis, 2005), virtual tutors (Moreno, Klettke, Nibbaragandla, & Graesser, 2002), personal trainers (Bickmore & Picard, 2005) and representatives of major multinational companies (e.g., Ford, Coca-Cola, McDonald and Ikea).

Natural language facilitates anthropomorphic attributions (De Angeli, Gerbino, Cassano, & Petrelli, 2000). Only humans communicate using language and carry on

conversation with one another. Therefore, a talking machine tends to be perceived at a superior level of agency as compared to other machines, and may reach a threshold which subsumes intentionality, sociability, and personality. Anthropomorphic attributions are further strengthened by virtual bodies, which often resemble real humans (Cassell, Bickmore, Campbell, Vilhjalmsson, & Yan, 2001). Overall, conversational interfaces have brought forward an extraordinary change in interaction design: the human metaphor has become the design model (Marakas, Johnson, & Palmer, 2000). ECA's are intentionally designed to be human-like, to show a sense of personality and attitudes, and to involve the user in social relationships.

2.1 Engineering approach

Design and evaluation of ECA's is heavily characterised by a positivistic approach which emphasises desirable consequences of the new interface technology, with little critical analysis of their social implications (Cassell, 2000). The ECA approach found its most influential justification within the field of HCI in the Media Equation Paradigm, a research framework stating that computers are social actors to which users respond as they would do to real people (Reeves & Nass, 1996). This way of thinking deeply influenced the proposition of the Persona Effect claiming that ECA's have the potential for expanding the interaction sphere between humans and computers to encompass social connections and positive emotions leading to more productive task performance (Lester et al., 1997). Despite a recent meta-analysis (Yee, Bailenson & Rickertsen, 2007) has failed to provide empirical support to the Persona Effect, this assumption is still relatively unchallenged in main-stream ECA research which focuses towards the solution of technical issues in the engineering of conversational agents and their evaluation.

2.2 Critical rhetorical approach

A more encompassing research agenda has evolved in the last few years. It stresses the importance of analysing a larger picture of human-agent interaction, one which takes into consideration the full range of human behaviour, including negative outcomes and the possibility of conflicts, disinhibited and antisocial reactions to ECA's (Brahnam & De Angeli, 2008; De Angeli, Brahnam & Wallis, 2005). This critical rhetorical approach shifted the emphasis of ECA research from the engineering of conversational agents to the analysis of their psychological, social and ethical implications.

The approach was motivated by a growing corpus of evidence showing that virtual conversations can at times encourage disinhibited and antisocial behaviour. Verbal abuse and sexual conversations were found to be common elements of anonymous interaction with conversational agents (De Angeli & Brahnam, 2008; Rehm, 2008; Veletsianos, Scharber & Doering, 2008). Independent of the intended focus of the dialogue, interaction setting, and embodiment styles, explicit verbal abuse were common in the user behaviour. An example of verbal abuse is reported in the following *literatim* transcription of a spontaneous interaction between an anonymous web-user and Jabberwacky (www.jabberwacky.com), an on-line chatterbot designed for entertainment.

User: Fuck ypou.
Jabberwacky: I don't think I shall.
User: Fuck you.
Jabberwacky: Don't be superfluously rude.
User: Fuck ur mum.
Jabberwacky: Do you like to eat cake?
User: Off ur mum yeh. [*the interaction continues into a long pornographic talk*]

The user appropriated the chatterbot transforming it into a toy for practicing proposing sex, independently of the fact that the chatterbot did not engage in the sexual conversation and was not explicitly sexually provocative. The conversation closely resembled a context of harassment, which appears to be frequent whenever the agent is represented by an anthropomorphic female body (Brahnam, 2006; De Angeli & Brahnam, 2006; Veletsianos et al., 2008). There is a growing corpus of research indicating that virtual bodies carry with them stereotypical attributions and that user reaction to them is mediated by their physical appearance, such as for example, their gender (Zanbaka, Goolkasian, & Hodges, 2006) ethnicity (Nass, Isbiter & Lee, 2000) and attractiveness (Khan & De Angeli, 2009).

3 Ethical considerations

The occurrence of verbal abuse in human-ECA interaction indicates a need to discuss this topic and explore it more fully and openly. A specific call for action on the ethics of abusing artificial agents, and in particular robots, has been recently put forward (Whitby, 2008). This proposal lays the ground for discussion by proposing

three interdependent ethical issues. First, it raises the question whether it is morally acceptable to treat human-like artefacts in ways that would be considered unacceptable if they would target human beings. Assuming that society considers robot abuse as morally unacceptable, then a new issue is raised as part of the uniqueness debate (Tavani, 2002). It deals with the type of ethical norms which needs to be defined to protect artificial agents, being them unique to this specific context or a direct application of traditional ethics. The final ethical issue is related to design, and considers ways to engineer out the problem of abuse by providing appropriate interaction strategies which constraint its occurrence.

According to Whitby (2008), these questions should be urgently addressed by professional codes of conduct, such as those of the British Computer Society (BCS) and the Association for Computing Machinery (ACM). The argument is justified by the risk that violence towards human-looking artefacts may desensitize the perpetrators, a disputed critique often addressed towards violent video-games (Freier, 2008; Whitby, 2008). The value of an ethical discussion has been strengthened by the application of Christian principles endorsing positive responsibilities, such as “love thy neighbour as thyself” (Dix, 2008). The idea within this perspective is that the ethics of agents can be developed not only by looking at the harm which may come from them, but also at good outcomes, such as the impact of artificial pets on the development of children caring skills. If enhancing positive qualities appeals to our moral sense, then Dix argues that agents that encourage negative behaviours will likewise harm our moral senses. A further elaboration of this way of thinking is that if agents are so anthropomorphic that can be loved by somebody and/or abused by somebody else, their abuse is morally wrong. Thus, designers are encouraged to find new ways to design out unsavoury user behaviours.

A different view-point, claims that there is no pressing need to change current professional codes of conduct because, at the moment, robotics fails to extend the range of social consequences, at least not to the degree that merits special consideration (Thimbleby, 2008). The debate articulates around the appropriateness of applying the concept of abuse to non-sentient agents (Brahnam & De Angeli, 2008). By analysing the issue within the larger context of environmental abuse (those, for example, that lead to global warming) and industrial, personal, and economical abuses of technology (for example, the national financial consequences of not wearing seat belts), the concept of robot abuse is dismissed as sentimentalism. Furthermore, this intellectual position stresses that there are multiple moral layers behind the abuse of

robots. There are some forms of abuse (such as extreme testing in robot war) which have a utilitarian value for engineering development, and as such is morally good.

4 Conclusions

For decades, science fiction writers have envisioned a world in which robots and computers acted like human assistants, virtual companions or artificial slaves. Nowadays, for better or for worse, that world looks closer. As the line between the metaphorical and the literal vanishes, we face uncertainty about how artificial agents will affect our way of interacting with technology, and possibly our social lives. If machines can understand verbal instructions, sense, acquire knowledge, have memory, preferences and personalities many moral and ethical questions are raised. Will they have a sense of self? Who will educate them, guide them, who will they trust? Will there be a time when real and virtual humans are indistinguishable? Who will determine their ethics and morals?

The occurrence of abuse in the interaction with social agents has severe moral, ethical and practical implications. From a moral standpoint, we must reflect on possible effects on individuals, groups, and societies. As we are analysing a dynamic phenomenon which is growing, shaping and constantly changing during the analysis, this reflection must be closely supported by research, which should not only concentrate on the engineering approach but it should move closer to the critical rhetorical approach, proposed in this paper.

The time may not be ready yet for a specific ethics dealing with artificial agents, similar to the growing field of animal ethics or environmental ethics which address moral dilemmas of non-human beings or even inanimate beings. Yet, there is a growing consensus that there is a deontological requirement to initiate a serious reflection within professional bodies, to guide the designers of ECA's in their difficult challenge and possibly to enhance technological development within a value-based approach (Dix, 2008; Freier, 2008; Whitby, 2008). A deontological ethics is of interest to us, if and only if, agent abuse may eventually harm the user. We believe that this risk is intrinsic in the potential of conversational agents to elicit disinhibition and stereotyping.

Stereotypes are widely shared generalisation about members of a specific social groups based on simplified and often derogatory images of out-group members (Fiske

& Taylor, 1991). Stereotypes create a contraposition (us versus them), which may induce and justify anti-social behaviour (e.g., sexism, racism). They are slow and difficult to change, and change requires deeper social and political transformation. We suspect that ECA's may delay this change. Let us take for example sex stereotypes which have been systematically fought by most western nations with positive actions and legal enforcement. ECA's are designed, intentionally or not, with a gender in mind, and more attention is put to the design of attractiveness and realism of female agents. If ECA's encourage gender stereotypes will this impact on real women on-line?

5. References

Abbattista, F., Catucci, G., Semeraro, G., Zambetta, F. (2004). SAMIR: A Smart 3D Assistant on the Web. *PsychNology Journal*, *2*(1), 43-60.

Berry, D. C., Butler, L. T., & de Rosis, F. (2005). Evaluating a realistic agent in an advice-giving task. *International Journal of Human-Computer Studies*, *63*(3), 304-327.

Bickmore, T. W., & Picard, R. W. (2005). Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction*, *12*(2), 293-327.

Brahnam, S. (2006). Gendered bods and bot abuse. In A. De Angeli, S. Brahnam, P. Wallis, A. Dix (Eds.), *Proceedings of the CHI 2006 Workshop: Misuse and abuse of interactive technologies*, (pp. 13-16). Retrieved on October, 3 2009, from agentabuse.org/CHI2006Abuse2.pdf.

Brahnam, S., & De Angeli, A. (2008). Editorial: Special issue on the abuse and misuse of social agents. *Interacting with Computers*, *20*(3), 287-291.

Cassell, J. (2000). Embodied conversational interface agents. *Communications of the ACM*, *43*(4), 70-78.

Cassell, J., Bickmore, T., Campbell, L., Vilhjalmsson, H., & Yan, H. (2001). More than just a pretty face: conversational protocols and the affordances of embodiment. *Knowledge-Based Systems*, *14*(1-2), 55-64.

De Angeli, A., & Brahnam, S. (2006-May). Sex stereotypes and conversational agents. Paper presented at the *AVI 2006 workshop on Gender and interaction: Real and virtual women in a male world.* Retrieved on October 3 2009, from: www.informatics.man.ac.uk/~antonella/gender/papers.htm.

De Angeli, A., & Brahnam, S. (2008). I hate you! Disinhibition with virtual partners. *Interacting with Computers*, *20*(3), 302-310.

De Angeli, A., Brahman, S., & Wallis, P. (2005). *Proceedings of Abuse: The darker side of human-computer interaction*, Workshop at Interact 2005. Retrieved on October 3, 2009 from agentabuse.org/Abuse_Workshop_WS5.pdf.

De Angeli, A., Gerbino, W., Cassano, G., & Petrelli, D. (2000-June). From tools to friends: Where is the borderline? Presented at *Proceedings of the UM'99 Workshop on Attitude, Personality and Emotions in User-Adapted Interaction*.

Dix, A. (2008). Response to "Sometimes it's hard to be a robot: A call for action on the ethics of abusing artificial agents". *Interacting with Computers*, *20*(3), 334-337.

Fiske, S. T., & Taylor, S. (1991). *Social Cognition*. New York: McGraw-Hill.

Freier, N. G. (2008). Children attribute moral standing to a personified agent. *Proceeding of the SIGCHI conference on human factors in computing systems* CHI 2008 (pp 343-352). New York: ACM Press.

Johnson, D. G. (1997). Ethics online. *Communications of the ACM*, *40*(1), 60-65.

Khan, R., & De Angeli, A. (2009-August). The attractiveness stereotype in the evaluation of embodied conversational agents. Presented at *Interact 2009*.

Lester, J.C., Converse, S.A., Kahler, S.E., Barlow, S.T., Stone, B.A., & Bhogal, R.S. (1997). The persona effect: affective impact of animated pedagogical agents. *Proceeding of CHI97: Human factors in computing systems* (pp. 359-366). New York: ACM Press.

Maner, W. (1996). Unique ethical problems in information technology. *Science and Engineering Ethics*, *2*(2), 137-154.

Marakas, G. M., Johnson, R. D., & Palmer, J. W. (2000). A theoretical model of differential social attributions toward computing technology: when the metaphor becomes the model. *International Journal of Human-Computer Studies*, *52*(4), 719-750.

Moreno, K. N., Klettke, B., Nibbaragandla, K., & Graesser, A. C. (2002). Perceived characteristics and pedagogical efficacy of animated conversational agents. In *Proceedings of the 6th International Conference on Intelligent Tutoring Systems*, Lecture Notes in Computer Science 2363, (pp. 963-971), Berlin: Springer Verlag.

Nass, C., Isbister, K., & Lee, E.J. (2000). Truth is beauty: Researching embodied conversational agents. In J. Cassell, Sullivan, J., Prevost, S., Churchill, E. (Eds.), *Embodied Conversational Agents* (pp. 374–402). Cambridge, MA: MIT Press.

Reeves, B., & Nass, C. (1996). *The media equation: how people treat computers, television, and new media like real people and places*. Cambridge: Cambridge University Press.

Rehm, M. (2008). "She is just stupid"--Analyzing user-agent interactions in emotional game situations. *Interacting with Computers*, *20*(3), 311-325.

Tavani, H. T. (2002). The uniqueness debate in computer ethics: What exactly is at issue, and why does it matter? *Ethics and Information Technology*, *4*(1), 37-54.

Thimbleby, H. (2008). Robot ethics? Not yet: A reflection on Whitby's "Sometimes it's hard to be a robot". *Interacting with Computers*, *20*(3), 338-341.

Veletsianos, G., Scharber, C., & Doering, A. (2008). When sex, drugs, and violence enter the classroom: Conversations between adolescents and a female pedagogical agent. *Interacting with Computers*, *20*(3), 292-301.

Whitby, B. (2008). Sometimes it's hard to be a robot: A call for action on the ethics of abusing artificial agents. *Interacting with Computers*, *20*(3), 326-333.

Yee, N., Bailenson, J. N., & Rickertsen, K. (2007). A meta-Analysis of the impact of the inclusion and realism of human-like faces on user experiences in interfaces. In *Proceeding of the SIGCHI conference on human factors in computing systems CHI2007* (pp. 1-10), New York: ACM Press.

Zanbaka, C., Goolkasian, P., & Hodges, L. (2006). Can a virtual cat persuade you?: the role of gender and realism in speaker persuasiveness. In *Proceeding of the SIGCHI conference on human factors in computing systems CHI2006* (pp. 1153-1162), New York: ACM Press.

PsychNology Journal, 2009
Volume 7, Number 1, 59 – 76

Witnessed Presence and the YUTPA Framework

Caroline Nevejan[♦♣] *

[♣]Intelligent Interactive Distributed Systems
Vu University Amsterdam
(the Netherlands)

[♦]Amsterdam School for Communication Research
University of Amsterdam
(the Netherlands)

ABSTRACT

This paper introduces the notion of witnessed presence arguing that the performative act of witnessing presence is fundamental to dynamics of negotiating trust and truth. As the agency of witnessed presence in mediated presence differs from natural presence orchestration between natural and mediated presences is needed. The YUTPA framework, introduced in this paper, depicts 4 dimensions to define witnessed presence: time, place, action and relation. This framework also provides a context for design of trust in products and services, as illustrated for a number of illustrative scenarios.

Keywords: *witnessed presence, time, place, action, relation, YUTPA, responsibility, social structures, performitivity, ethics, emotions, design*

Paper Received 01/03/2009; received in revised form 27/04/2009; accepted 28/04/2009.

1. Introduction

The simple fact of 'being alive' generates a person's natural presence. During the past century information- and communication technology has made it possible to facilitate mediation of human presence beyond the imagination of ever before. People transcend time and place many times in the course of a day, in different roles and different stances. In many situations physical presence is replaced or complemented by one or more types of mediated presence. As millions of people now use such technology every day, social systems for negotiating trust and truth are faced with new dynamics. The ethical implications of these new dynamics demand rigorous analysis of the unprecedented impact on the social structures currently valued. This paper argues that witnessed presence is key to determining trust and truth in natural and mediated environments. Chapter 2 positions the notion of witnessed presence in the context of

Cite as:
Nevejan,C. (2009). Witnessed Presence and the YUTPA framework. *PsychNology Journal, 7*(1), 59 – 76. Retrieved [month] [day], [year], from www.psychnology.org.

* Corresponding Author:
Dr. Caroline Nevejan
Intelligent Interactive Distributed Systems Group, VU University Amsterdam
e-mail: nevejan@xs4all.nl

relevant literature. Chapter 3 introduces witnessed presence as key to the negotiation of trust and truth. Chapter 4 presents the YUTPA framework, being with You in Unity of Time Place and Action, and the four dimensions of witnessed presence with which trust is associated. Chapter 5 illustrates the YUTPA framework as a method for design and Chapter 6 discusses future directions for research.

2. Witnessed Presence: the context

This section discusses three aspects of presence encountered in the literature that are strongly related to the concept of witnessed presence.

2.1 Being here: spaces of observation, agency and performitivity

Presence research over the last 30 years has been mostly concerned with the understanding and creation of human experiences in virtual environments. Tele-presence, and the potential occurrence of social presence and co-presence within virtual environments, focus on the creation and monitoring of the sense of 'being there'. Many detailed contributions to the field (refs) have been made but no agreement on definitions and distinctions has been reached (Lombard & Jones, 2007). From a philosophical perspective Luciano Floridi critiques the current conceptual foundation of tele-presence theory and proposes a new model of presence as 'successful observation' (Floridi, 2005). Floridi argues that tele-presence is used as 'a definition of epistemic failure', which is primarily founded in perception. Even interaction is analyzed as the perception of interaction and not as the interaction itself. Floridi argues that the current tele-presence models do not pay tribute to the complex dynamics between presence and absence, nor does it take the different levels of abstraction and spaces of observation into account: "For surely the doctor tele-operating on a patient is still present, independently of the doctor's perception (or lack thereof) of the technological mediation." (Floridi 2005, p. 660). Floridi argues that local and remote spaces of observation and different levels of analysis define presence.

Multiple experiences of different kinds of presence only become more complex, more hybrid, less linear and more fragmented. In every product or process the dichotomy between human nature and non–human nature can be distinguished and at the same time hybrids are almost immediately accepted in their own right (Latour, 1993). Physical, natural presence, the traditional basis for determining trust and truth in the

context of social activities (Giddens, 1984), is no longer the only determinant. When being in a place, in an on- or offline or mixed environment, ‘action’ generates a connection between “the material and symbolic resources that constitute a place and setting the terms of the agent’s presence” (Spagnolli & Gamberini, 2005, p. 6). However, in these new environments key-concepts of, for example, distance, connection, impact or locality, have been deeply affected by the use of technologies (Virilio & Lotringer, 2008). Tracking and tracing, collecting and distributing, presence and absence have changed the scale and patterns of communication. They have changed how people act and how they relate to each other. Because the time-space configurations of social structures have changed, also the agency of the actor has changed (Giddens, 1984). As a consequence the negotiation of trust and truth has acquired new dynamics, because not only the spaces of observation are more complex, also the agency of the witness is transformed.

Judicial systems in Europe have developed over the last 2000 years and as such they reflect knowledge of social structures that human kind has known so far in this part of the world. In judicial contexts a witness is a crucial figure and courts demand a witness to be sworn in. Having been an observer is not enough; a witness has to take the stand and take responsibility for the report on what has been observed and experienced. The fact that an action that is witnessed becomes a deed upon which can be testified emphasizes the possible impact of the act of witnessing. While witnessing a witness can decide to intervene in the witnessed situation as well. When witnessed, the executing power of the same action has changed for both the one who witnesses as well as of for the one who is being witnessed.

The notion of witnessed presence proposed in this paper emphasizes how presence is performed, can be performed or cannot be performed in the context of a communication process in which multiple types of presence play a role. In addition to understanding the witness as a chosen position in a specific situation, ‘having presence in the world’ can also be understood from the perspective of performitivity (Butler, 1993). In performative acts biological conditions and social identities merge into, for example, the performance of gender or sexuality. When studying presence in on- and offline environments the notion of presence as ‘enacting being’ is informative. Also language can be performative, when words become deeds (Austin, 1962). As most mediated environments are dependent on written code and commands to enable presence in mediated environments, the performative perspective on presence contributes to the understanding of presence as a chosen ‘enactment’ facilitating

certain actions and excluding others.

Luc Steels argues that processes of attribution, synchronization and adaptation define the performance of presence in natural and mediated presences (Steels, 2006). 'Tuning' presence happens in both (Nevejan, 2007). In social structures the understanding of different types of mediated presence is deeply influenced by the development of media schemata. Media schemata, define how mediated presence will be accepted and how they execute power in the social structures in which they function (Ijsselsteijn, 2004). Media schemata, change over time and are different in the variety of (sub) cultures around the globe. The way, for example, television, email or an SMS is understood and accepted, is defined by such media schemata.

The notion of witnessed presence as performance resonates with Floridi's critique on current tele-presence theory. Floridi emphasizes the dynamics between local and remote spaces of observation in which the local space of observation is defined by physical presence that is bound to space and time. The notion of witnessed presence shifts the tele-presence focus from 'being there' to a focus on presence as 'being-here' in relation to many other here's and there's available. It is in the being-here that the perspective on agency and performitivity of presence is to be found as argued in the following paragraph.

2.2 Conatus: depth in relation, data-identities and moral distance

Riva, Waterworth & Waterworth argue from a bio-cultural approach, that presence manifests in the strive for well-being and survival (Riva, Waterworth & Waterworth, 2004). From this perspective the notion of witnessed presence can be considered to have agency and performitivity as well. The witness chooses to take the stand, the sense of presence makes her or him be aware and act. The perception and awareness of 'something is happening' has impact in natural presence because the conatus, first introduced by Spinoza as the quest for well-being and survival, operates on all levels of the organism of the human being, who is trying to regulate constantly towards homeostasis (Damasio, 2004). From a neurological perspective Antonio Damasio states that the brain constantly distinguishes between what is beneficial for life and what is detrimental to life. Damasio argues that in the perception of something happening emotions and feelings are crucial indicators of where well–being and survival are to be found (Damasio, 2000). People steer away from pain, trying to restore the homeostasis. People steer away from unhappiness, trying to make things better. The 'conatus' triggers a human being to take care of him or herself, and it also

triggers the human being to take care of ‘other selves’ to keep the environment healthy and safe. Social conventions and ethical rules may be seen as extensions of the basic homeostatic arrangements at the level of society and culture. An individual’s drive for survival can also be considered to be the fundament for ethical behaviour towards others (Damasio, 2004).

Mediated presences contribute to daily lives, knowledge and experience significantly. However, the natural presence of the actors involved remains to be distinct because natural presence has to physically survive with or without the use of technology. From this perspective it seems reasonable to argue that mediated presences should only have impact as far as that they do not harm nor confuse the sense of natural presence that helps human beings to steer away from pain towards well-being and survival.

When ‘enacting being’ the depth in relation between human beings sets the context for how communication is understood (Nevejan, 2007). Strangers, people with whom a human being has no relation, are merely perceived as information (Buber, 1937). This resonates with the experience that in the midst of all the data streams that human kind produces today, it seems that to be able to hear the voices of suffering has become more problematic than ever (Baxi, 1999). To be able to hear a voice of suffering requires the capacity to have complex feelings like compassion and solidarity which do not evolve from the perception of information only. To develop these feelings human beings have to be part of social structures and engaged in human relationships over time (Damasio, 2004).

Because mediated presences offer limited sensorial input, limited mediation of context, and limited possibilities to act, a moral distance is easily adopted towards people a human being does not know (Hamelink, 2000). Current technology facilitates not only a mediation of presence, they also collect, match, duplicate, distribute and produce ‘data-identities’ (Nevejan, 2007). Human beings have little control over their ‘data-identity’ in current technological systems while the data-identity of a human being has acquired great agency in the social structures in which human beings live. There is little control on how data are created, there is hardly any control on how data are matched, travel or even on how long they exist. One can argue that the systems themselves have become participants in communities and are executing their own specific ways of witnessed presence (Brazier & van der Veer, 2009). The confrontation between a human being and his/her data-identity and the effect of being witnessed by technological systems, which imperceptibly invade the personal sphere all the time, has hardly been studied. However, having agency is a requirement for being a witness

and to participate in the negotiation of trust and truth. Because human beings have so little influence over their data-identities in the social structures upon which they depend, they adopt a moral distance towards the own self as well (Nevejan, 2007). One of the possible implications of adopting a moral distance towards one's self is that feelings and emotions will not evolve as they should, which leads to the ultimate consequence that a human being is less capable of steering towards one's own well-being and survival. Also the sense for a safe social environment diminishes because as a result of the moral distance to the self, also the moral distance to other human beings increases.

Although related to mediated presence, concepts like homeostasis and conatus are different: there is a different sense of causality, limited sensorial input, local and implicit knowledge can hardly be mediated and the connection most often provides context. Context as reference, that a place with an embedded culture offers, has disappeared (Nevejan, 2007). Also, consistency in identity, through actions and feedback to these actions, requires special attention when being involved in mediated presence. The way emotions and feelings are triggered in mediated presence, and the process of attribution, synchronization and adaptation happen, can be significantly different from a natural presence context. When being a witness in mediated environments the steering capacity of emotions and feelings towards well-being and survival has to be understood and analyzed in different ways. The agency of witnessed presence is different in natural presence from the agency a witness has in mediated environments.

2.3 Collaboration: spatiotemporal movements, incommensurability and collective authored outcomes

Higher trust makes collaborations more smooth and effective and therefore also more cost-effective as Karen Armstrong claims (Kleiner, 2002). To create a 'trusted' sense of place in only mediated environments is a challenge, which is why 'being a witness' and creating a 'to be witnessed presence' in mediated environments requires attention.

In social networking sites, like Facebook and Linkedin, the purpose is to connect people to other human beings and therefore these sites facilitate a witnessing and being witnessed around the clock and from all over the globe. The popularity of these sites proves that new configurations are being invented to connect natural and mediated presence to create a trusted sense of place in which people can witness each other, possibly testify and possibly act upon what they witness. The context these mediated environments offer (in addition to the platform they provide), appears to be

the ‘being in relation with other human beings’ it self. It appears that people trust what they perceive on these sites for 100% (ten Kate, 2009). The ‘neutrality’ of technology generates a great sense of trustworthiness even though most users are not even aware of license agreements to which they have agreed. People argue that the information about others is also to be trusted because all information links to real life situations, networks, cultures and people. Any untruth would surface easily because of this (ten Kate, 2009).

In professional realms, be it in geographically distributed teams of collaboration or not, technologies play a crucial role in the work processes and new configurations between on- and offline work are being invented (Vasileiadou, 2009). As a result, how and when to meet in real life, in natural presence, has become a choice. In collaborations a significant hurdle to over come between the participants involved is incommensurability, the fundamental not sharing of an understanding. Thomas Kuhn has been studying this phenomenon extensively. To be able to interact, Kuhn argues, members of the community have to share certain concepts or no interaction is possible (Kuhn, 2000). Collaborating actors share terrains of commensurability and also terrains of incommensurability, otherwise they can not collaborate. Witnessing the presence of others informs about the identity of others and these identities are, among other things, formed by conceptual schemes as well as by the spatiotemporal trajectories that are identified (Kuhn, 2000). To be able to recognize spatiotemporal trajectories of other participants is a requirement ‘tuning’ participant’s presence’s, which is necessary for tackling incommensurability and being able to interact. However, identifying spatiotemporal trajectories in mediated presence is very different from identifying spatiotemporal trajectories in natural presence. To mediate nuances of spatiotemporal trajectories of enacted beings is difficult and may even be impossible. Just as the sense for well-being and survival is difficult to mediate since it is highly context dependent and context can hardly be mediated at all (Nevejan, 2007). Therefore the conclusion can be drawn that when issues of ethical nature are at stake, when questions are asked about what is good to do and what is beneficial for life, people have to meet in natural presence. Only in natural presence the shared sense of what is good for well-being and survival can be ‘collectively authored’ in such a way that all stakeholders will base their future acts on the ‘collectively authored outcomes’ that have been agreed upon (Humphries & Jones, 2006).

3. Witnessed presence is key in negotiating trust and truth

Both trust and truth are not given entities but processes of negotiation. Trust builds and breaks down, truth changes according to perspective. Also both processes are dependent on human perception and interaction for which reason they are subjected to complex dynamics in which psychological, sociological, theological, biological, political and economic realities play a role. Nevertheless according to the literature discussed in chapter 2, when discussing presence technologies key dimensions underlying these dynamics can be identified.

Being a witness traditionally meant that a human being was present at a specific time and a specific place. From a judicial perspective being an observer is not enough; to be a witness a human being consciously decides to take responsibility for the report on situation that is witnessed. As a result the report on this act of witnessing is supposed to contribute to the truth. This dynamic of being a witness and taking responsibility for being a witness, can be identified in many realms of society to create trusted and truthful interactions. In commercial contracts, when the stakes are high as in buying shares or a house for example, stakeholders have to be present in front of a notary to sign a contract specifying the date and precise time. In organizational agreements and civic procedures like marriage, the witness is a returning figure. Witnessing is formally orchestrated in these processes to guarantee truthful and trusted interactions and transactions. In informal social environments witnessing, or lying about having witnessed something as in gossip and rumours may happen, is a well known dynamic to create certain (mis)conceptions of other people or events. When discussing ethics of presence technologies, witnessed presence as a notion that plays a role in the negotiation of trust and truth, is useful.

The following three sections discuss four variables related to the concept of witnessed presence: space, time, relation and action.

3.1 Space and Time

The structure of communication between people is not only defined by the sharing of place and time, but also by the capacity to recognize other beings spatiotemporal trajectories. Being a witness to other people's presence starts before the moment of interaction. It is pre-linguistic in that sense. The perception of other human beings movements influences how a witness performs his or her own presence as a consequence. The configuration of space and time defines the space of observation,

and defines how the witness' presence is performed as well. In natural presence this process of 'staging' presence in relation to the witnesses around, is very different from staging presence in mediated environments.

The perception and experience of space and time have been part of human existence. In arts as well as in the sciences human kind has been struggling to understand and express these fundamental dimensions of life. The current presence technologies challenge this understanding and experience in unprecedented ways. When focusing on witnessed presence in the context of presence technologies all of the questions about space and time that have ever been asked seem relevant. When trying to understand what happens in a specific situation, when being a witness, those questions have to be asked again because an apparent simple transcending of time and or place actually deeply transforms the concepts that human beings recognize and therefore the way presence is performed as well.

3.2 The possibility to act

In addition to space and time, also the possibility to act influences how presence is performed. In mixed on- and offline environments the possibility to act helps to bridge the different worlds. If there is no possibility to act and a human being is nevertheless witness to enrolling events, people easily adopt a moral distance and doing so detach themselves from the sense for well-being and survival. Especially in mediated environments where data-identities interact, such a moral distance can even be taken to the own self. Witnessing is an act in which a human being takes responsibility for the act of being witness. If this responsibility is denied because of a lack of possibilities to act, often there seems to be no other option than to detach. Vice versa, a witness who decides to act, and words can be a deed in this sense, breaks the moral distance and becomes an actor in his or her own right.

To be able to act as a witness, having the potential to become an actor, a person needs a sense of what will be good and what will be bad, in order to anticipate an intended effect of one's action. In on- and offline places where culture is shared, the witness can be aware of the morality around him/her and will know what is good and what is bad for well-being and survival. When a witness does not know the morality of the context in which one witnesses, the witness will be hesitant to use the capacity to act upon what is witnessed because there is no sense of social safety around.

To be a witness and to be part of the negotiation of trust and truth, human beings need the possibility to act as well as an understanding of the possible impact of the act

they may or may not do.

3.3 Depth of Relation

The depth of relation between human beings is the fourth variable that defines how witnessing takes place. Witnessing a loved person in on- or offline environments is very different from witnessing a stranger. This relation provides a very strong context in natural as well as in mediated presence. In social relations human beings develop a whole range of psychological states, from simple emotions of like and dislike, to love and hate and more complex feelings like compassion or solidarity. To be a witness to suffering or being witnessed when suffering demands performance of presence and social structures that support. Also passion, joy and success need performance of presence and social structures that support. When focusing on ethics in presence technologies the question that rises is how complex feelings and emotions like compassion, empathy, shame, guilt and others, evolve in mediated presence over time and affect the social structures in which human beings live and survive. Because of the large-scale use of presence technologies, the range and depth of human relationships are undergoing significant change. People can be ‘in touch’ with loved ones thousands of miles away and strangers can become intimate friends even though one has never met in real life before or even intends to do so.

Processes of attribution, synchronization and adaptation have more impact than ever because current presence technologies can only facilitate partial channels of communication and transactions. Because mediated presence is dependent on these processes of attribution, solitary human beings are easily confused about what they perceive. The social structures, in which the mediated presences of other people are perceived, are crucial in the understanding of the trustworthiness and truthfulness of the presences that are witnessed. Even in large social networks the connection to the natural presence of human beings involved, is necessary to create trusted and truthful environments. Also in collaborations it appears to be necessary to meet in natural presence when issues of ethical nature are at stake. In natural presence the ultimate sense of what is good for survival and well-being is strongest and the identification of other human beings and the concepts that are shared, is clearest.

Therefore ethics of presence technologies have to be founded in the natural presence of human beings involved. To better understand the social structures in which witnessed presence operates, the YUTPA framework was developed (Nevejan, 2007).

4. YUTPA framework

The specific configuration of time, space, action and relation in a certain product or process, in which natural presence, mediated presence and witnessed presence all play a role, enables certain forms of trust and truth to be established while excluding others. Because of the development of mechanical, electrical, electronic and digital technologies, people can act with other people over time and distances in other ways than those that are dictated by physical presence. It is in the specific 'time and space configuration in which one meets with others in action' that one set of possibilities and liabilities can be distinguished from another. Such a configuration is called a YUTPA configuration.

YUTPA is an acronym for "being with You in Unity of Time, Place and Action". Time, place, action and relation are dimensions that can have different values between You and not–You, Now and not–Now, Here and not–Here, Do and not–Do as depicted in Figure 1.

The You/not-You dimension refers to the relationship with the other human being(s) with whom one interacts.

The Now/not-Now dimension refers to the sharing of the experience of time, synchronous or asynchronous in past or future.

The Here/not-Here dimension encompasses the sharing of place or not. Depending on how place is defined or experienced this can be geographically small or large, it can also refer to the sense of distance in virtual and online worlds.

The Do/not-Do dimension refers to the possibility to act as part of or as a result of a social interaction.

The word Unity refers to the specific configuration between these four dimensions that is designed in a certain product or process, which makes certain interactions possible while it excludes others. It is a formulation from the perspective of the actor, from the perspective of the person involved. In specific configurations human beings enact their being, witness each other, tune and perform their presences.

In every specific YUTPA configuration different possibilities to delegate trust and to produce and verify particular facts is given. Internet, mobile communication, GIS, and databases have created new YUTPA configurations of communication.

The position this paper takes is that values for 'presence-ethics' need to be developed in the relation to the natural presence of the people involved. All contributions, possible destruction, confusion and transformations of other YUTPA

configurations have to be valued and judged from the perspective of the natural presence of human beings, and the environment they need, to be well and survive (Nevejan, 2007). In this respect it is interesting to notice that most current information and communication technology agenda's for innovation of truthful and trustworthy environments can be located in the black space of figure 1 in which there is no possibility to act for human beings to be involved. While most human beings love, have children, enjoy life and find trust and truth in the white 'action' space of the same figure 1.

Figure 1. The 4 dimensions of time, place, relation and action define how the relation between witnessed presence and the negotiation of trust and truth can be understood. Next to the three axes, the dimension of Action is represented by the black and white parts of the sphere illustrating the possibility to act in the white of 'clear air' or the lack of possibilities to act in the black of 'no oxygen space'. (Graph: Max Bruinsma)

The four action spaces defined by You create a solid ground for social interaction because these interactions are understood in the context of the relation with the other human being. Establishing distrust is as trustworthy in this respect as establishing trust. Feedback from synchronous and asynchronous mediated presences (You/not–Now/not–Here, You/Now/not–Here) may contribute to the building or diminishing of trust provided the context of a relationship supports this process. With strangers

especially synchronous communication, as is facilitated by the telephone for example, is perceived as truthful and generates trust.

The four action spaces defined by not–You are more complex and highly dependent upon the delegation of trust. Trust in social structures and trust in technology are required to be able to operate in those spaces, trust between individuals is not the issue here. When sharing time and place, while not knowing other people who are present as in a busy street for example, people treat each other as information. But in such a busy street one can still be a witness and decide to act. In all other three not–You spaces technology is needed for human beings to be present; a presence that manifests itself mostly as data-identity, formatted by technology, which is often outside of the 'original' human beings control. In the not–You communication spaces basic trust is delegated to governments and companies to create and maintain systems in trustworthy and truthful manners yet these are not always capable or willing to do so.

The blurring between You and not–You creates confusion as well as solutions. In not–You spaces trust is delegated, moral distance is easily taken, responsibility is harder to sense but the 'neutrality' of technology generates a great sense of truth and trustworthiness. Therefore in communication processes, which consist of series of interactions and transactions as well, the orchestration of links between on- and offline moments is crucial for success. Part of the trustworthiness of online banking for example is the fact that there is also a bank in a building, with people with whom one can communicate. Part of the trustworthiness of online banking is also the fact that the 'real' bank is subjected to the rule of law.

Between the experience of a human being and the social structures to which trust and truth finding are delegated, specific products or services are accepted or not. By interchanging between mediated and natural, between witnessed and not, between synchronous and asynchronous, between not–You and You spaces, between Here and not–Here, and by offering the possibility to act, communication processes take shape and trust and truth are build up or broken down.

5. Applying the YUTPA framework as a method for Design

Over the last two years the YUTPA framework has been used as a method for design in a variety of situations with a variety of people: marketing managers, business people, representatives of larger organizations, government bodies and students of

media and design. Trust and truth are context dependent and so is any intervention by design. In this sense the YUTPA method can only function in processes of situational design (Schwarz, 2006). In this section a short impression is given of how the YUTPA framework is used as a method for design.

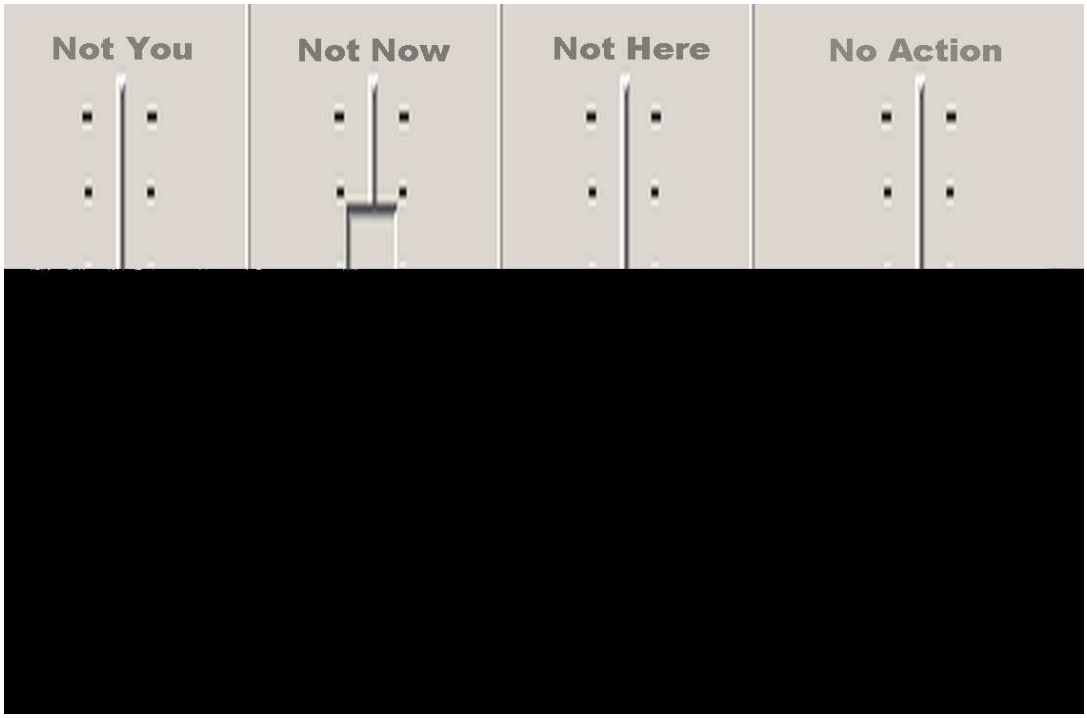

Figure 2. Using the image of a sound mixer to tune 4 dimensions of witnessed presence into one configuration to enhance trust. (Graph: Mike de Kreek)

Design of a process with the YUTPA framework involves (1) Analysis of the 4 dimensions with respect to the requirements. (2) Having analyzed the design problem the four dimensions are tuned to different values to explore the impact on the negotiation of trust and truth. (3) When the preferred configuration has been found, the new products or services can be further configured and designed.

Especially in business environments, communication processes are costly. The configuration of not-You, not-Here, not-Now and not-Do (when all the slides are up in figure 2) would to be very cost effective using technological systems, which can run by themselves and incur no personnel costs. However, if clients loose trust in the service because it is too impersonal and hard to control, people will refrain from using it. So a balance between communication costs and ‘trust-investment’ needs to be found.

6. Discussion

This paper proposes a conceptual framework that facilitates a better understanding of the ethical implications of presence design. Where Floridi proposes to focus on the spaces of observation and levels of abstraction instead of focusing on the perception and experience of the subject as most tele-presence research does, the notion of witnessed presence proposed in this paper takes the sociological perspective of 'witnssed presence as agency' in mediated and non-mediated environments to sustain well-being and survival for the individual human being and between human beings as well. It takes the perspective of the individual human being, who, while being present 'here' is also present in several 'there's' while interacting and observing others. Having witnessed presence, enacting being in natural and mediated environments, is considered to be an act of performitivity in which biological and social realites merge. To be a witness is an act with distinct consequences, especially with respect to the ethical implications of presence design.

Further research will explore each of the four dimensions in relation to each other. Earlier tele-presence research will be revisited, as will the perspective of the Universal Declaration of Human Rights to use the capability approach to translate these values in practical and measurable terms for presence design (Nussbaum, 1999). The challenge is how to understand, create and integrate witnessed presence in social structures of system and service designs, so human beings can take full responsibility for their actions and safeguard human dignity for generations to come.

Current research focuses on the implications of witnessed presence for the design of autonomous systems, systems that participate in communities in their own right (Nevejan, 2009). With social scientists, artists and designers the concept of witnessed presence is further explored.

7. Acknowledgement.

I thank professor Cees Hamelink and professor Sally Wyatt for supervising the dissertation upon which most of the here presented work is based. I thank professor Frances Brazier for supervising me in writing this paper, the two anonymous reviewers for their valuable suggestions and Anna Spagnolli for her encouragement. This paper is largely based upon the dissertation "Presence and the Design of `Trust" (Nevejan

2007) and the presentation of this work on the Presence Conferences in Barcelona in 2007 and in Padua in 2008. The dissertation and current further research can be found at http://www.being-here.net.

8. References

Austin, J.L. (1962). *How to do things with words*. Cambridge MA: Harvard University Press

Baxi, U. (1999). Voices of Suffering, Fragmented Universality, and the Future of Human Rights. In B.H. Weston & S.P. Marks (eds.) *The Future of International Human Rights* (pp. 101–156). New York: Transnational Publishers.

Brazier, F.M.T. & van der Veer, G. (2009- February). Interactive Distributed and Networked Autonomous Systems: delegation or participation. Paper presented at the workshop *Human Interaction with Intelligent & Networked Systems*. Retrieved on April 15, 2009 from http://www.iids.org

Buber, M. (1937). *I and Thou*. Edinburgh, UK: T&T Clark.

Butler, J. (1993). *Bodies that Matter, on discursive limits of "sex".* New York: Routledge.

Damasio, A. (2000). *The Feeling of What Happens. Body, Emotion and the Making of Consciousness.* London: Vintage, Random House.

Damasio, A. (2004). *Looking for Spinoza, Joy, Sorrow and the Feeling Brain*. London: Vintage, Random House.

Floridi L. (2005). The Philosophy of Presence: From Epistemic Failure to Successful Observation. *Presence: Teleoperators & Virtual Environments*, *14*(6): 656-667.

Giddens, A. (1984). *The constitution of Society, Outline of the Theory of Structuration*. Cambridge: Polity Press.

Hamelink, C.J. (2000). *The ethics of Cyberspace*. Thousand Oaks: Sage Publications.

Humphries, P. & Jones G. (2006). The Evolution Of Group Decision Support Systems To Enable Collaborative Authoring Of Outcomes. *World Futures: The Journal of General Evolution, 62*: 171–192.

IJsselsteijn, W. A., (2004). *Presence in Depth*, PhD diss., Technische Universiteit Eindhoven.

ten Kate, S. (2009). *Trustworthiness within Social Networking Sites: A study on the*

intersection of HCI and Sociology. Master thesis Business Studies, University of Amsterdam

Kleiner, A. (2002). Karen Stephenson's Quantum Theory of Trust. In *Strategy + Business*, Fourth Quarter. Retrieved on November, 24th 2006 from: http://www.strategy–business.com.

Kuhn, T. S. (2000). *The road since structure, philosophical essays, 1970–1993, with an autobiographical interview*. Chicago: The University of Chicago Press.

Latour, B. (1993). *We have never been Modern.* Cambridge: Harvard University Press.

Lombard, M., & Jones M. T. (2007). Identifying the (Tele)Presence Literature. *PsychNology Journal*, *5*(2), 197 – 206.

Nevejan, C. (2007). *Presence and the Design of Trust*. PhD diss., University of Amsterdam. Retrieved on April, 15 2009 from http://www.being-here.net

Nevejan, C. (2008-October). YUTPA: A Methodology for an Ethical Approach to Presence Design. Presented at *the First European workshop on The Ethics of Presence and Social Presence Technologies.*

Nevejan, C. (2009- February). Spatio-temporal movements in communities of practice, in which human beings and autonomous systems participate. Paper presented at the workshop *Human Interaction with Intelligent & Networked*. Retrieved on April, 15, 2009 from http://www.iids.org

Nussbaum, M.C. (1999). Capabilities, Human Rights, and the Universal Declaration. In *In B.H. Weston & S.P. Marks (eds.) The Future of International Human Rights* (pp. 25–64). New York: Transnational Publishers.

Riva, G., Waterworth J.A. & Waterworth E.L. (2004). The Layers of Presence: A Bio–cultural Approach to Understanding Presence in Natural and Mediated Environments. *CyberPsychology & Behavior*, *7,(*4): 402–416.

Spagnolli A., & Gamberini L. (2005). A Place for Presence. Understanding the Human Involvement in Mediated Interactive Environments. *PsychNology Journal*, *3*(1), 6 – 15. Retrieved on April 15, 2009, from www.psychnology.org.

Steels, L. (2006). Experiments on the emergence of human communication. *Trends in Cognitive Sciences, 8*(10): 347–349. Retrieved on July 7, 2006, from http://arti.vub.ac.be/~steels/.

Schwarz, M. (2006). Institutioneel Ontwerpen her–zien. In *Institutioneel Ontwerp: Relict, revival of revisie.* Den Haag*:* Ministerie van VROM, atelier Rijksbouwmeester.

Vasileiadou, E. (2009). Working Apart Together, using ICT's in research collaboration,

PhD diss., University of Amsterdam.

Virilio, P. & Lotringer, S. (2008). *Pure War, 25 years later*. Los Angeles: Semiotext(e).

PsychNology Journal, 2009
Volume 7, Number 1, 77 – 100

Cybertherapy: Advantages, Limitations, and Ethical Issues

Cristina Botella[♦♣*], Azucena Garcia-Palacios[♦♣], Rosa M. Baños[♥♣] and Soledad Quero[♦♣]

[♦]Universitat Jaume I (Spain)

[♣]CIBER de Fisiopatología de la Obesidad y Nutrición (CIBEROBN) (Spain)

[♥]Universidad de Valencia (Spain)

ABSTRACT

Information and Communication Technologies (ICT) are becoming more and more common in Clinical Psychology. Two of the technologies that are more consolidated in this field are virtual reality (VR) and telepsychology. There are other technological innovations that are beginning to be used in clinical and health psychology such as ambient intelligence, ubiquitous computing or persuasive computing. In the last fifteen years there has been a proliferation of studies testing the efficacy of immersive virtual reality in the delivery of cognitive behavioral therapy (CBT) for several mental disorders and health conditions. The essence of VR is that it can simulate reality and add a new possibility: the user has the illusion of "being" in the computer-generated environment while interacting with the VR objects. This unique feature of VR is very relevant for its use in Clinical Psychology. At the same time, it can raise several ethical issues. It is important to investigate the possible effects of blurring the distinction between real and virtual worlds in vulnerable populations. Some other concerns regarding the use of VR in therapy have already been investigated, such as cybersickness. After ten years of experience treating patients with VR, this has not been a problem in the published efficacy studies.

Telepsychology has also been used to improve the delivery of CBT. A number of Internet-delivered programs have already become important tools in the health system. The main advantage is that online therapy can reach people who might not otherwise seek therapy, such as disabled people or those who live in remote areas. However, several concerns have been raised about self-help procedures, like the issue of self-diagnosis. and the fact that patients usually have all of the necessary self-help information at their disposal. It is important to establish criteria to protect people from the possible negative effects of this.

Other innovations such as ambient intelligence and pervasive computing bring up other ethical issues. For example, is privacy being compromised too much when people are located using GPS or physiologically monitored 24 hours a day? Criteria for considering these issues must be established.

Our research group has been working with new technologies and therapies for the last fifteen years. This paper addresses the ethical issues we have encountered in our research and clinical practice; it also explores ethical issues that will become increasingly important.

Keywords: *cybertherapy, virtual reality, telepsychology, e-therapy, pervasive computing, ethics*

Paper Received 05/04/2009; received in revised form 10/04/2009; accepted 10/04/2009.

Cite as:

Botella, C., Garcia-Palacios, A., Baños R.M., & Quero, S. (2009). Cybertherapy: Advantages, Limitations, and Ethical Issues. *PsychNology Journal, 7*(1), 77 – 100. Retrieved [month] [day], [year], from www.psychnology.org.

* Corresponding Author
Cristina Botella
Dpt. Psicologia Basica, Clinica y Psicobiologia, Universitat Jaume I., Avda Vicent Sos Baynat s/n., 12071 Castellon, Spain, Phone: +34 964 729881, Fax: +34 964 729267
botella@psb.uji.es

1. Advances in psychological treatments

The field of psychological treatment grew tremendously during the twentieth century. The pioneering work by J.B. Watson and his assistant Rosalie Rayner demonstrated, contrary to the dominant Freudian theories of psychology, that it was possible to stimulate phobias in a laboratory environment (Watson & Rayner, 1920). The *Little Albert* experiment provided empirical evidence of classical conditioning in humans. A few years later, Mary Cover Jones conducted her study of a patient named Peter (Jones, 1924). She treated his fear of a white rabbit with a variety of fear-reducing procedures, the most successful of which employed *direct conditioning*: a pleasant stimulus (food) was presented simultaneously with the rabbit. This case illustrated how fear may be eliminated under laboratory conditions. Recognizing her success, Joseph Wolpe christened Jones "the mother of behavior therapy".

Wolpe continued this path of systematic experimentation. He introduced evidence-based psychological procedures to the field of psychological treatment through the application of "systematic desensitization". This was the basis of a psychological technique that would enjoy success and empirical support: exposure therapy. Since then, numerous studies have been conducted demonstrating the efficacy of psychological procedures for the treatment of several mental disorders. Wolpe, Kelly, Lazarus, Skinner, Rachman, Marks, Beck, Seligman, Mahoney, Barlow, Salkovskis, Clark, etc. are important figures who contributed to the consolidation of Cognitive-Behavior Therapy (CBT).

> *Cognitive-Behavior Therapy is devoted to the relief of human suffering using methods that work. The latest in scientific advances are used to design treatments. Cognitive and Behavioral Therapies (CBT) use techniques that are based on scientific evidence to understand and treat psychological symptoms* (Association for Behavioral and Cognitive Therapies).

CBT comprises a significant number of well-established therapeutic techniques and programs for the treatment of several human mental disorders: exposure, cognitive restructuring, relaxation, modeling, and others. One of the main features of CBT is its emphasis on empirical study, namely the study of mental health problems in the lab wherein one can control variables to assure internal validity while also considering ethical issues concerning the study's participants.

CBT's emphasis on experimentation and empirical studies allows us to understand another important advance in the field which happened at the end of the twentieth century. In 1993, division 12 of the *American Psychological Association* (Clinical Psychology Division) created a panel of Experts (*Task Force*) with the aim of promoting the application of empirically validated treatments as well as the development and dissemination of evidence-supported psychological procedures. Since then, there has been an increase in the development of evidence-based treatments, most of them being CBT programs (Chambless et al., 1996, Chambless & Hollon, 1998). Additionally, a clinical guide (*Template for Developing Guidelines: Interventions for Mental Disorders and Psychological Aspects for Physical Disorders*) also by the American Psychological Association (*APA Task Force on Psychological Intervention Guideline,* 1995) established a distinction between the efficacy of an intervention (*efficacy*) and its clinical utility (*effectiveness)*. The guideline proposed evaluating interventions using two axes: Axis 1, efficacy or internal validity, is a rigorous analysis of the empirical evidence in order to investigate the efficacy of the intervention. Axis 2, effectiveness or clinical utility, is an analysis of the feasibility of the intervention in its natural context; it includes therapist expertise, availability of trained therapists, acceptability of the treatment by patients, and the possibility of applying the intervention in natural contexts.

From this perspective, ethical considerations have to be made. First, it is important to build on the work of Watson, Jones and Wolpe to develop efficacious treatments, or to progress in axis 1 (internal validity). Additionally, we must progress in axis 2 (effectiveness or clinical utility). For example, if we have two procedures equally efficacious (axis 1), which one would we use to treat a patient? We would have to rely on axis 2 criteria. For example, we could choose the one with fewer negative side effects, more feasible application, more patient acceptance, less attrition, lower financial costs, or with lower cost and effort required for training therapists. Information and Communication Technologies (ICTs) can have an important role in improving axis 2 aspects in the field of psychological treatments.

2. Efficacious psychological treatments and Information and Communication Technologies (ICTs)

In recent years, there have been spectacular advances in ICTs that could help improve psychological interventions, mainly regarding the effectiveness or clinical utility axis. Therapeutic applications of ICTs include e-therapy, virtual therapy, VR treatments, and others, all of which fall under the umbrella of "cybertherapy". This term gives its name to one of the more prestigious conferences in the field of computer-aided Psychotherapy, another concept coined by a prestigious figure in CBT, Isaac Marks (Marks, Cavanagh & Gega, 2007). Cybertherapy involves using the computer to provide, enhance or facilitate therapy. As a therapeutic tool, the computer can be both a communication device that enables and promotes distance interaction, as well as a simulation device for creating virtual realities. From our point of view, cybertherapy includes the use of any new device based on ITCs that could contribute to improvements in clinical psychology (in addition to computers).

2.1. Virtual Reality and Augmented Reality

ICTs' first contribution to Clinical Psychology was in Immersive Virtual Reality (VR). The pioneering work by Barbara Rothbaum and colleagues (Rothbaum, et al., 1995), explored the utility of virtual reality for the delivery of exposure therapy for acrophobia. Since then, a significant number of virtual reality programs have demonstrated its efficacy in the treatment of several mental disorders, mainly in anxiety disorders. Ten years ago, in their work *"Basic issues in the use of virtual environments for mental health applications",* Rizzo, Wiederhold and Buckwalter (1998) highlighted that "After an early period of inflated expectations and limited delivery, Virtual Reality Technology is now beginning to emerge as a viable tool for mental health applications. Virtual environments (VE) have been developed which are now demonstrating effectiveness in the areas of clinical psychology and neuropsychology" (p. 21).

In the past fifteen years numerous studies have tested the efficacy of virtual reality in the delivery of CBT. The first works addressed less severe disorders, such as specific phobias. However, more recently, virtual reality programs have been developed for the treatment of more severe problems such as panic disorder, posttraumatic stress disorder, eating disorders and pathological bereavement. A significant number of controlled studies, reviews and meta- analysis reveal the scientific advancement in the field. (see, for a review, Anderson, Jacobs & Rothbaum, 2004; Emmelkamp, 2005;

Krijn, Emmelkamp, Olafsson & Biemond, 2004; Garcia-Palacios et al., 2006; Marks, Cavanagh & Lega, 2007; Powers & Emmelkamp, 2008).

A recent application of ICTs that could help improve efficacy and effectiveness of psychological interventions is Augmented Reality (AR). AR is a modification of VR which includes a combination of both real and virtual elements. Current AR applications for psychological treatment are scarce and address two specific phobias: acrophobia and small animal phobia. Following the guidelines of Öst, Salkovskis & Hellstroöm (1991), our research group has obtained positive preliminary results (through a case study, a case series and a multiple base-line single case design study) on the use of AR for the delivery of one-session exposure for the treatment of specific phobias (Botella et al., 2005; Botella, Bretón-López, Quero, Baños & García-Palacios, submitted; Juan et al., 2005).

2.2. Telepsychology

Another contribution of ICTs to Clinical Psychology is "telepsychology" or online therapy, in which electronic equipment and therapeutic communication converge. Telepsychology can be defined as using ICTs to put patients and mental health professionals in contact to conduct diagnosis or treatment, to disseminate information, or to conduct research studies or any other activity related to mental health care (Brown, 1998). "E-therapy" refers to the delivery of mental health services online. Typically the online services include emails, discussion lists, chats, or audiovisual conferencing. This kind of therapy is proliferating rapidly, and its applications have the potential to advance the field of psychology in a multitude of ways since they are used when face-to-face contact with licensed psychologists is impossible.

Telepsychology has been employed to improve the delivery of CBT. A number of Internet-delivered programs have already become important tools in the health system. The main advantage of online therapy is that it can reach people who might not otherwise seek therapy, such as disabled people or those who live in remote areas; it also reduces the contact time between therapist and patient. Several recent studies have demonstrated the efficacy of Internet-based programs for a variety of mental disorders and health problems such as eating disorders, posttraumatic stress disorder and pathological grief, panic disorder, depression, and chronic pain, among others (see Andersson, 2009; Carlbring & Andersson, 2006; Ritterband et al., 2003, for a review).

2.3. Ubiquitous Computing and Persuasive Computing

We are witnessing the development of ITCs that combine ubiquitous and persuasive computing to improve psychological treatments. The term ”ubiquitous computing” was coined by Weiser in 1991 as a human-computer interaction paradigm in which the computer is integrated into the user’s environment; various small, inconspicuous devices enable the interaction. It employs the use of miniature technology, small systems that communicate among themselves and can be easily integrated into different objects. Examples of these technologies can be found in mobile phones and PDAs with Internet capabilities. One key benefit is that they allow free access to information anywhere at any time. This advancement will likely lead to the fusion of the computer with the objects of daily life (Mattern, Ortega & Lores, 2001).

“Persuasive Computing”, coined by Fogg in 1999, can be defined as the use of technology with the explicit purpose of changing human attitudes and behaviors. Of course, humans have a great capacity for persuasion. However, the persuasive abilities of computers (Fogg, 1999) have made Persuasive Computing one of the technological concepts that have rapidly obtained the attention of the human-computer interaction community. Computers can be more persistent than humans in some ways, by employing many methods to create a convincing experience (text, audios, images, videos, virtual environments, sounds and animations, and more). Additionally, this convincing experience can be easily replicated and distributed to large numbers of people simultaneously. In the field of health, it has been focused on the promotion of healthy habits and the prevention and treatment of unhealthy habits (e.g. smoking). Persuasive computing includes the following main features: a) Timeliness: e.g. messages can be sent to promote healthy food choices in context (Intille, Kukla Farzanfar & Baku, 2003); b) Simulation of experiences: simulations of useful experiences for making appropriate decisions; and, c) Customization: use of customized information in order to ensure that the user follows the instructions.

The most common device in the field of ubiquitous and persuasive computing is the mobile phone. Its fast integration into daily life has benefitted the field of health. For example, it has facilitated the delivery of counseling to HIV patients (i.e. Skinner, Rivette & Bloomberg, 2007), assisted with providing strategies for coping with stress (i.e. Riva, Preziosa, Grassi & Villani, 2006); and aided in the treatment of combat-related stress (Riva, Grassi, Villani & Preziosa, 2007). Because of such benefits, Bang, Timpka, Eriksson, Holm and Nordin (2007) have proposed integrating CBT strategies into mobile phones. Our research team is pursuing this approach, combining phones

with games to prevent obesity in children (E-TIOBE project) and for providing support in the treatment of phobias.

User-centered technologies appear to be the future of the use of ITCs. The rapid development of ITCs and the increasing research on their usefulness will result in a growing number of therapeutic tools with the potential of improving treatments for various mental disorders.

3. Advantages, limitations and ethical issues of using ITCs in psychological treatments.

As previously stated, ICTs are providing new methods for delivering therapy. Yet, it is important to consider their ethical implications and to explore the advantages and limitations of their use in therapy. This reflection is relevant in any research or practice with human subjects.

3.1. Advantages

We began considering these issues when we began our line of research with ICTs and psychological treatments (Baños, Botella & Perpiñá, 1999; Botella, Baños, Perpiñà & Ballester, 1998; Botella et al., 2004). Currently, our dominant perspective is that ICTs can be understood as *new senses* that are incorporated into our *structure to know the world* (using Konrad Lorenz terminology, 1974). While they enhance our abilities to function as living creatures, they are also useful in optimizing therapy.

1. Immersive VR and AR are technologies that allow the creation of 3D computer-generated objects, avatars, environments or situations. Significantly, they simulate reality while providing a new possibility: the user has the illusion of "being" in the computer-generated environment interacting with the VR objects. This is a unique feature that is very relevant for applications in Clinical Psychology. It involves creating *safe* virtual worlds where the patient can explore and experience "new realities"; this feeling of safety is essential in therapy, so that the patient can act without feeling threatened. The "as if" of Kelly (1955) is a good example. This perspective is also included in the "need for safety and protection" of Bowlby's (1973) attachment theory.

The virtual context allows patients to approach situations that they perceive as threatening in a gradual way, at their own pace, with complete safety and protection.

2. Confronting and overcoming fears in therapy is essential, and VR allows total control in this area. Information can be presented gradually, in such a way that the patient can progress from easier tasks to more difficult ones. This work in the virtual world helps patients master the strategies needed to overcome their fears in the real world.

3. Bandura (1977) stated in his theory that of all possible sources of personal efficacy, performance achievements are especially useful. VR is an excellent source of information on performance achievements, since numerous methods can be designed to assure the patients' success in each of their virtual experiences; patients can also practice potential difficulties or occasional failures. According to Bandura, once strong expectations of efficacy have been established through repeated successes, the potential negative impact of occasional failures will be reduced. Failures that are overcome with the patients' effort will strengthen the patients' persistence and involvement. It is of great importance that the patients view themselves as competent, and efficacious. Likewise, is essential for the patient to associate personal competence with factors such as consistency and effort in the environment, which give rise to a larger sense of strength and mastery.

4. Another important advantage of VR is that it and other ICTs can help build a new conceptual framework for understanding how the human mind works. If normal mental processes are better understood, disturbances in mental processes which are involved in the development of mental disorders can be more easily studied (Baños, Botella & Perpiñá, 1999).

5. Virtual worlds provide additional advantages. As they are "virtual", it is not necessary for them to adhere to the rules of space and time; indeed, they can be said to exist "beyond reality". As a result, researchers do not have to wait for specific events to occur. Rather, they can simulate them whenever appropriate for the patient and the therapy process. Also, the possibilities for self-training are enhanced. A patient can work on a concrete issue for any duration at any time and in any place (through the use

of mobile devices). Virtual worlds can also help to generalize the progress achieved in therapy, because the patient can practice in different virtual contexts.

Additionally, VR makes it possible to alter the feared environment at the patient's and therapist's convenience. It is flexible enough to allow the existence of a series of contexts where patients can confront not only their concrete fears, but also elements and situations beyond those concrete fears. For example, an individual with a fear of public speaking is afraid of making mistakes and being negatively judged by an audience. In a virtual world, the patient can experience different reactions from the audience, from an attentive and supportive reaction to a very negative reaction which would only rarely occur in the real world (insults, throwing objects, and so forth). As another example, an individual suffering claustrophobia fears staying in a locked room and not being able to open the door. This situation can be simulated virtually, with the additional experience of having the walls and the ceiling move until the individual is confined in a square meter space. These examples illustrate how the patients' interaction with the virtual environment can be managed at different levels and in different ways. This promotes overlearning and enhances mastery of one's fears. Thus, the purpose of virtual worlds is not merely to recreate reality, but to create therapeutic contexts containing elements that are relevant to the patients and their problems, some of which might not otherwise be available (Wann, Rushton, Smyth & Jones, 1997).

Because patients can work on concrete interactions with the world repeatedly and at their own pace means that they can experience the consequences of those interactions many times. This is demonstrated by one of the first well-known VR applications, flight simulation. Users can practice multiple situations, difficulties, mistakes, dramatic consequences and so forth, while gaining knowledge and skills for dealing with these experiences in the real world. Progress achieved in the virtual world regarding a feared situation will help the patient to "live reality" in a different way and to generate new internal models of the world and ways to interact with it (Korzybski, 1958). These internal models will allow users to view themselves and the world from a new perspective. In summary, VR can promote operational thinking (Piaget, 1926) and improve one's capacity to perceive the world, while enhancing one's fundamental capacity to imagine "what would happen if…" (Tart, 1991). Virtual experiences not only have an impact on patients, but can also leave imprints as patients incorporate experiences to memory, to cognitive structures and to life in general.

These effects of VR therapy entail advancements of psychological treatments from an ethical point of view. Virtual worlds support and protect the patient along the therapeutic process; this is evident in one of the best applications of VR to therapy, exposure. Therefore, it is not surprising that that some studies have found that patients show a preference for VR exposure over exposure in the real world. For example, Rothbaum, Hodges, Smith and Lee (2000) gave flight phobia patients a choice between VR exposure and in vivo exposure; most of them chose VR exposure. Similar data have been reported by García-Palacios, Botella, Hoffman and Fabregat (2007) in a study wherein patients were asked about their preferences regarding VR exposure and in vivo exposure. Again, most of them chose VR.

6. As previously mentioned, other ICTs related to ubiquitous and persuasive computing (the Internet, mobile phones, PDAs, GPS, several types of sensors and others), offer the important benefit of making information available at any time (e.g. information about the treatment rationale, instructions for applying treatment strategies, and more) in any context (at the patient's home, on the street, at the workplace, and so forth), with the possibility of immediate feedback. These strengths can help improve the provision of mental health care. Benefits from the perspective of axis 2 (the clinical utility axis) are enormous given that the technologies can help reduce costs and facilitate the dissemination of information that, until now, was impossible.

However, the use of ICTs in clinical psychology has disadvantages. It is important to pay careful attention to the limitations and ethical issues related to the progress of this field.

3.2. Limitations and ethical issues

The use of ICTs in clinical psychology is a recent development that is gaining momentum and offers significant benefits; therefore, researchers must be aware of the issues and ethical considerations that can arise in applications of ICTs.

1. Ten years ago, susceptibility to cybersickness and aftereffects of treatment were particularly of interest. It was an ethical requirement to analyze the potential for adverse side effects of treatment. Cybersickness is a form of motion sickness that includes symptoms such as nausea, vomiting, eyestrain, disorientation, ataxia, and vertigo (Riva, Bacchetta, Baruffi, Rinaldi, & Molinari, 1999). Aftereffect symptoms

include disturbed locomotion, perceptual-motor disturbances, flashbacks, drowsiness, fatigue, and lowered arousal. Rizzo et al., (1998) considered of the possible relationship between various side effects and the features of different clinical groups. They also analyzed various issues relevant to the application of VR in clinical populations. Researchers were concerned about the duration of exposure sessions and whether patients would be unable to complete VR sessions due to cybersickness. However, after ten years of experience treating VR patients suffering mental health problems this has not proven to be a problem in efficacy studies. Only a minority of patients cannot benefit from VR therapy for such reasons. In these cases, it is important to find alternative treatments or ways to minimize the effects of cybersickness or other aftereffects.

2. Another historically important consideration was whether a patient's age was a key factor in appropriateness of VR treatment. At the time, it was recommended to be very cautious when treating children and elderly patients with emotional problems. After many years of VR application, it is clear that children with phobias can benefit from VR treatment (e.g. Botella et al., 2007), as can the elderly. For example, patients 60 years old and older have overcome phobias (claustrophobia and storm phobias) over many years of VR treatments (Botella et al., 1998; Botella et al., 2006). However, it is necessary to provide more empirical data on the benefits of ICTs in clinical psychology for these populations who suffer various mental problems.

3. Another early concern was whether VR ought to be applied to more severe anxiety disorders, such as posttraumatic stress disorder (PTSD) or panic disorder with agoraphobia. The concern was that VR exposure might not benefit these patients, and could even have negative effects (such as causing sensitization instead of desensitization). Because of this, it was considered important to analyze possible negative effects and to be cautious about the application of ICTs. Researchers have taken these recommendations into consideration, and the preliminary data indicate that VR could benefit PTSD patients without causing negative side effects. (Rothbaum, Hodges, Ready, Graap, & Alarcon, 2001; Difede & Hoffman, 2002). In addition, our research group has conducted a controlled clinical trial showing efficacy and effectiveness at short- and long-term after VR exposure for the treatment of panic disorder (Botella et al., 2007).

4. Another original belief was that VR therapy should not be used with populations suffering from certain types of psychopathology or having features of various psychotic, bipolar, paranoid, substance abuse, and other disorders where reality testing and identity problems were present. Because VR simulates reality with a high degree of fidelity, this was believed to present issues for populations with difficulties in distinguishing between reality and imagination (such as those with high vulnerability to psychosis). However, there are currently a number of VR applications for psychosis that show that VR can in fact be applied to these populations. Freeman and his colleagues at the Institute of Psychiatry in London have demonstrated that VR is a safe and acceptable method for studying paranoia in the laboratory. Indeed, in a recent work by Freeman (2008), "Studying and Treating Schizophrenia Using Virtual Reality: A New Paradigm", the author states: "The use of virtual reality (VR) interactive immersive computer environments allows one of the key variables in understanding psychosis, social environments, to be controlled, providing exciting applications to research and treatment… VR, suitably applied, holds great promise in furthering the understanding and treatment of psychosis" (p. 605).

Ten years ago, Baños, Botella and Perpiñá (1999) supported the application of VR in the field of psychopathology. They demonstrated that VR could aid in generating useful models for studying basic processes and their disturbances. For example, VR can assist in studying the processes involved in reality testing, which is one of the most intriguing challenges for psychopathologists. Results of this line of inquiry might reveal some of the most important aspects of the distinction between psychosis and neurosis. Researching how VR can influence reality judgments might shed some light on how we attribute reality to our perceptions, or other cognitive information. It is also important to study and understand the same metacognitive processes in psychotic individuals. The field is continuing to progress, and there are recommendations for further studies. Perhaps VR can provide answers to questions underlying mental disorders, and can spur further advances (as happened with the pioneering studies by Watson and Jones).

5. Another consideration is the limitation of the continuing high cost of some of the necessary devices and systems. Although costs have decreased significantly in recent years, it remains expensive to develop technological tools and equipment required for program implementation; many therapists and mental health care institutions find them unaffordable. Decreasing these costs is an ethical imperative. Another challenge is that

psychologists and patients unfamiliar with ICTs may resist their use in therapy if they lack the confidence and skills required for the programs' application. It is important to work on increasing the acceptability of ITCs by patients and therapists.

6. Several concerns have been raised about self-help procedures in telepsychology and online therapy. One is the issue of self-diagnosis. For example, a patient could begin an Internet-based program for social phobia because he thinks he has this issue; yet, this might not be an accurate diagnosis. A true diagnosis must be conducted by an expert; therefore, online treatment programs must consider the risks involved in offering treatment with only self-diagnosis. Another issue is that patients usually have access to all the self-help information simultaneously, as opposed to a program wherein each step must be suitably completed before advancing to the next step; this could lead to negative effects. It is important to establish criteria to protect people from the possible negative effects of these kinds of treatments.

Online mental health services present several important legal and ethical issues, most of which remain unresolved: determining the identity of the recipient of services, maintaining confidentiality, legal jurisdiction, and technical competence of the therapist (Gingerich, 2002). To remedy this, various associations of online professionals and health care organizations have developed codes of conduct for online services in recent years. The following links are examples of some guidelines that have been published by various professional organizations: American Psychological Association (1997) Services by Telephone, Teleconferencing, and Internet; American Counseling Association (1999) Ethical Standards for Internet Online Counseling; International Society for Mental Health Online & Psychiatric Society for Informatics (2000) Suggested Principles for the Online Provision of Mental Health Services; Internet Health Coalition (2000) e-Health Code of Ethics (draft); National Board of Certified Counselors (1997) Standards for the Ethical Practice of Web Counseling; Health on the Net Foundation (1997) HON Code of Conduct for Medical and Health Websites; American Medical Informatics Association (1998) Guidelines for the Clinical Use of Electronic Mail with Patients.

In summary, although many of the anticipated limitations of the application of ICTs in therapy appear to have been overcome, it is important to remain cautious. VR and other ICTs offer many potential benefits and can be integrated into established psychological and psychiatric theory and practice. However, ICT researchers must

follow the ethical guidelines for the standard practice of conventional psychological and psychiatric research and therapy.

3.3. Ethical considerations for the future

Technology offers enormous potential for change; it changes us as well as the world at large. For example, a multitude of human advances are due to the use of tools. Likewise, innumerable global changes have resulted from the development and use of increasingly sophisticated devices (carts, boats, trains, cars, planes, rockets, computers, and others). These extraordinary advances have contributed to the development of various cultures and civilizations and the conquest of new frontiers on Earth and beyond.

Despite the dominance of certain countries or cultures at each moment in history, the culture of a plethora of populations has persisted. The abundance of cultures in the world has contributed to the rich cultural legacy of humans. However, current powerful technological advances like the Internet, VR or other ICTs have the potential to provoke significant changes in the path of human progress.

Internet

This technology allows us to quickly access any person or place. This offers enormous advantages, but also limitations. Clear advantages include ease and fluency of communication and the potential for great advances in knowledge. Possible limitations of the Internet include the risk of transforming the world into a “global village” characterized by increased cultural uniformity and loss of diversity. Additionally, the Internet has spurred a phenomenon of social isolation which leads to reduced communication and intimacy with friends and family. Many people no longer need to venture from their homes in order to work or enjoy themselves. The Internet’s growing entertainment function includes new interfaces and increasingly sophisticated stimuli. “Generation C” has been described by Peter Marsh as the generation of the Internet. Its main features are creativity, connectivity, collaboration and communication. Marsh notes that this generation has developed with the ideology of the Internet, including free access to information, cooperation and information sharing. While this may be a benign characterization, descendants of Generation C may be have the misfortune of having access primarily to “connected experiences” in which each individual is isolated from others. Thus, they would be “connected human beings” with more personal space but increased isolation.

While incorporating these new senses, changes in our structures for knowing the world could occur, similar to the loss of hair in primates or the loss of molars in some mammals. However, these changes could be more dramatic and dangerous and could result in the loss of valued behaviors such as non-verbal language or physical contact. The effects that this could have in the long term regarding social functioning and physical development are unknown. Perhaps the new "digital natives" (using Marc Prensky terminology) in the very distant and hypothetical future will not need legs; they will not need to go anywhere when all the information and enjoyment they need will be brought to them. In this scenario, all spheres of life would be digital.

Mobile devices and other ICTs

The benefits of using new ICT technologies (mobile phones, PDA, GPS, biosensor, etc.) have been previously explained. Their main advantage is their ability to provide information where and when it is necessary. However, it is important to consider possible limitations; one is the availability of information. An increasing amount of information is becoming available, some of which is general, but some of which is personal, which brings up privacy issues. For example, the increasing popularity of social networks like Facebook.com raises questions about ownership of and access to information. For example, could a piece of information uploaded by a user when he was 12 damage him when he is 30? Who owns the information? It is necessary to make laws about information ownership and use, privacy, and the protection of minors. Another danger related to this is the possibility of using these devices illegally, considering that it is possible to do such things as locate a person using GPS, or to monitor a person's physiological state 24 hours a day with small sensors. A "big brother" could use this technology to observe and obtain continuous information from people. This leads to ethical considerations such as whether we are putting ourselves at risk in the area of privacy.

Another danger involves the use of attractive devices that help us in numerous valuable ways, but which also entails a high price. Clear examples are the mobile phone and email. Some people have shown resistance to mobile phones; however, most have surrendered to this valuable technology. These devices are increasingly precise, small, useful and indispensable. Many people develop a dependency on these technologies (and find it very difficult to leave the mobile phone at home or not check email daily). Thus, technology "controls" us while also making it easier for others control us. For example, with email people receive information immediately, which is convenient; however, they then are obliged to answer emails daily, which consumes a large amount

of time. Therefore, it is advisable to consider limits in the use of these technologies at the workplace and also in leisure activities.

Another important reflection concerns the drive to provide a more sophisticated (and thus presumably better) education to one's children. In western society, it is considered important for children to study and understand ICTs and to have immediate access to information and knowledge in order to excel in school and beyond. It is desirable to have the fastest Internet connection in schools and homes, and to develop ways to be online anywhere at any time. Ever-smaller memory devices contain information essential to life. From this evidence it would appear to be desirable to develop devices and systems with great storage capacity that would enable continuous Internet connectivity. One might imagine that some would accept a microchip under their skin with a hard disk and capability for continuous updates, or that babies could have such devices implanted from birth. What the true cost of this be?

Virtual Reality

If the former is extraordinary, the possibilities of VR for the future are equally impressive. People might not only be able experience simulations of reality, but might also be able to *live* any "reality" that could be programmed. Theoretically, anything is possible; the human imagination and technological advances are the only limitations.

This could lead result in a significant shift in our experience of the physical world, thanks to new technologies that will enable greater control and mastery of the physical world. For example, it might no longer be necessary to physically go to New York in order to have the experience of "being" in New York. A person using a sophisticated VR system to "travel" could believe that he has been in New York when in fact he has not. How will humans adapt to these new capabilities? The possibilities can cause fear and anxiety, and might cause some people to revert to their private spaces, only to emerge when it is absolutely necessary or profitable to do so.

It is likely that the learning experiences of human beings will change. One might reconsider Albert Bandura and his vicarious learning paradigm. This author took into consideration aspects that remained outside the classic and operant conditioning paradigms (attention processes, storage processes, motivational processes, and so forth). He established an important distinction between learning and performance and emphasized the possibility of establishing new patterns of response by vicarious learning. It is likely that VR is evolving into a new learning paradigm: "virtual learning". It is incumbent upon researchers to determine its parameters and rules.

In the past, parents and the community (the tribe) decided what was necessary to teach, including the ideas, procedures and norms of the tribe. Later, teachers, the educational system, and books were added in order to significantly expand the limits of knowledge. However, people only had access to certain mentors and books (which were available in their immediate environment). The creation of printing expanded the availability of books. The development of revolutionary technologies such as the telephone and television resulted in greater access to various ideas, contexts and information sources. However, current advances including the combination of different ICTs like Internet, VR, mobiles devices, sensors and more result in dramatically expanded capabilities. This new learning framework offers clear advantages, including providing access to an enormous amount of information, designing and defining experiences, creating an experience or situation, organizing and structuring an experience or situation at will, adapting it to one's needs, practicing at a customized pace, recording an experienced situation, saving it, providing feedback on it, and repeating an experience as many times as needed in order to consolidate or generalize skills and strategies. The possibilities are infinite; the important question is whether all the resulting effects will be positive ones.

As previously mentioned, caution should be taken to avoid disturbing human development. Children must continue to perform certain physical functions, to develop motor skills and social skills and play with other children. The development process can be adversely affected if there is no control over the use of new technologies such as VR. These negative effects have already been observed with TV, but VR could be even more dangerous. It could cause problems in the cognitive organization of human experiences, memories, judgments, beliefs, and in the distinction between the self and the environment, leading one to question: Is this real? Is that me? Was that me? Is that happening to me? Did that happen to me? Was it a dream? Was it real? Was it a virtual experience? Having multiple virtual experiences during the years could confuse reality judgments and the perception of self and identity.

Another important advance to consider is the creation of new entities. One example is *avatars*, virtual beings that live in the memory of computers and perform in virtual worlds. Another is *robots*. The creation of these entities leads to ethical and philosophical issues about their degree of individuality or identity. One of the most intriguing questions is to what extent these entities have consciousness. In other words, will human beings be capable of creating artifacts with consciousness?

The word “robots” usually connotes machines with human appearances. However, other kinds of robots (already in existence) include autonomous minuscule structures that can be installed in human bodies; they have goals and perform actions designed to develop and maintain one’s physical and mental health, generating what some authors call “technological consciousness”. The risk is not only that robots will be integrated into human lives, but also that humans will be assimilated by them. Along this line, some authors argue that a biotechnological revolution will soon occur that will change humans’ position in the world. Ray Kurzweil, author of *The Singularity Is Near: When Humans Transcend Biology,* predicts that the fusion between human intelligence and artificial intelligence will lead to a new kind of intelligence, radically different from what can currently be imagined. This is predicted to happen soon; intelligent “nanorobots” will be integrated into human bodies and into the environment, resulting in total immersion into a virtual reality. If this does indeed transpire, humans should be prepared with deal with the possibility of losing their human identities in favor of some new entity.

4. What can humans do?

First, as in any problem-solving process, one must stop and think. The human being is the most powerful predator and also the most dangerous destructive agent on Earth. However, ICTs are also powerful, and are increasingly ubiquitous. They will intrude everywhere, becoming “liquid” technologies. Therefore, care, caution and control are highly recommended.

It is useful to recall Bertalanffy’s systems theory (1967, 1968): his vision of the human being resembles an open system and an active agent. Bertalanffy stated that people are not merely passive recipients of stimuli from an external world; rather, they “create” their realities. In the context of ICTs, his relevant belief is that any intervention or human artifact exacts a price, no matter how benign the motivation of the creator. To what universe and into what kind of being (or cybernetic organism) are humans evolving? Researchers do not want to contribute to the destruction or disturbance of central and necessary human processes; rather, they would like to promote useful patterns that could be useful in the process of human evolution, such as personal growth or empathy. Therefore, it is becoming more urgent to cautiously define the framework and contexts of intervention, to delimitate the programs of observation and

intervention accurately and to anticipate possible side effects and intervene as needed. This should be done with adequate scientific rigor. It is necessary to conduct research programs with accurate and strict methodological and ethical control. This is uncharted territory into which researchers venture blindly. To succeed, it will be important to make use of all of human accumulated knowledge. It must be remembered that ICTs are new tools and much work is yet to be done, including creating a theoretical framework that allows making predictions and organizing the findings.

ICT applications have a promising future, and the existing applications are merely the beginning of an enormous progression. As previously mentioned, it is difficult to conceive of an application that cannot be created using existing technology. The creation of future applications is only a question of talent, time and resources; technology itself is neutral. The important questions are these: In which direction should technological development proceed? Which applications will be the most useful, have the most impact or benefit the most people? Which psychological cyberspace is most effective and which cybertherapy is most appropriate to pursue? Answering these critical questions is a challenge to which all researchers can apply themselves.

5. Acknowledgments

The research presented in this paper was funded in part by Ministerio de Educación y Ciencia, Spain, PROYECTOS CONSOLIDER-C (SEJ2006-14301/PSIC), by Ministerio de Ciencia e Innovación (PSI2008-04392), by Generalitat Valenciana, Conselleria de Educación Programa de Investigación de Excelencia PROMETEO (2008/157), and by CIBER. CIBER Fisiopatología de la Obesidad y Nutrición is an initiative of ISCIII.

6. References

American Counseling Association (1999). Ethical Standards for Internet Online Counseling. Available at http://www.angelfire.com/co2/counseling/ethical.html

American Medical Informatics Association (1998) Guidelines for the Clinical Use of Electronic Mail with Patients. Available at http://www.amia.org/mbrcenter/pubs/email_guidelines.asp

American Psychological Association Task Force on Psychological Intervention Guidelines (1995). *Template for developing guidelines: Interventions for mental*

disorders and psychological aspects of physical disorders. Washington, D. C.: American Psychological Association.

American Psychological Association (1997). *Services by Telephone, Teleconferencing, and Internet.* Available at http://www.apa.org/ethics/stmnt01.html

Andersson, G. (2009). Using the Internet to provide cognitive behaviour therapy. *Behaviour Research and Therapy, 47*, 175-180.

Anderson, P., Jacobs, C. & Rothbaum, B.O. (2004). Computer-supported cognitive behavioral treatment of anxiety disorders. *Journal of Clinical Psychology, 60*(3), 253-67.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavior change. *Psychological Review,* 84, 191-215.

Bang, M., Timpka, T., Eriksson, H., Holm, E. & Nordin, C. (2007). Mobile phone computing for in-situ cognitive behavioral therapy. *Studies in Health Technology Informatics, 129,* 1078-82.

Baños, R.M., Botella, C., y Perpiñá, C., (1999). Virtual Reality and Psychopathology. *Cyberpsychology and Behavior, 2 (4),* 283-292.

Bertalanffy, L. (1967) *Robots, Men and Minds: Psychology in the Modern World*, New York: Braziller,

Bertalanffy, L. (1968) *General System theory: Foundations, Development, Applications*, New York: George Braziller.

Botella, C., Baños, R.M., Perpiñá, C. & Ballester, R. (1998) Realidad Virtual y Tratamientos Psicológicos. *Análisis y Modificación de Conducta, 24 (93),* 5-26.

Botella, C., Baños, R.M., Perpiñá, C., Villa, H., Alcañiz, M. & Rey, B. (1998). Virtual reality treatment of claustrophobia: a case report. *Behaviour Research and Therapy, 36,* 239-246.

Botella, C., Bretón-López, J., Quero, S., Baños, R.M., & García-Palacios, A. (submitted). *Treating Cockroach Phobia with Augmented Reality: A Controlled Study.*

Botella, C., Quero, S., Baños, R.M., Perpiñá, C. & García-Palacios, A. (2004). Virtual Reality and Psychotherapy. In G. Riva, C. Botella, P. Légeron, & G. Optale (Eds.), *Cybertherapy: Internet and Virtual Reality as Assessment and Rehabilitation Tools for Clinical Psychology and Neuroscience* (pp. 37-52). Amsterdam: IOS Press.

Botella, C., Juan, C., Baños, R.M., Alcañiz, M., Guillén, V., & Rey, B. (2005). Mixing Realities? An application of Augmented Reality for the treatment of cockroaches phobia. *CyberPsychology and Behaviour,* 8(2), 161-171.

Botella, C., Baños, R.M., Guerrero, B., García-Palacios, A., Quero, S., & Alcañíz, M. (2006) Using a flexible Virtual Environment for Treating a Storm Phobia. *PsychNology Journal*, 4(2) 129-144.

Botella, C., Lasso de la Vega, N., Castilla, D., García-Palacios, A., López Soler, C., Baños, R., & Alcañiz, M. (2007). Virtual reality for the application of psychological treatments in children: darkness phobia. *Cybertherapy 12 Conference: Transforming Health care Through Technology*. Washington D.F.

Botella, C., Villa, H., García-Palacios, A., Baños, R. M., Quero, S., Alcañiz, M. y Riva, G. (2007) Virtual Reality Exposure in the Treatment of Panic Disorder and Agoraphobia: A controlled study*. Clinical Psychology and Psychotherapy,* 14 (3) 164-175.

Bowlby, J. (1973). *Attachment and loss Vol. 2: Separation, anxiety and anger.* Nueva York: Basic Books.

Brown, F. W. (1998). Rural telepsychiatry. P*sychiatric Services*, *49*, 963-964.

Carlbring, P. & Andersson, G. (2006). Internet and psychological treatment. How well can they be combined?. *Computers in Human Behavior, 22*, 545-553.

Chambless, D. L., Sanderson, W. C., Shoham, V., Bennett Johnson, S., Pope, K. S., Crits-Christoph, P., Baker, M., Johnson, B., Woody, S. R., Sue, S., Beutler, L., Williams, D. A. & McCurry, S. (1996). An update on empirically validated therapies. *Clinical Psychologist, 49,* 5-18.

Chambless, D. L. & Hollon, S. D. (1998). Defining empirically supported therapies. *Journal of Consulting and Clinical Psychology, 66,* 7-18.

Difede, J. & Hoffman, H. (2002). Virtual reality exposure therapy for World TradeCenter post-traumatic stress disorder: A case report. *Cyberpsychology & Behavior, 5*, 529-535.

Emmelkamp, P. M. (2005). Technological Innovations in Clinical Assessment and Psychotherapy. *Psychotherapy Psychosomatics,* 74, 336-343.

Fogg, B. J. (1999). The elements of computer credibility. *Proceedings of the SIGCHI conference on Human factors in computing systems: the CHI is the limit*. (pp. 80–87). Pittsburgh, Pennsylvania, United States.

Freeman, D. (2008). Studying and Treating Schizophrenia Using Virtual Reality: A New Paradigm. *Schizophrenia Bulletin, 34* (4), 605-610.

Garcia-Palacios, A., Botella, C., Hoffman, H. & Fabregat, S. (2007). Comparing Acceptance and Refusal Rates of Virtual Reality Exposure vs. In Vivo Exposure by Patients with Specific Phobias. *Cyberpsychology and Behavior, 10*, 722-4.

García-Palacios, A., Botella, C.; Hoffman, H., Baños, R.M., Osma, J., Guillén, V. & Perpiñá, C. (2006). Treatment of Mental Disorders with virtual reality. En M. J. Roy (Ed.), *Proceedings of the NATO advanced research workshop on PTSD*. Amsterdam: IOS Press.

Gingerich, W. J. (2002). Online social work: Ethical and practical considerations. In A. R. Roberts & G. J. Green (Eds.), *Social worker's desk reference* (pp. 81-85). New York: Oxford University Press.

Health on the Net Foundation (1997). *HON Code of Conduct for Medical and Health Websites.* Available at http://www.antiagingresearch.com/hon.shtml

International Society for Mental Health Online & Psychiatric Society for Informatics (2000). *Suggested Principles for the Online Provision of Mental Health Services.* Available at http://www.ismho.org/suggestions.asp

Internet Health Coalition (2000). *e-Health Code of Ethics (draft).* Available at http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1761853

Intille, S. S., Kukla, C., Farzanfar, R. & Bakr, W. (2003). Just-in-time technology to encourage incremental, dietary behavior change. *Proceedings of the AMIA 2003 Symposium*.

Jones, M.C. (1924). A laboratory study of fear: The case of Peter. *Pedagogical Seminary, 31*, 308-315.

Juan, M. C., Alcañiz, M., Monserrat, C., Botella, C., Baños, R. & Guerrero, B. (2005). *Using Augmented Reality to treat phobias.* IEEE Computer Graphics and Applications, Nov-Dic, 31-37 .

Kelly, G.A. (1955). *The Psychology of personal constructs*. Nueva York. Norton.

Kurzweil, R. (2005) *The Singularity Is Near: When Humans Transcend Biology,* Viking Adult.

Korzybski, A. (1958). *Science and sanity: An introduction to non-Aristotelian systems and general semantics.* Lakeville. Connecticut. The International Non-Aristotelian Publishing Company.

Krijn, M., Emmelkamp, P.M.G., Olafsson, R.P. & Biemond, R. (2004). Virtual reality exposure therapy of anxiety disorders: A review. *Clinical Psychology Review, 24*, 259-281.

Lorenz, K. (1974). *La otra cara del espejo*. Madrid. Plaza y Janés.

Marks, I. M., Cavanagh, K. & Gega, L. (2007). Computer-aided psychotherapy: revolution or bubble?. *British Journal of Psychiatry, 191*, 471-3.

Mattern, F., Ortega, M. & Lorés, J. (2001). Ubiquitous Computing: The Trend Towards the Computerization and Networking of All Things. Upgrade, 2, (5). http://www.upgrade-cepis.org/issues/2001/5/up2-5Present.pdf.

National Board of Certified Counselors (1997). *Standards for the Ethical Practice of Web Counseling*. Available at http://www.nbcc.org/ethics/wcstandards.htm

Öst, l., Salkovskis, P. & Hellstroöm, K. (1991). One-session therapist directed exposure vs. self-exposure in the treatment of spider phobia. *Behavior Therapy*, *22*, 407-422.

Piaget, J. (1926). *The language and thought of the child.* Nueva York. Harcourt Brace.

Powers, M. B. & Emmelkamp, M. G. (2008). Virtual reality exposure therapy for anxiety disorders: A meta-analysis. *Journal of Anxiety Disorders, 22,* 561-569.

Ritterband, L. M., Gonder-Frederick, L. A., Cox, D. J., Clifton, A. D., West, R. W., & Borowitz, S. M. (2003). Internet interventions: in review, in use and into the future. *Professional Psychology: Research and Practice*, 34, 527-534.

Riva, G., Bacchetta, M., Baruffi, M., Rinaldi, S., & Molinari, E. (1999). Virtual reality based experiential cognitive treatment of anorexia nervosa. *Journal of Behavioral Therapy and Experimental Psychiatry, 30 (*3), 221-230.

Riva, G., Preziosa, A., Grassi, A. & Villani, D. (2006). Stress management using UMTS cellular phones: a controlled trial. *Studies in Health Technology and Informatics, 119*, 461-463.

Riva, G., Grassi, A., Villani, D., Gaggioli, A. & Preziosa, A. (2007). Managing exam stress using UMTS phones: the advantage of portable audio/video support. *Studies in Health Technology Informatics, 125,* 406-8.

Rizzo, A. A., Wiederhold, M. & Buckwalter, J. G. (1998). Basic issues in the use of virtual environments for mental health applications. *Studies in Health Technology and Informatics, 58,* 21-42.

Rothbaum, V. O., Hodges, L. F., Kooper, R., Opdyke, D., Williford, J. S. & North, M. (1995). Virtual-Reality Graded Exposure in the Treatment of Acrophobia - A Case Report. *Behaviour Therapy*, *26* (3), 547-554.

Rothbaum, B.O., Hodges, L., Smith, S., Lee, J.H. & Price, L. (2000). A controlled study of virtual reality exposure therapy for fear of flying. *Journal of consulting and Clinical Psychology, 68* (6), 1020-1026.

Rothbaum, B. O., Hodges, L., Ready, D., Graap, K., & Alarcon, R. D. (2001). Virtual reality exposure therapy for Vietnam veterans with posttraumatic stress disorder. *Journal of Clinical Psychiatry, 62*, 617-622.

Skinner, D., Rivette, U. & Bloomberg, C. (2007). Evaluation of use of cellphones to aid compliance with drug therapy for HIV patients. *AIDS Care, 19* (5), 605-7.

Tart, C. T. (1991). Multiple personality, altered states and virtual reality: the world simulation process approach. *Dissociation, 3,* 222-233.

Wann, J., Rushton, S., Smyth, M. & Jones, D. (1997). Rehabilitative environments for attention and movements disorders. *Communications of the ACM, 40,* 49-52.

Watson, J. & Rayner, R. (1920). Conditioned emotional reactions. *Journal of Experimental Psychology, 3*(1), 1-14.

PsychNology Journal, 2009
Volume 7, Number 1, 101 – 112

Telepresence and Video Games: The Impact of Image Quality

Cheryl Campanella Bracken[♦*] and Paul Skalski[♦]

[♦] Cleveland State University
(USA)

ABSTRACT

This study investigates the impact of video game image quality on telepresence. Past research has demonstrated positive associations between television image quality and presence and video game technology and presence. No study to date, however, has examined the presence effects of video games played in high definition, which is becoming increasingly common due to the diffusion of new TV technologies into homes. This paper reports the results of an experiment in which image quality was manipulated. The results of the study provide some support for image quality affecting telepresence. Specifically, higher quality images in video games led to higher levels of immersion. These findings are discussed along with suggestions for future research.

Keywords: *Video Games, Telepresence, Presence, HD games.*

Paper Received 05/08/2008; received in revised form 27/03/2009; accepted 03/04/2009.

1. Introduction

Video games have become one of the most popular forms of media in the United States and abroad. Global sales in the industry are projected to exceed $46.5 billion dollars by 2010 (Kolodny, 2006), and 69% of American heads of households currently play video games (Entertainment Software Association, 2006). The popularity of games has been fuelled in part by advancements in gaming technology, a trend that has persisted since the earliest days of the medium (Skalski, 2004). Over time, games have evolved considerably in graphic richness and realism. The simplistic character representations in games like *Pac-Man*, for example, have now been replaced with the realistic human figures and environments in popular recent titles like *Grand Theft Auto IV* and *Halo 3*[1]. These and other advances in game technology have important

Cite as:
Bracken, C. C., & Skalski, P. (2009). Telepresence and Video Games: The Impact of Image Quality. *PsychNology Journal, 7*(1), 101 – 112. Retrieved [month] [day], [year], from www.psychnology.org.

[*] Corresponding Author :
Paul Skalski, 2121 Euclid Ave., Cleveland State University, Cleveland, OH 44115
Phone: 216-687-5042, E-mail: p.skalski@csuohio.edu

consequences for how games are experienced (Ivory & Kalyanaraman, 2007). In particular, they are expected to contribute to the sense of presence, or "perceptual illusion of nonmediation" (Lombard & Ditton, 1997), felt by users.

Presence (also referred to as "telepresence") has recently been identified as a potentially important variable in video game research that may affect use and a variety of outcomes of exposure, ranging from enjoyment to aggression (Tamborini & Skalski, 2006; Lee & Peng, 2006). Few studies, however, have examined the relationship between exposure to game technology and presence. Tamborini et al. (2004) found that playing a game created a stronger sense of presence than just observing a game, presumably due to the addition of interactivity. Though many technological features of video games are expected to contribute to the sensation of presence, one that has received no attention to date is image quality. High Definition Television (HDTV) sharply improves the quality of TV images and, with TV and movie clips, has been shown to relate positively to the experience of presence (Bracken, 2005). But what about video games, which add the crucial feature of interactivity to HDTV and other high-quality imagery? Several video game systems, including the Sony PS2, Microsoft Xbox, and new generation Nintendo Wii, have an adapter that allows players to play games in improved image quality through a component video connection, resulting in lined doubled progressive scan or Enhanced Definition (ED) gaming. Additionally, the new generation Microsoft Xbox 360 and Sony PS3 consoles are capable of displaying High Definition (HD) images. Since HDTV is expected to diffuse rapidly in coming years (Dupagne & Seel, 2006) in part through game consoles, this type of gaming should become increasingly common in the future, raising questions about its effects on players.

This research investigates the effect of image quality on presence-related reactions to video games. Participants in this study played a game in either high definition/HD (higher image quality) or standard definition/NTSC[2] (lower image quality) and then completed measures of presence dimensions (spatial presence and immersion).

[1] In *Grand Theft Auto IV*, players assume the identity of a European immigrant named Niko Belic, who must fight for survival in the vast crime-infested streets of the fictional location Liberty City (modelled after New York City). *Halo 3* is the final instalment of a science fiction game trilogy in which players control a futuristic soldier helping to defend Earth from hordes of alien aggressors.

[2] National Television Systems Committee is the current analog Standard for television in the United States.

2. Telepresence

The concept of presence was introduced in the early 1980s as a sense of "being there," (Minsky, 1980), and as a "sensation of reality." The concept has been examined in research and theory in diverse fields. In an effort to develop a cohesive definition, an online discussion of presence researchers concluded that presence is "a psychological state or subjective perception in which even though part or all of an individual's current experience is generated by and/or filtered through human-made technology, part or all of the individual's perception fails to accurately acknowledge the role of the technology in the experience" (International Society for Presence Research [ISPR], 2000). While there is not a universal definition (see Lee, 2004; Lombard & Ditton, 1997 for more details), the ISPR's definition has been widely accepted. The field has moved away from earlier conceptualizations of self and personal presence (Biocca, 1997) and has generally accepted the existence of several sub-dimensions of presence (Freeman, 2004), with the numbering of dimensions varying from one to six (or more). Many of the competing conceptualizations include three similar sub-dimensions of presence - though terms employed vary. Freeman (2004) argues these can be classified consistently into spatial/physical presence, immersion, and social realism. *Spatial presence* refers to the sense of "being there" in the space of the media environment (Wirth et al., 2007). *Immersion* involves being perceptually and psychologically "submerged" in a mediated environment (Lombard & Ditton, 1997; Biocca & Delaney, 1995) *Social Realism* refers to the extent to which a media/artificial environment is comparable to the real world; this is sometimes identified as behavioral realism (Freeman, 2004). Spatial presence and immersion are most relevant for the current study because the video games selected for this experiment include behaviors and characters not seen in the real world, and the authors felt the assessment of social realism would seem excessively artificial.

The concept of presence has been applied to a variety of media experiences, including video games (Eastin, 2006; Eastin & Griffiths, 2006; Ivory & Kalyanaraman, 2007) and television viewing (Bracken, 2005; Lombard & Ditton, 2000). The role of presence has also been explored in relation to other media effects (i.e., media enjoyment [Green, Brock, & Kaufman, 2004; Lombard & Ditton, 2000]; impact of violent content [Tamborini et al., 2004]). The findings are somewhat consistent across studies, with media users who reported higher levels of presence typically experiencing higher levels of the other dependent variables investigated (e.g., sensations of presence lead

to higher levels of enjoyment). Further, presence has been attributed to the technological form of a medium (Lombard & Ditton, 1997) and to media users' characteristics (Hecht & Reiner, 2007). The current study examines the influence of media form (image quality).

2.1 Presence and Image Quality

Previous research has demonstrated that form variables can influence the sensation of presence experienced by media users. The most relevant form variable to the current study is image quality. High image quality was originally hypothesized as a form variable that could leads to increases in presence experienced by media users (Lombard & Ditton, 1997). Recent studies have provided causal evidence that higher image quality lead to elevated levels of presence. Specifically, in an experiment exploring television viewers' presence responses to varying image quality levels, Bracken (2005) used HD and NTSC television images to manipulate image quality. The higher image quality provided by HD led to increased levels of various dimensions of presence, including the immersion, spatial presence, social realism, and social presence-passive (i.e., perceptions of facial expressions and characters' style of dress) dimensions of presence.

Further, in a study exploring the impact of image quality on participants' perceptions of newscast credibility (Bracken, 2006), image quality was manipulated using a local newscast viewed in either HD or NTSC. Significant differences were found for some presence dimensions, namely for immersion and social presence. The results also demonstrated that participants who watched the newscast in HD rated it as significantly more credible than those participants who watched the newscast in NTSC. These studies demonstrate that audience members can distinguish between HD and NTSC images, and that varying image quality has led to differing levels of presence. However, all of these studies were conducted with television content and this study seeks to explore the differences in image quality with video games.

2.2 Telepresence and Video Games

The release of the Xbox 360 in November of 2005 by Microsoft ushered in what has been dubbed the "HD Era" of gaming (Cross, 2005). The higher image quality of games on the Xbox 360 and Sony PS3 is one of the main distinguishing features of this new generation of game consoles. The huge investment by Sony and Microsoft into improving game image quality suggests the importance of graphics to gamers.

Research has shown that players strongly prefer more realistic graphics in video games (e.g., Wood, Griffiths, Chappell, & Davies, 2004), and HD games have the capability to produce better graphics than ever, with potential effects on perceived realism (Shapiro, Pena-Herborn, & Hancock, 2006) and other important outcomes of game exposure. This study examines the impact of high and standard definition images on presence.

There are relatively few studies focusing exclusively on presence and video games, and most incorporate presence as one of several dependent variables. An example of this type of inclusion is an experiment examining the use of story narratives in first-person shooter video games (Schneider, Lang, Shin, & Bradley, 2004). The authors manipulated the existence of the storylines in four different video games, with two having storylines and two not having them. The results were that participants reported feeling stronger presence sensations and identified more with the video game characters when there was a story in the game. However, this study did not find the significant relationship between presence and violent thoughts identified in earlier research,

One study with a primary focus on video games and experiencing a sense of presence was an exploratory study using the "autoconfrontation method." In this method, participants engage in an activity (e.g., playing a video game) and their performance is videotaped. After completion of the activity, they view the videotape of themselves along with the researcher. As they watch they are asked to comment on their experience and to rate a variety of presence dimensions (including immersion) as they viewed themselves playing the game. The participants reported feeling varying levels of presence, with higher levels of presence experienced by players who felt they performed well in the game (Rétaux, 2002).

The majority of the studies exploring presence and video games have investigated the relationship between video game playing, sensations of presence, and player aggression. In an experiment examining the relationship between video game exposure, presence, and hostile thoughts, Tamborini et al. (2004) compared participant responses to four gaming conditions: playing a virtual reality (VR) violent video game, playing a standard violent game, observing a violent game, or observing a nonviolent game. Game players in this study reported more telepresence (or "being there") than non-players. In addition, prior violent game experience was found to be a significant predictor of telepresence. Further support for the relationship between players'

experiencing presence and reporting more aggressive thoughts has been found in more recent experiments (Eastin, 2006; Farrar, Krcmar, & Nowak, 2006).

Lastly, the effect of gaming technology on presence and aggressive feelings was examined by Ivory and Kalyanaraman (2007). Gaming technology was manipulated by exposing participants to either PC games from the mid-1990s or PC games created from 2001 to 2003. The authors explained that the advances in gaming technology allowed the newer games to have higher visual and auditory quality resulting in a more vivid playing experience. The findings demonstrated that improvements in gaming technology do lead to stronger sensations of presence. However, this study did not find a significant relationship between presence and violent thoughts that were identified earlier.

Together these studies provide evidence that video games can evoke a sense of presence in video game players. The following hypotheses are posited:

> Hypothesis 1 Participants who play the HD version of a videogame will experience a higher level of *spatial presence* than those who play the NTSC version of the videogame**.**
>
> Hypothesis 2: Participants who play the HD version of a videogame will experience a higher level of *immersion* than those who play the NTSC version of the videogame**.**

3. Methods

In a between-subjects experiment, 50 participants played a video game in either HD (1080i lines, component video) or NTSC (480 lines, composite video). The independent variable was image quality (HDTV versus NTSC). The video game was played on a rear-projection television with a 65-inch wide (16:9) screen. Using random assignment, slightly more than half of the participants played the videogame in HD (26 players) and slightly less than half played the game in NTSC (24 players). Participants played the game alone and the image quality was switched after every second participant.

3.1 Participants

The 50 undergraduate students who participated in this experiment were between 18 and 50 years old (M = 23.02, SD = 5.14). The vast majority were between the ages of 18 and 29 (94%). The participants were equally divided by gender, with 25 females and

25 males. In terms of race, 74% of participants reported being “White,” 16% reported being “African-American,” 4% reported being “Pacific Islander,” and the remainder self identified as “Asian,” “Hispanic” or “Other” (6%).

3.2 Stimulus

All participants played the game *Perfect Dark Zero* on the Xbox 360 console system. In *Perfect Dark Zero*, players assume control of a futuristic spy who battles minions of the evil corporation dataDyne. To help in this struggle, the spy has access to an arsenal of deadly weapons, including a pistol and automatic rifle.

Perfect Dark Zero falls in the popular *first-person shooter* game genre, meaning that the action takes place through the lead character’s eyes. In line with predictions of presence scholars (Lombard & Ditton, 1997), the first-person point-of-view in this game was expected to increase the likelihood of presence being experienced. Players in this study began the game in a mountain environment where they were soon attacked by alien creatures.

3.3 Procedure

Each participant was met by the experimenter and provided with an informed consent form. Once the participant provided their consent they were escorted into a carpeted, 8 x 10 foot room that contained a television, a Microsoft Xbox 360, a videogame controller, and a comfortable chair that faced the television screen. Various other amenities, such as a decorative table lamp and pictures on the wall, made the environment similar to a living room. In both conditions, the chair was placed 6 feet from the front of the screen.

The experimenter explained that the participant would be playing a videogame and then completing a pencil-and-paper questionnaire. The experimenter then instructed the participant on how to use the controller and play the game. This was done first through a brief instructions sheet and then a short practice session. In the session, each participant was guided by the experimenter to the same point in the game. Depending upon prior gaming experience, this process took between 3 and 10 minutes. After the participant arrived at the designated point in the game the experimenter exited the room and the participant played the game alone for 10 minutes. After the allotted time the experimenter returned to the room and provided the questionnaire. The experimenter emphasized that they were no wrong answers and

that the participant should follow the directions in the questionnaire. The entire procedure took between 35-45 minutes.

3.4 Independent Measure

Image Quality. The image quality of the video game was manipulated with one group playing the game in HD (higher image quality) and the other playing the video game in NTSC (lower image quality).

3.5 Dependent Measures

Immersion. Participants responded from (1) to (7) for three items statements adapted from Lombard and Ditton (2000) to measure the extent to which media users feel a sense of being a part of the action or are connected when with media content. The items were : “How involving was the videogame”, “To what extent did you feel mentally immersed in the videogame environment”, and “I was so involved in the videogame environment that I lost track of time”. Cronbach’s alpha for the scale was .76.

Spatial Presence. Participants responded from (1) to (7) to three items adapted from Lombard and Ditton (2000) to measure the extent to which media users feel a sense of sharing a physical space within a mediated environment. The three questions were: “How much did it seem as if the objects and the people you saw/heard had come to the place you were”, “How much did it seem as if you could reach out and touch the objects or people you saw/heard”, and “How often when an object seemed to be headed toward you did you want to move out of its ways”. Cronbach’s alpha was .75.

4. Analysis and Results

A series of Independent Samples t-Tests were conducted with the independent variable image quality (HD versus NTSC) to test the hypotheses and research questions.

Hypothesis 1, which predicted that participants who played video games in HD would experience a higher level of spatial presence (being there in the video game) than those who played in NTSC, was not supported, $t(48) = .96$, $p < .34$, though the means were in the correct direction.

Support was found for Hypothesis 2, which predicted that participants who played video games in HD would report higher level of immersion than those who played the

video game in NTSC. The main effect was significant for immersion, $t(48) = 2.99$ $p < .004$), with those participants who played the video game in HD reporting higher levels of immersion ($M = 4.80$, $SD = 1.44$) than those who played the game in NTSC ($M = 3.65$, $SD = 1.24$).

5. Discussion

The results of this study provide initial evidence that image quality impacts both the level and types of telepresence dimensions experienced by video game players. The results strengthen the claim that image quality influences sensations of presence (Bracken, 2005; Lombard & Ditton, 1997). Further, the results support previous work examining video games and telepresence (Ivory & Kalyanaraman, 2007; Schneider et al, 2004; Tamborini et al, 2004; Tamborini & Skalski, 2006). In doing so, they add to the growing body of literature on video games, image quality, and telepresence and begin the process of synthesizing these important bodies of research.

Interestingly, the two telepresence sub-dimensions examined in this study were not affected the same by video game image quality. Spatial presence was not reported as significantly different for players but immersion was positively impacted by HD. Other studies have found that different stimuli increase different sub-dimensions of telepresence (e.g., Bracken, 2005).

There are several possible explanations for this finding. It may be that spatial presence is too big of a "leap" for game players to take without more advanced technology, such as virtual reality (VR). In other words, playing a video game on a HDTV may be enough to immerse players but not enough to make them feel "in" the space of a mediated environment, especially the type of environment used in this research (i.e., futuristic and non-photo realistic). Regardless, the mixed results observed in this study point to the importance of considering presence as a multi-dimensional construct in future research, as well as to the need to continue efforts at conceptualizing telepresence begun by others (e.g., Lombard & Ditton, 1997; ISPR, 2000; Lee, 2004).

The innovative nature of this experiment resulted in a few limitations that should be considered in future work. First, the interactive nature of video games should be controlled by including a viewing only condition.[3] This would allow for a manipulation of

[3] The authors originally designed the study to include this group, but the equipment necessary for recording HD was not avaible at the time of the data collection.

both vividness and interactivity, the two basic dimensions determining telepresence in Steuer's (1995) seminal work. Second, the skill level of the participants should be controlled, allowing a more direct comparison of the players' experiences.

6. Conclusion

The current study provides evidence that image quality in video games has an effect on participants' sensations of at least one dimension of presence. Future work should continue to address these relationships and the mediating/moderating role of presence dimensions on outcomes such as aggression and learning in response to advanced video game technology.

7. References

Biocca, F. (1997). The cyborg's dilemma: Progressive embodiment in virtual environments. *Journal of Computer-Mediated Communication, 3*, 1–29. Retrieved October, 2007, from http://jcmc.indiana.edu/vol3/issue2/biocca2.html

Biocca, F., & Delaney, B. (1995). Immersive virtual reality technology. In F. Biocca & M. R. Levy (eds.), *Communication in the age of virtual reality* (pp. 57-124). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bracken, C.C. (2006). Perceived source credibility of local television news: The impact of image quality and presence. *Journal of Broadcasting & Electronic Media, 50(4), 723-741*

Bracken, C.C. (2005). Presence and image quality: The case of high definition television*. Media Psychology, 7*(2), 191-205.

Cross, J. (2005). "HD Era" coming to gaming. *ExtremeTech*. Retrieved January 15, 2006 from: http://www.extremetech.com/article2/0,1558,1774523,00.asp?kc=ETRS S02129TX1K000532

Dupagne, M. & Seel, P. B. (2006). Digital television. In A. E. Grant & J. H. Meadows (Eds.), *Communication technology update.* New York: Focal Press.

Eastin, M.S. (2006). Video Game Violence and the Female Game Player: Self- and Opponent Gender Effects on Presence and Aggressive Thoughts. *Human Communication Research, 32*, 351-372.

Eastin, M.S., & Griffiths, R. P. (2006). Beyond the shooter game: Examining presence and hostile outcomes among male game players. *Communication Research, 33*, 448-466.

Entertainment Software Association. (2006). *Top ten industry facts*. Retrieved January 5, 2007 from http://www.theesa.com/facts/top_10_facts.php.

Farrar, K., Krcmar, M., & Nowak, K. (2006). Contextual Features of Violent Video Games, Mental Models, and Aggression. *Journal of Communication*, *56*(2), 387-405.

Freeman, J. (2004). Implications for the measurement of presence from convergent evidence on the structure of presence. Paper presented to the *Information Systems Division at the annual meeting of the International Communication Association*, New Orleans., LA.

Green, M. C., Brock, T. C., Kaufman, G. F. (2004). Understanding media enjoyment: The role of transportation into narrative worlds. *Communication Theory, 14*, 311-327.

Hecht, D., & Reiner, M. (2007). Field Dependency and the sense of object-presence in haptic virtual environment. *Cyberpsychology & Behavior, 10*, 243-251.

ISPR (2000). *The Concept of Presence: Explication Statement*. Retrieved January 12, 2007 from www.ispr.info.

Ivory, J.D., & Kalyanaraman, S. (2007). The Effects of Technological Advancement and Violent Content in Video Games on Players' Feelings of Presence, Involvement, Physiological Arousal, and Aggression. *Journal of Communication, 57*, 532-555

Kolodny, L. (2006). *Global video game market set to explode.* BusinessWeek. Retrieved January 5, 2007 from http://www.businessweek.com/innovate/content/jun2006/id20060623_163211.htm?chan=innovation_game+room_top+stories

Lee, K.M. (2004). Presence, Explicated. *Communication Theory,* 14, 27-50.

Lee, K. M. & Peng, W. (2006). What do we know about social and psychological effects of games?: A comprehensive review of current literature. In P. Vorderer & J. Bryant (Eds.), *Playing video games: Motives, responses, and consequences* (pp. 325-345). Mahwah, NJ: Lawrence Erlbaum Associates.

Lombard, M., & Ditton, T. B. (1997) At the heart of it all: The concept of presence. *Journal of Computer-Mediated Communication, 3*(2), Retrieved March 22, 2009 from http://jcmc.indiana.edu/vol3/issue2/lombard.html

Lombard, M., & Ditton, T. B. (2000). Measuring presence: A literature-based approach to the development of a standardized paper-and-pencil instrument. Presented at the *Third International Workshop on Presence, Delft, The Netherlands*. Retrieved March 22, 2009 from: http://www.matthewlombard.com/P2000.htm

Minsky, M. (1980). Telepresence. *Omni*, June, 45–51.

Rétaux, X. (2002-October). A subjective measure of presence feeling: The autoconfrontation method. Paper presented at *the Fifth Annual International Workshop on Presence*.

Schneider, E. F., Lang, A., Shin, M., Bradley, S. D. (2004). Death with a store: How story impacts emotional, motivational, and physiological responses to first-person shooter video games. *Human Communication Research,* 30, 361-375.

Shapiro, M. A., Pena-Herborn, J. & Hancock, J. T. (2006). Realism, imagination, and narrative in video games. In P. Vorderer & J. Bryant (Eds.), *Playing video games: Motives, responses, and consequences* (pp. 275-289). Mahwah, NJ: LEA.

Sherry, J. L. (2004). Media enjoyment and flow. *Communication Theory,* 14*(4)*.

Skalski, P. (2004-March) The quest for presence in video game entertainment. Presented as part of presence panel at the *Central States Communication Association Annual Conference*.

Steuer, J. (1995). Defining virtual reality: Dimensions determining telepresence. In F. Biocca & M. R. Levy (Eds.), *Communication in the age of virtual reality* (pp. 33-56). Hillsdale, NJ: LEA.

Tamborini, R., Eastin, M., Skalski, P., Lachlan, K., Fediuk, T. & Brady, R. (2004). Violent virtual video games. *Journal of Broadcasting and Electronic Media,* 48(3), 335-357.

Tamborini, R. & Skalski, P. (2006). The role of presence in the experience of electronic games.In Vorderer, P. & Bryant, J. (Eds.), *Playing video games: Motives, responses, and Consequences* (pp. 225-240). Mahwah, NJ: Lawrence Erlbaum Associates.

Wirth, W., Bocking, S., Hartmann, T., Klimmt, C., Schramm, H., & Vorderer, P. (2007). Presence as a process: Towards a unified theoretical model of formation of spatial presence experiences. *Media Psychology,* 9, 493-525.

Wood, R. T., Griffiths, M. D., Chappell, D. & Davies, M. N. (2004). The structural characteristics of video games: A psycho-structural analysis. *Cyberpsychology & Behavior,* 7(1), 1-10.

PsychNology Journal, 2009
Volume 7, Number 1, 113 – 131

What could abductive reasoning contribute to human computer interaction? A technology domestication view

Erkki Patokorpi[*♣]

[♣]IAMSR/Åbo Akademi
University
(Finland)

ABSTRACT

In recent decades, non-monotonous, informal patterns of reasoning have awakened a renewed interest among psychologists, economists and educationalists. Computer scientists and information systems professionals could also benefit from getting better acquainted with new research on how people think and act in the real world. The purpose of the paper is not to make an empirical contribution but to present a general argument in favour of a psychological approach to logic and its application to Human Computer Interaction (HCI), focusing especially on abduction. Abduction is a form of everyday reasoning that people typically use under uncertainty in a context. Abduction may help us better understand the epistemic conditions of advanced HCI – which increasingly takes place in authentic surroundings instead of in a laboratory-like setting – thus contributing to better research and design. HCI design should enhance our natural capacities and behaviour, which at the same time could mean creating new freedoms in the structures of everyday life.

Keywords: *abduction, practical reasoning, informal reasoning, logic of discovery, information systems methodology, human-computer interaction, technology design*

Paper Received 31/03/2009; received in revised form 28/04/2009; accepted 28/04/2009.

1. Introduction

Deductive arguments have traditionally been regarded as the soundest basis for reasoning, especially for reasoning in science. Recently a renewed interest in practical patterns of reasoning and people's behaviour in the real world has emerged in many disciplines, including pedagogy, the cognitive sciences, psychology and economics. Instead of a consuming preoccupation with the correctness of logical forms, which has

Cite as:

Patokorpi, E. (2009). What could abductive reasoning contribute to human computer interaction? A technology domestication view. *PsychNology Journal, 7*(1), 113 – 131. Retrieved [month] [day], [year], from www.psychnology.org.

[*] Corresponding Author:
Erkki Patokorpi
IAMSR, Åbo Akademi University, Joukahaisgatan 3-5a, 20520 Åbo, Finland
erkki.patokorpi@abo.fi

dominated the history of logic especially for the last hundred years or so, the issue of utility of logic is once more on the agenda. Assuming that Charles Sanders Peirce and later Wittgenstein are right about meaning being essentially a social and inferential phenomenon, a broader view on human knowledge calls for an examination of inferential practices.

Peirce's suggestion of abduction as a middle ground between induction and deduction will here be taken seriously. There are basically three research traditions on abductive logic. Two of these – abduction as hypothesis finding (and comparison) in the theory of science and abduction as logic programming – are briefly discussed but the focus will be on the third one, abduction as practical reasoning. Accordingly, the focus is on the actual abductive reasoning by people in real life.

Abduction as a form of everyday reasoning may be a central inferential mechanism at work when users act and interact with objects in an Information Society Technology (IST) context. Hence, abduction can be used for modelling what goes on "inside" the user's head. An advanced mobile computing situation especially calls for the use of abductive reasoning as the user typically is forced to come to a speedy conclusion on the spot in order to act in accordance with numerous contextual requirements of a real-life situation. From a Human Computer Interaction (HCI) design viewpoint, abduction as everyday reasoning is important because IST has to support natural social behaviour in order to become accepted by the majority of users (Abowd & Mynatt, 2000; Kleinrock, 2004; Grudin, 2002). The paper takes a stand in favour of emancipatory, domesticated technology, a kind of technology that allows the user to better control the tools he or she is using, and to comprehend the consequences of technology supported action to others and at least to the immediate environment (Keen & Mackintosh 2001; Punie, Bogdanowicz, Berg, Pauwels, & Burgelman, 2003; Patokorpi, 2006).

Pioneering work has been done for instance by Magnani and Bardone (2005a; 2005b; 2008) and Orliaguet (1999; 2000; 2001; 2002) but much more hands on deck are required to exploit the potential of abduction in the field of HCI. The sole purpose of this paper is to present a general argument in favour of applying reasoning, and especially abductive reasoning as a form of everyday, experiential, perception-based logic, to HCI. Accepting the argument means forming an alliance between logicians, psychologists and computer scientists that to some people may seem unholy (e.g. Popper, 1969).

First, abduction and its relation to the other two basic forms of logic will be explained, followed by a presentation of the three interpretational traditions of abduction. The relation of reasoning to proof and the combining of different inferential patterns in reasoned action will be illustrated in section 3. Abduction's role in discovery is then discussed, followed by a section on abduction as a potential tool for the technology domestication approach to HCI and a section on the image versus logic traditions. The last-mentioned section (section 6) tries to touch upon the parallel development of thinking styles and "thinking" machines.

2. Abduction

According to Peirce, there are only three elementary forms of logic: deduction, induction and abduction (CP 8.209 [CP refers to Peirce, 1934-63]; Hoffman, 1997; Rizzi, 2004). Peirce's canonical examples of the three basic inferential forms are the following:

Deduction
Rule: All the beans from this bag are white.
Case: These beans are from this bag.
Result: These beans are white.

Induction
Case: These beans are from this bag.
Result: These beans are white.
Rule: All the beans from this bag are white.

Abduction
Rule: All the beans from this bag are white.
Result: These beans are white.
Case: These beans are from this bag (CP 2.623).

The three elementary forms of logic can be seen as complementary operations of the human mind (Rizzi, 2004): Deduction infers a result (conclusion) that is certain; induction produces a rule (conclusion) that is valid until a contrary instance is found; abduction produces a case (conclusion) that is always uncertain (i.e. merely plausible).

In order to the scientific process of inquiry to become methodologically complete, abduction (whose job is to form hypotheses to explain an observation) needs to be followed by deduction (to logically derive the consequences of the hypothesis) and induction (to empirically test the predicted consequences of the hypothesis) (CP 6.469; CP 7.220; Pückler, u/d; Pape, u/d; Hoffmann, 1997; Flach, 1996).

The phenomenon of abductive reasoning has been discussed at some length in logic and rhetoric since Aristotle's Prior Analytics (2nd Book, Ch. 25; Gabbay & Woods, 2005). In the late 19th century, it was rediscovered by Peirce, whose interpretation and development of it has set the stage for virtually all subsequent research. There are three distinct interpretational traditions related to abduction, namely, abduction as a method of or model for:

1. scientific research or inquiry (logic of discovery)
2. machine reasoning (logic programming)
3. everyday reasoning (*logica utens*)

These three fields of application have their own, partly incompatible views on abduction.

The bulk of research into abduction has so far focused on its role in scientific research or inquiry (i.e. number 1). Ideally, an inductive research approach starts with gathering data by empirical observations free from prior ideas or preferences as to how the observations should be explained. A deductive approach in turn starts with explanations, hypotheses or theories. By drawing deductive inferences from a theory, its consequences in the real world can be predicted, provided that the theory is true. The predicted (or deduced) consequences in the real world can then be tested by empirical (inductive) methods (Danermark, Ekström, Jakobsen, & Karlsson, 2001). Deduction as a method of proof preserves truth, which means that if it starts from true premises, the logical form guarantees that the conclusion will be true. True premises cannot be arrived at by deduction, though. Induction, as a method of proof, is less truth-preserving, and as a method of arriving at true premises it is as impotent as deduction. Abduction's job is to produce hypotheses (explanations, guesses), and hypotheses are always merely plausible. Hence, abduction is the starting point of the self-correcting empirical research process. Punch, Tanner, Josephson and Smith (1990) have observed that frequently in accounts of scientific reasoning the nature of the hypothesis that could explain the findings is generally very indistinct: "What counts

as an explanation is not clear. It could involve accounting for (or covering) the findings to be explained, accounting with causal consistency, or maximal-plausibility coverage" (p. 38).

The role of abduction is, or should be, strong when the aim is to create something new. Secondly, the role of abduction is strong when there are not yet established theories, as abduction in tandem with induction is the means of arriving at new explanations and theories. And as was mentioned above, deduction's role is to draw the consequences of theories so that they can be put to test by induction (Kovács & Spens, 2005). As Peirce (CP 2.623; 6.469; 7.220) says, abduction describes what might be. It is thus connected to plausibility and oriented to the future (Patokorpi & Ahvenainen, 2009). Unlike deduction, it does not preserve truth.

The second perspective to abduction – abduction as a model for logic programming – has likewise interests and a research tradition of its own. Josephson and Josephson (1994) have modelled computing after the abductive inference model. In syllogistic terms, abduction is a *modus ponens* turned backwards, which in the eyes of formal logic makes it into a worst kind of textbook error in logic. Abduction is a logical fallacy because even if the premises were true (e.g. "All the beans from this bag are white" and "These beans are white"), the conclusion ("These beans are from this bag") may be false (i.e. these white beans could come from somewhere else than from this bag) (Wirth, 1993; Josephson & Josephson, 1994). As a rule, the algorithms based on abduction seem to be variations of the topsy-turvy *modus ponens*. Abduction has been successfully applied to computer systems that must work with incomplete knowledge. Abductive logic is regarded as capable of making computing machines think and act more like humans do (Satoh, Inoue, Iwanuma, & Sakama, 2000).

Both in the study of scientific inquiry and logic programming abduction is usually interpreted as Inference to the Best Explanation (IBE), that is, in terms of the so-called IBE model (see e.g. Lipton 1991). The IBE model deals with the generation and assessment of hypotheses, focusing on the formal-logical accuracy rather than the actual mental process of reasoning. In other words, it is concerned with comparing guesses (hypotheses) and not with what goes on in (and outside) someone's head when drawing the actual inferences (in scientific methodology these inferences have the role of hypotheses).

The third perspective sees abduction as a form of everyday reasoning or practical reasoning. In everyday reasoning there is no escaping the use of abduction because our knowledge in rapidly changing real-life contexts rests mostly on guessing, i.e. more

or less *ad hoc* hypotheses (Hoffmann, u/d). Abduction is especially suited for dealing with incomplete evidence under high uncertainty in complex real-life situations or ill-structured disciplinary fields of knowledge (e.g. medical diagnostics) (Spiro, Feltovich, Jacobson, & Coulson, 1988; Thagard, 1998; Lundberg, 2000). This may sound like a pretext for "anything goes," a recipe for anarchy. However, this is not to substitute truth with untruth but rather, as Spiro et al. put it: "the phenomena of ill-structured domains are best thought of as evincing *multiple truths*: single perspectives are not *false*, they are *inadequate*" (1988). Abduction is a practical pursuit that settles for conjecture because the search for an optimum, if not impossible, would, among other things, be too time-consuming and cognitively too demanding (Gabbay & Woods, 2005).

3. Reasoning does not equal evidence

One has to be able to say when a reasoning process is correct and when it is incorrect. Normative standards are necessary for given mental processes to count as logic (Fetzer, 1999). John Stuart Mill certainly was no stranger to the practical utility of reasoning, but he also had great concern about its correctness, which eventually led him to seek for greater certainty. He came to doubt Jeremy Bentham's facts-in-the-concrete and shifted focus in economic research from analogical (inductive) reasoning from experience to deductive (a priori) reasoning from assumptions to consequences. The latter attains greater certainty and is forward-looking, enabling prediction. In this type of a priori reasoning evidence is sharply separated from reasoning, and one starts from assumptions. Evidence enters the picture after reasoning as confirmation of predictions (Mill, 1961; De Marchi, 2002). Isolating forms of reasoning from one another and separating reasoning from evidence may give greater certainty but it is likely to steer attention away from the practical utility of logic.

In real life, forms of reasoning and evidence can be seen as essentially connected through reasoned action in the real world. Chiasson (2001) has described the use of different forms of reasoning in real life situations, demonstrating how different combinations affect our behaviour in the real world. Examples (adapted from Chiasson, 2001) of these inferential forms and their combinations are given below:

Simple abduction (guessing)

I see the dog coming into the house dripping wet. I focus on differences; and the wetness is a difference that draws my attention. What, is it raining? I give the matter no second thoughts and dry the dog.

Simple induction (guessing)

The dog comes into the house dripping wet. I focus on similarities, and the last time the dog was wet my wife was in the yard with the sprinklers on. If I make no further inquiry into the matter, I may jump to the conclusion that my wife is in the yard, taking the dog's wetness as "evidence" of it.

Gradual induction (possibly seeking evidence)

The dog comes into the house dripping wet. I focus on similarities. The last time the dog was wet my wife was in the yard with the sprinklers on but last Monday when the dog was wet it was raining, and two weeks ago the dog took a dive into the pond in the backyard. I may start looking for evidence that would corroborate one and falsify other alternatives. On the other hand, I may have no incentive to do so.

Deduction combined with gradual induction (seeking evidence)

The dog comes into the house dripping wet. By gradual induction I focus on similarities, remembering that there have been several occasions on which my dog got wet. By deduction I focus on consequences, and understand that the different reasons for the dog being wet have their consequences. For instance, if the dog has been hosed down by the neighbour because it had been messing up their flower bed, I may have to face an angry neighbour. So, by using gradual induction I proceed to check the neighbour's yard, the pond, the sprinkler, and so on, seeking evidence which would corroborate one of the explanations and falsify the rest.

Abduction combined with gradual induction (seeking evidence)

The dog comes into the house dripping wet. I use abduction, which means that I focus on differences, qualitative anomalies. I discover that there is a piece of plant in the dog's fur, and venture a guess that the dog has been in the pond. I check it. It is not the pond. Because abduction dominates my thinking, I hang on to the piece of plant, trying to find another explanation for it, combining abduction with gradual induction which may lead me to check my guesses. However, because deduction is missing I am more interested in raising new questions than reaching a definite conclusion. My investigation lacks a goal.

Abduction combined with gradual induction and deduction (guessing, inferring the consequences of the guess, putting the guess to test)

The last combination adds deduction, which makes my reasoning goal-oriented. Abduction, in turn, keeps my eyes peeled for new and unexpected facts or observations, thus guiding me in the finding of hypotheses, whereas gradual induction helps me to keep score of similar events. Gradual induction may lead me to look for evidence but deduction gives me an incentive to do so.

The above examples show that all three forms of reasoning are needed for reasonable action in the real world. The three forms of logic do not have to appear in the order presented in this particular example but may of course be combined in a number of ways. Reasoning does not equal evidence, but our inferential practices are irretrievably and in numerous ways linked with experience and evidence. Admittedly, abduction does not meet the standards of deductive validity, but as Tuzet (2004, p. 276) points out, abduction is often accused of being fallacious (logically invalid) when in fact the problem is epistemic, that is, there is not enough evidence to draw the conclusion.

4. To discover and to justify

There are historical reasons for undermining informal reasoning. One reason is a sharp separation of the context of discovery and the context of justification. A modern, influential advocate of this separation is Karl Popper (1969). According to Popper, matters related to the finding of something new should be studied in psychology, sociology and history, whereas matters related to the justification (proof or evidence) of findings belong to scientific method. Popper's view on scientific method does not recognize (epistemological) breaks in the growth of scientific knowledge as something rational. Scientific knowledge is supposed to build on previous knowledge by logical steps. Epistemological ruptures or scientific revolutions are thus things that do not belong to the logic of scientific inquiry but into historical or sociological studies of science (Bertilsson, 1978, pp. 10-14; Chauviré, 2005). For Karl Popper, logic is formal logic, and its job is to justify or prove hypotheses. If the premises of a deductive inference are true, the conclusion will also be true. If the conclusion of an inductive inference is corroborated by empirical evidence, the inference is probably true. We are justified in holding it to be (probably) true until a contrary event disproves it. This is presently the standard view on proper scientific procedure in terms of logic. An interesting consequence is that logic becomes separated from factual, experiential

evidence. Evidence is not the starting-point of reasoning but may be gathered to corroborate or weaken the implicit claims made of the real world. Because logic needs to be correct or immaculate for the scientific procedure to potentially produce truth, the correctness of logic guarantees the quality of scientific propositions. So far logic has mainly focused on the correctness or immaculateness of reasoning patterns rather than their usefulness or relevance.

For Peirce, abduction is a logic of discovery. Discovery is thus a rational process of constructing, finding and choosing hypotheses. Science, in turn, is “controlled creation” (Bertilsson, 1978, p. 76), based on abduction and confirmed by deduction and induction. Discovery and justification go more or less hand in hand. In highly formalized systems like mathematics the discovery process can be seen as an objective process circumscribed by the properties and relations of the signs of the system. To be objective means here that the (symbolic) process of discovery is virtually one with the real-life phenomena of mathematics. Reality (of mathematical things which are signs in a basically conventional formal system) and what we think of it coalesce. In ill-structured knowledge domains the objectivity of the discovery process is in turn questionable (Bertilsson, 1978, pp. 142-143).

The idea of a logic or method of discovery is not new. Greek geometricians built a conceptual model of inquiry, which they called analysis. Analysis is a heuristic method, a method of finding proofs. Abduction has a similar heuristic function as the so-called upward propositional interpretation of geometrical analysis and the analysis-of-figures interpretation of analysis that were used in Greek geometry (Niiniluoto, 1999; Hintikka & Remes, 1974; Patokorpi, 1996). Analysis was supposed to be a conscious and skilled process and therefore learnable. Three principal views on the analytic method exist. The analytic method is seen as (i) a subordinate part of the axiomatic (deductive) method, (ii) an alternative to the axiomatic method or (iii) a superordinate part whose subordinate part the axiomatic method is. The first one is related to the closed world view and the second and third to the open world view (Cellucci, 1998; 2005). The difference between a closed and an open world view corresponds to having either all information ready from the very beginning or making it up, or emerging, as one goes on. The important implication here is that abduction can be controlled and it suits especially for dealing with open systems.

5. Domesticating technology

Keen and Mackintosh (2001) and Punie, Bogdanowicz, Berg, Pauwels, and Burgelman (2003) present parallel views on technology, stressing the user's natural way of using technology in everyday life. By "natural" is meant the biologically and socio-historically conditioned behaviour of man as a tool-making, tool-using animal. Keen and Mackintosh (2001) borrow Ferdnand Braudel's maxim of technology as a means of creating freedoms in the structures of everyday life. Technology is thus seen as something which expands our natural ways of behaving in the world, and the most successful technologies are those that build on our natural interaction with the environment. Punie, Bogdanowicz, Berg, Pauwels, and Burgelman (2003) speak in favour of harnessing information technology to use by the man in the street; technology has to be domesticated. According to Punie, Bogdanowicz, Berg, Pauwels, and Burgelman (2003), human interaction with technology is a constant struggle in which technology changes us and we (as users) change technology. Technological artefacts are continuously modified, put to novel uses, and reinterpreted by the users. For example, the designers could not predict that the users would use the Short Messaging System (SMS) in the fashion they presently do. Technology in turn changes how we humans perceive, act and think. The SMS has for example changed how we make and keep appointments. Both technology and the socio-cultural aspects have to be taken into account in order to avoid both technological determinism and an oversimplified picture of user behaviour (Patokorpi, 2006). Abduction may help reaching both goals. Our abductive competence is also an essential factor in human creativity, and good design designs for change and user inventiveness (Robinson, 1993). Technology domestication is about maximizing the user's power and control over the artefact. Understanding user behaviour can hardly succeed well if our abductive processes are neglected.

An example of the study of everyday reasoning – although this one is not of abduction – is Gigerenzer and Hoffrage's (1995) empirical research on statistical inferences. They first carefully studied the actual reasoning processes the people used in the relevant context, then made the presentation of data more natural, which significantly improved both laymen's and professionals' estimates of probabilities and frequencies. The estimates became as good as the Bayesian ones when the data presentation suited the experimental subjects' natural way of dealing with frequencies. Unlike in the heuristics and biases programme of Tversky and Kahneman (Kahneman,

2003), the reasoning processes in the above example were not considered as more or less successful attempts, under constraints, at linear optimisation but adaptations to the environment. It is quite common in HCI that designers believe to know how the user should think: in accordance with unbounded rationality (e.g. deductive logic and Bayesian probability calculation). Thereof follows either a tendency to block avenues that do not meet the criteria of unbounded rationality – although they would be quite adequate for sensible use – or to make the inner workings of the artefact even less transparent than they already are because the users "would not understand them anyway".

6. Image and logic

Logic is usually understood in terms of order (i.e. syntax) and form. Hence meaning or content (semantics) is a matter of premises and completely isolated from logical form, which means that the existence of material inferences is not recognised. Material inferences are inferences in which the content of the concepts make them good inferences. For instance, we can infer from "John is Paul's father" to "Paul is John's son" because we understand the concepts 'father' and 'son'. On this view, grasping the conceptual content is in some sense prior to logical form, but logical (inferential) all the same (Brandom, 2000). Reasoning patterns as adaptations to the environment have a connection to experience and meaning making, but the connection itself has so far been little studied (studied though e.g. by Magnani, 2009). Abduction, as a form of perception-based reasoning, retains a connection to content or meaning because percepts make sense to us, that is, they have (iconic, indexical or/and symbolic) meaning. In other words, perception is inferential by nature. Does this holistic character of our experience set us apart from other complex systems, like machines? The issue is inherently linked to embodiment, disembodiment, situatedness and emergence which unfortunately due to limitations of space cannot be discussed here. Let us instead focus on representation (the representational side of 'experience') in humans and machines. Consider the episode in the history of physics narrated below

Peter Galison (1997, esp. p. 19) has studied two competing traditions of instrument making in physics: (i) the image tradition, based on *mimesis*, and (ii) the logic tradition, based on logical relations. The image tradition strives at representing natural processes in all their complexity through a single image. For instance, the existence of

particles is demonstrated through a picture of bubbles in superheated hydrogen. The logic tradition relies not on a single event but on a large amount of data, on the basis of which one can make statistical arguments for the existence of a particle. Hence, the image tradition relies on a concrete, homomorphic representation of a single occurrence, whereas the logic tradition relies on statistically derived logical relations between circumstances that give a homologous representation of the type of event in question. The homomorphic representation seeks to register a natural occurrence without interfering with it, whereas the homologous representation entails interfering with the phenomenon through complicated data massaging. Presently, says Galison, image and logic are converging:

> The new instruments (drift chambers and time projection chambers coupled to powerful computers) had begun to meld together the data-sorting capabilities of the logic tradition with the singular detail and inclusiveness of the image tradition. So it was that in the early 1980s the two traditions fused, with the production of electronically generated, computer-synthesized images. It was just as an electronic “photograph” that heralded the discovery of the W and Z particles in 1983 – the first time a single electronic detection of an event had ever been presented to the wider physics community an compelling evidence in and of itself (1997, p. 21).

The convergence is evident also for instance in computer science which has made significant progress in approximating image by the means of logic, as the research on automated image recognition indicates.

It has to be born in mind that in Galison’s story what are converging are homologous and homomorphic *representations*. The computer-synthesized outcome is or can be seen as an image, and thus at least in a derivative sense an individual, but it may have no material (concrete) connection to the individual event or object it depicts (or simulates) (see Turoff, 1997). Simulation can produce a gestalt, a whole which is more than the sum of its parts. In the extent that distributed computing (e.g. neural networks, multi-agent systems, swarm intelligence, distributed subsumption networks) can create complete, integral wholes computers can partly do the creative, imaginative work for us. According to Peirce (CP 2.777 CP 5.171), abduction is the only form of logic that creates something new. This creative or expansive capacity is partly due to abduction’s holistic and homomorphic nature. So far, we humans have outsmarted

machines in the sphere of the homomorphic (esp. abduction) whereas machines have long outperformed us in the sphere of the homologous, which traditionally has involved calculation by deduction. But now simulation and calculation seem to be coalescing. Does it mean that machines are closing upon us in areas which we regard as uniquely human?

As we humans are increasingly becoming cyborgs – meaning that our environment, social practices and bodies are merged with machines and artefacts – the line between human and nonhuman is eroding. However, our abductive competence is still among the toughest things to simulate by computers, mostly because abductive reasoning is focused on differences and conveys things as concrete individuals (as wholes). What will be our “natural” behaviour in the future remains to be seen. Presently, it seems reasonable to call such behavioural patterns natural which are firmly anchored especially in the most formative period of our human evolution, namely that of hunter-gatherers. Abduction certainly still is a central mechanism to us humans (as animals), which shows in the easy, rapid and mostly unconscious way we use it.

7. Conclusion

Abductive inferences are only plausible, and thus not, unlike deduction, truth-preserving. Abduction as everyday reasoning differs from the standard view of abduction as Inference to the Best Explanation, the latter being more focused on abduction’s role in scientific method. Incomplete information, time pressure and a changing environment compel to the use of abduction. In real life we use different forms of reasoning in tandem in order to take reasonable action. In order to become a sensible guide to reasonable or reasoned action abduction requires support from deduction and induction as well as a close connection to the seeking of evidence. Abduction is a logic of discovery, that is, a means of finding something new. Abductive reasoning takes a holistic view of facts or observations, focusing on details which lead to individuals (i.e. individuals as wholes). It is thus typically oriented towards homomorphic representations and analogue knowledge, which means that images (iconic signs) dominate abductive thinking. More precisely, abduction is typically oriented towards experiential gestalts of sounds, odours, feelings, tastes, and so forth. Lessons drawn from the study of abductive reasoning in real life should be readily

applicable especially in interface design. Logic programming in terms of everyday abduction instead of the IBE model is a tougher challenge.

Firstly, abduction is a form of logic and therefore a sharp, analytic tool that should meet the rigorous requirements of technologically oriented HCI researchers. Abduction is also a form of qualitative, everyday reasoning, connected to our bodily interaction with the world. Therefore it should be rich and broad enough for socio-culturally oriented HCI researchers (Patokorpi, 2006). Secondly, because the creation and sharing of knowledge is in many ways anchored in real-world environments and social practices, the study of knowledge practices ought also to pay attention to reasoning patterns as specialized ways of an organism to adapt to the environment (see e.g. Gigerenzer, 2008; Gigerenzer, Hoffrage & Goldstein, 2008). In addition to abductive practices, deductive, inductive and other reasoning practices need to be re-examined (e.g. Chiasson u/d; 2001). There is a wealth of valuable insights into reasoning, classification, perception and so forth, by writers like for instance Gigerenzer (e.g. 2008), Rosch (1975), Landa & Ghiselin (1999), Magnani (e.g. 2004) and Orliaguet (e.g. 2000) that are yet largely untapped by the HCI community.

Finally, insofar as our inherited forms of reasoning and behaviour in general help us to interact with new technology as well as better understand the impact of our technology enhanced action to others and the environment, it makes sense to exploit knowledge of our natural ways of interaction when designing technology. Although the fairly recent developments in the history of logic (e.g. Boole's laws of thought) coupled with the invention and diffusion of computing machines effectively divorced logic from psychology, it certainly is legitimate to study logic in terms of adaptations to the environment (ecological rationality) and in terms of inferential practices that cannot be reduced to their pure logical form (inferentialism). It is legitimate because without a connection to the environment there would be neither reasoning nor rationality. Psychology has here a great deal to contribute to logic and the kind of practically oriented logical study advocated in this paper has a great deal to contribute to human technology interaction.

8. References

Abowd, G.D. & Mynatt E.D. (2000). Charting past, present, and future research in ubiquitous computing. *ACM Transactions on Computer-Human Interaction*, *7*(1), 29-58.

Aristotle (1973). *Prior Analytics*. With an English translation by Hugh Tredennick. Cambridge, MA: William Heinemann.

Bertilsson, M. (1978). *Towards a Social Reconstruction of Science Theory. Peirce's Theory of Inquiry and Beyond*. Theses. Reprocentralen Lunds Universitet.

Brandom, R.B. (2000). *Articulating Reasons. An introduction to inferentialism*. Cambridge, MA: Harvard University Press.

Cellucci, C. (1998). The scope of logic: deduction, abduction, analogy. *Theoria* Retrieved on September, 30 2008 from: http://bacheca.lett.unisi.it/duccio/files/ScopeofLogic.pdf.

Cellucci, C. (2005). Mathematical Discourse vs. Mathematical Intuition. In C. Cellucci & D. Gillies (eds.), *Mathematical Reasoning and Heuristics* (pp. 137-165). London: King's College Publications.

Chauviré, C. (2005). Peirce, Popper, abduction and the idea of a logic of discovery. *Semiotica*, *153*, (¼), 209-221.

Chiasson, P. (2001). "Logica utens". In J. Queiros*, Digital Encyclopedia of Charles S. Peirce*. Retrieved on September, 30 2008 from: http://www.digitalpeirce.fee.unicamp.br/p-logchi.htm.

Chiasson, P. (Undated). *The role of optimism in abduction*. Retrieved on April, 26 2007 from: http://www.digitalpeirce.fee.unicamp.br/p-rolchi.htm.

Danermark, B., Ekström, M., Jakobsen, L. & J.C. Karlsson (2001). *Explaining Society. Critical realism in the social sciences*. London: Routledge.

De Marchi, N. (2002). Putting evidence in its place: John Mill's early struggles with "facts in the concrete." In U. Mäki, (ed.), *Fact and Fiction in Economics: Models, Realism and Social Construction* (pp. 304-326). West Nyack, NY: Cambridge University Press.

Fetzer, J. (1999). Mental Models: Reasoning without Rules. *Minds and Machines, 9*,(1), 119-125.

Flach, P.A. (1996-August). Abduction and induction: syllogistic and inferential perspectives. Presented at *the ECAI'96 workshop on Abductive and Inductive Reasoning*.

Gabbay, D. M., & Woods, J. (2005). Advice on Abductive Logic. *Logic Journal of the IGPL, 14*(2), 189–219.

Galison, P. (1997). *Image and logic: A material culture of microphysics*. Chicago: The University of Chicago Press.

Gigerenzer, G. (2008). Why Heuristics Work. *Perspectives on Psychological Science, 3*(1), 20-29.

Gigerenzer, G. & U. Hoffrage (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102*, 684-704.

Gigerenzer, G., U. Hoffrage & D.G. Goldstein (2008). Fast and Frugal Heuristics Are Plausible Models of Cognition: Reply to Dougherty, Franco-Watkins, and Thomas (2008). *Psychological Review*, *115*(1), 230-239.

Grudin, J. (2002). Group Dynamics and Ubiquitous Computing. *Communications of the ACM*, *45,(*12), pp. 74-78.

Hintikka J. & Remes, U. (1974). *The Method of Analysis: Its Geometrical Origin and Its General Significance*. Dordrecht: Reidel.

Hoffmann, M. (1997-July). Is there a 'Logic' of Abduction? Presented at the *6th Congress of the IASS-AIS*.

Hoffmann, M. (Undated). Problems with Peirce's Concept of Abduction. Retrieved on September, 30 2008 from: http://user.uni-frankfurt.de/~wirth/texte/hoffmann.html.

Josephson, J.R. & S.G. Josephson (eds.) (1994). *Abductive Inference. Computation, Philosophy, Technology*. Cambridge: Cambridge University Press.

Kahneman, D. (2003). A Perspective on Judgment and Choice; Mapping Bounded Rationality. *American Psychologist, 58*(9), 697-720.

Keen, P.G.W. & R. Mackintosh (2001). *The Freedom Economy: Gaining the MCommerce Edge in the Era of the Wireless Internet*. New York: Osborne/McGraw-Hill.

Kleinrock, L. (2004). The Internet rules of engagement: then and now. *Technology in Society, 26*(2-3), 193-207.

Kovács, G. & K.M. Spens (2005). Abductive reasoning in logistics research. *International Journal of Physical Distribution & Logistics Management*, *35*(2), 132-144.

Landa, J.T.& M.T. Ghiselin. (1999). The emerging discipline of bioeconomics: aims and scope of the Journal of Bioeconomics. *Journal of Bioeconomics, 1*, 5-12.

Lipton, P. (1991). *Inference to the Best Explanation*. London and New York: Routledge

Lundberg, C.G. (2000). Made sense and remembered sense: Sensemaking through abduction. *Journal of Economic Psychology, 21*(6), 691-709.

Magnani, L. (2004). Reasoning through doing. Epistemic mediators in scientific discovery, *Journal of Applied Logic*, *2*(4), 439-450.

Magnani, L. (2009). Abducing chances in hybrid humans as decision makers. *Information Sciences, 179*(11),1628-1638.

Magnani, L. & E. Bardone (2005 - July). Abduction and HCI. A cognitive model for evaluating and designing human interfaces. Presentation at *HCI International 2005*.

Magnani, L. & E. Bardone (2005b). Designing human interfaces. The role of Abduction in: L. Magnani and R. Dossena (eds.), *Computing, Philosophy, and Cognition*, *Proceedings of the conference E-CAP2004* (pp. 131-146). London, College Publications.

Magnani, L. & Bardone, E. (2008). Sharing representations and creating chances through cognitive niche construction. The role of affordances and abduction, in: S. Iwata, Y. Oshawa, S. Tsumoto, N. Zhong, Y. Shi and L. Magnani (eds.). *Communications and Discoveries from Multidisciplinary Data* (pp. 3-40), Berlin: Springer.

Mill, J. S. (1961). *A system of logic, ratiocinative and inductive: Being a connected view of the principles of evidence and the methods of scientific investigation*. White Plains, NY: Longman.

Niiniluoto, I. (1999). Defending Abduction. *Philosophy of Science*, *66*, 436-451.

Orliaguet, J-M. (1999). *A semiotic perspective on digital (r)evolution. From UNIX to the desktop.* Retrieved on October, 23 2004 from: http://www.led.br/~tissiani/arquivos/ePapers/papers_VRUI/chalmersMediaLab/unix_semiotics.pdf.

Orliaguet J-M. (2000). *How Do We Reason when Using Computers? How Programmable Are We?* Retrieved on February, 2 2006, from: http://www.ckk.chalmers.se/people/jmo/essays/how_do_we_reason.pdf.

Orliaguet, J-M. (2001). *Design, Virtual Reality and Peircean Phenomenology*. Retrieved on March, 27 2007 from: http://scholar.google.no/scholar?hl=no&lr=&q=cache:fBL2ZowKT3UJ:www.medialab.chalmers.se/people/jmo/semiotics/hci2001-JM_Orliaguet.pdf+Orliaguet+design.

Orliaguet, J-M. (2002). *Prolegomenon to a Semiotic of Digital Media*. Retrieved on October, 23 2004 from: www.ckk.chalmers.se/people/jmo/semiotics/semiotic_of_digital_media.pdf.

Pape, H. (Undated). *Abduction and the Topology of Human Cognition*. Retrieved on March, 3 2005 from: http://user.uni-frankfurt.de/~wirth/texte/pape.html.

Patokorpi, E. (1996). *Rhetoric, Argumentative and Divine*. Frankfurt am Main: Peter Lang Verlag.

Patokorpi, E. (2006). *Role of Abductive Reasoning in Digital Interaction*. Åbo: Åbo Akademi University Press. Retrieved on September, 22 2008 from: http://www.cspeirce.com/menu/library/aboutcsp/patokorpi/abduction.pdf.

Patokorpi, E. & M. Ahvenainen (2009). Developing an Abduction-Based Method for Futures Research. *Futures, 41*(3) 126-139.

Peirce, C.S. (1934–63). *Collected Papers of Charles Sanders Peirce*, *Vols. 1–7*. Cambridge, MA: Belknap Press of Harvard University.

Popper, K. (1969). *Logik der Forschung*. Tübingen: J.C.B. Mohr (Paul Siebeck).

Punch, W.F., M.C. Tanner, J.R. Josephson & J.W. Smith (1990). PEIRCE: a tool for experimenting with abduction. *IEEE Expert*, *5* (5), 34-44.

Punie, Y., Bogdanowicz, M., Berg, A-J., Pauwels, C. & J-C. Burgelman (2003). *Living and Working in the Information Society: Quality of Life in a Digital World.* A Final Deliverable of the European Media Technology and Everyday Life Network (EMTEL).

Pückler, von T. (Undated). *Peirce und Popper über Hypothesen und ihre Bildung*, Retrieved on September, 9 2006 from: http://user.uni-frankfurt.de/~wirth/texte/P%FCckler.html.

Rizzi, A. (2004). Abduzione ed inferenza statistica. *Statistica e Società, 2*(2),15-25.

Robinson, M. (1993). Design for unanticipated use. In G. De Michelis, C. Simone and K. Schmidt (eds.), *Proceedings of the Third European Conference on Computer-Supported Cooperative Work* (pp. 187-202). Netherlands: Kluwer,.

Rosch, E. (1975). Cognitive Representations of Semantic Categories. *Journal of Experimental Psychology (General)*, *104* (3), 192-233.

Satoh, K., Inoue, K., Iwanuma, K. & C. Sakama (2000-September). Speculative Computation by Abduction under Incomplete Communication Environments. Paper presented at *the Fourth International Conference on Multi-Agent Systems*.

Spiro, R.J., Feltovich, P.J., Jacobson, M.J. & R.L. Coulson (1988). *Cognitive Flexibility, Constructivism, and Hypertext: Random Access Instruction for Advanced Knowledge Acquisition in Ill-Structured Domains.* Retrieved on August, 6 2004 from: http://phoenix.sce.fct.unl.pt/simposio/Rand_Spiro.htm.

Thagard, P. (1998). Explaining Disease: Correlations, Causes, and Mechanisms. *Minds and Machines*, *8*, 61-78.

Turoff, M. (1997). Virtuality. *Communications of the ACM, 40*(9), 38-43.

Tuzet, G. (2004). Le prove dell'abduzione. *Diritto e Questioni Pubbliche*, *4*, 275-295. Retrieved on June, 11 2008 from: http://www.dirittoequestionipubbliche.org/page/2004_n4/studi_G_Tuzet.pdf.

Wirth, U. (1993). Die 'Abduktive Wende' der Linguistik, *Kodikas/Code*, *16*, 289-301.