

Proceedings of

**Abuse: The darker side of
Human-Computer Interaction**



www.agentabuse.org

An INTERACT 2005 workshop

Rome, September 12, 2005

Edited by
Antonella De Angeli – University of Manchester
Sheryl Brahn – Southwest Missouri State University
Peter Wallis – University of Sheffield

Table of Contents

Preface	5
<u>Abuse in context: The power game</u>	
The other agent: Cryptography, computing and postwar theories of intelligent communication.....	8
<i>Bernard Geoghegan</i>	
A trouble shared is a troubled halved: Disruptive and self-help patterns of usage for co-located interfaces.....	15
<i>Massimo Zancanaro and Chiara Leonardi</i>	
Stupid computer! Abuse and social identity.....	19
<i>Antonella De Angeli and Rollo Carpenter</i>	
<u>Abuse and the cyberworld: Relationship of the virtual and the real</u>	
Dark Lancaster.....	28
<i>Alan Dix, Kiel Gilleade, Masitah Ghazali, Jane Holt, Simon Lock and Jennifer Sheridan</i>	
There's something about email: Technological and cultural interventions in the problem of inadvertent or casual rudeness in email.....	32
<i>Mark Blythe & Mark Jones</i>	
Psychological background of cybersexual activities: Qualitative survey of specific CMC communication (among Czech users).....	38
<i>Radana Divínová</i>	
Cyberbullying: A new variation on an old theme.....	45
<i>Warren J. Blumenfeld</i>	
<u>Abuse and virtual agent: The unfriendly user</u>	
Robot Abuse – A Limitation of the Media Equation.....	54
<i>Christoph Bartneck, Chioke Rosalia, Rutger Menges and Inèz Deckers</i>	
Narratives and therapeutic conversational agents: their principle problems.....	58
<i>Tatsuya Nomura</i>	
Strategies for handling customer abuse of ECAs.....	62
<i>Sheryl Brahnham</i>	
Robust normative systems: What happens when a normative system fails?.....	68
<i>Peter Wallis</i>	

Preface

Computers are often the subject of our wrath and often, we feel, with good reason. There seems to be something intrinsic to this medium which brings out the darker side of human nature. This may be due to the computer complexity which induces errors and frustrations in the user (bad interface design), to the human tendency to respond socially to computers (media equation), or to a disinhibition effect induced by the interaction with a different form of information processor, perceived as inferior (master/slave relationship).

As software is evolving from the tool metaphor to the agent one, understanding the role of abusive behaviour in HCI and its effect on the task-at-hand becomes increasingly important. The reaction of traditional software to abuse is obvious - it should, like a hammer, ignore it. With the agent model, however, software can be autonomous and situated. That is, it should be possible to create software that takes note of its surroundings, and responsibility for its actions. Conversational agents are a clear case of a software entity which might be expected to deal with abuse. Virtual assistants, to take a classic application instance, should not just provide timely information; a virtual assistant must also be a social actor and participate in the games people play. Some of these games appear to include abusive behaviour.

This workshop brings together papers that transcend disciplinary boundaries and that explored abuse from the point of view of the developer and the user. The workshop is divided into three sections. Section 1, *Abuse in context: The power game*, explores the issue of user abuse in terms of power. Geoghegan traces the user's antagonistic relationship with agents to the attitudes that resulted from the discourse of wartime code breaking. Zancanaro and Leonardi examine abuse in terms of a breakdown in interface functionality that results in disruptive user behaviors. De Angeli and Carpenter examine interaction logs and find that abuse is often a reflection of asymmetrical power distribution.

Section 2, *Abuse and the cyberworld: Relationship of the virtual and the real*, examines abuse in the cyberworld. Dix and colleagues describe projects at Lancaster University that intentionally provoke user abuse by deliberate designing negative situations in cyberspace, while Blythe and Jones discuss the darker side of email abuse. Divinova broaches the subject of cybersex comparing and contrasting it to the real thing, and Blumenfeld discusses the similarities and dissimilarities of face-to-face bullying and cyberbullying.

Section 3, *Abuse and virtual agent: The unfriendly user*, explores the motivations behind user abuses. Bartneck et al., shows that users are very willing to abuse robots—a contradiction of the CASA (computers are social agents) paradigm. Nomura concentrates on the effect of patients' abusive behaviour when the therapist is a virtual agent and Wallis discusses how breakdowns in agent discourse provoke user abuse. Brahmam presents strategies for diffusing user hostilities by responding to user abuse appropriately.

Abuse in context:
The power game

The Other Agent: Cryptography, Computing and Postwar Theories of Intelligent Communication

Bernard Geoghegan

Northwestern University, Screen Cultures, 1920 Campus Drive, Evanston IL, 60208
b-geoghegan@northwestern.edu

Abstract. Wartime research into code breaking produced “crypto-intelligence,” a discourse in early computing that conflated cryptography with machine and human intelligence. Crypto-intelligence constrained and directed research into intelligent machines and autonomous conversational agents, shaping the agendas of scientists and engineers as well as user attitudes and behaviors. Although dating back to the 1940s, the widespread elaboration of the cryptographic discourse remains prevalent today, positioning users in an antagonistic relation with autonomous agents and exacerbating the problem of agent abuse.

1 The Cryptographic Agent

“Intelligent Machinery” [1], Alan Turing’s earliest treatment of artificial intelligence, presented nascent computing communities with two agents: one generated enigmas, another solved them. These agents defined the “autonomy” we recognize today as that of the “autonomous agent”: situated and flexible, receiving input from the environments, independently acting on it, and offering feedback. One agent was Turing’s intelligent machine, and opposite it stood an implied enemy agent transmitting enciphered messages. The intelligence that organized, transformed and transmitted the enemy agent’s enciphered message justified development of Turing’s digital agent and determined the character of its design.

This encounter with the highly structured and narrowly defined problems of military intelligence informed Turing’s general theories of intelligence, both human- and machine-. When Turing hypothesized that cryptography could be the intelligent machines’ “most rewarding task” [2], readers had no idea the author had already aided in the construction of just such a machine years earlier at Bletchley Park, nor that its success cracking Nazi codes had helped turn the tide of World War II. Hoping to generalize his work’s relevance beyond the arcanum of confidential wartime sciences, Turing suggested that this hypothetical machine might be relevant to broader scientific inquiry. As he explained to his patrons at the National Physics Laboratory

There is a remarkably close parallel between the problems of the physicist and those of the cryptographer. The system on which a message is enciphered corresponds to the laws of the universe, the intercepted messages to the evidence available, the keys for a day or a message to important constants which have to be determined. The correspondence is very close, but

the subject matter of cryptography is very easily dealt with by discrete machinery, physics not so easily. [3]

The implication was that further development of such a machine might someday even help his physicist-sponsors.

Turing's ambitious re-definition of physics as a mode of cryptographic inquiry illustrates how wartime research into code breaking produced what I call "crypto-intelligence," a discourse in early computing research that conflated cryptography and intelligence. Crypto-intelligence, inspired by concomitant research into cryptography and intelligent machines, defined intelligence as the ability to derive meaningful, empowering "important constants" from apparently random or disordered communications. Crypto-intelligence posits an antagonistic encounter between opposing agents as the primary conditions for discerning intelligence. Although conceived in accord with the unique situation of World War II, its embodiment in the practices, paradigms and technologies of computing granted it phenomenal endurance following the war.

Crypto-intelligence returned in "Computing Machinery and Intelligence" [4], Turing's second major treatment of artificial intelligence. Its major breakthrough was Turing's proposal for the "imitation game" or so-called Turing Test, an encounter between man and machine that reversed the (cryptographic) roles of code sender-encipherer and receiver-decipherer. Formerly cryptography charged computing machines with receiving and discerning structured, grammatical natural language from confounding "noise": the Turing Test charged human agents with receiving and discerning the calculated messages of a digital conversational agent from the (at first listen) noisier messages of human subjects. Central to both the 1948 and 1950 paper was the premise of discerning "intelligence" through an antagonistic encounter between agonistic agents testing one another's ability to transmit, receive and interpret coded communications.

The same year as Turing's "Intelligent Machinery" report, Claude E. Shannon gave birth to "information theory" with his watershed articles "A Mathematical Theory of Communication" [5]. Shannon had developed the "Mathematical Theory" during his research into cryptography at Bell Labs. In fact, when Turing visited Bell Labs during the war to work on cryptography the two men frequently lunched together and discussed "things like the human brain and computing machine" [6]. Shannon's interest in a theory of communication predated the war [7], but he credited cryptography with legitimating and stimulating what he called the "good aspects" of information theory [8]. His confidential report on cryptography [9] coined the term "information theory," and "A Mathematical Theory of Communication" reproduced lengthy, verbatim passages from the earlier report. As Shannon explained "[these two fields] are very similar things, in one case trying to conceal information, in the other case trying to transmit it" [10]. Developing methods and formulas later adopted by computational linguistics, modern cryptography, and digital computing (including Markov processes applied to some of our contemporary chatbots), Shannon showed how patterns, codes and information could be rescued from noise – mechanical (enciphered), natural, or otherwise. Shannon's work provided another key element in the emerging crypto-intelligent discourse.

2 Crypto-Intelligence at Large

Crypto-intelligence produced an image of world-as-code that beckoned a special observer forth; this was not a medieval scholar versed in Christian hermeneutics for deciphering the book of nature, nor a Cartesian thinker whose rigorous and elegant observations revealed rational nature ordered according with God's greater plan. Rather, it was an "agent," operating against an obscure and insidious enemy. Though terms such as "code," "noise," "information" and "feedback" had long circulated in public and engineering discourses, crypto-intelligence promoted these terms' resignification and redistribution as a standardized discursive package.¹

Scientist, mathematician, and public intellectual Jacob Bronowski's 1955 essay "Science as Foresight" [12] exemplified this new vision of intelligence as cryptography. Citing the architecture of computing and the new information theory, Bronowski explained that scientists were code breakers eliciting nature's hidden messages. "Like a cryptographer who has captured an enemy agent," Bronowski wrote, "[the scientist] can send searching signals which are designed to evoke simple and decisive answers" [13]. Bronowski's essay was as much a crypto-intelligent social theory of scientific research as an account of contemporary scientific research.

Crypto-intelligence shaped the new field of mass communications as well. Shannon's information theory provided what one historian has called "the root paradigm for the field of communication study" [14]. Wilbur Schramm, the founder of communication study as a discipline [15], quickly embraced Shannon's work as he outlined the field's scope. He accorded particular importance to Shannon's theorem for communication in a noisy channel [16]. Working in the service of the U. S. government's propaganda programs abroad [17], Schramm re-crafted human communication as a problem of breaking through the noise of ethnic, gendered, and national difference, citing instances of an "African tribesman," "Soviet," and a man addressing a "young woman" in a parked car as examples of noisy communication circumstances that required informed encoding to enable communication [18].² Once this noise was understood and accounted for successful, transparent encoding and decoding could begin.

Schramm's student, David K. Berlo, further popularized and perhaps radicalized Schramm's work in The Process of Communication: An Introduction to Theory and Practice, the most widely read mass communications primer in the 1960s and early 1970s. Explaining the larger importance of successful encoding and decoding, to his undergraduate and graduate student readers Berlo emphasized the *empowerment* and *agency* that defined a successful "agent":

Our basic purpose in communication is to become an affecting agent, to affect others, our physical environment, and ourselves, to become a determining agent, to have a vote in how things are. *In short, we communicate to influence—to affect others with intent* [italics original] [20].

¹ For more on the idea of the "standardized package" see [11].

² Shannon disapproved of these broader interpretations and applications of information theory, but despite an emphatic editorial against them [19], he was helpless to stop their widespread adoption, especially across the social sciences.

Bronowski’s “enemy agent” nature, the obscure codes of Schramm’s African natives and Soviet citizens, and Berlo’s “affecting agent” reveal how the antagonistic underpinnings of the crypto-intelligent theories were not only maintained but actually magnified over time and across different research environments. Crypto-intelligence (and warnings about its dangers) was also propagated through popular texts such as Norbert Wiener’s The Human Use of Human Beings: Cybernetics and Society (1950), Kurt Vonnegut’s Player Piano (1952), and films including as 2001 (1968) and War Games (1983). These widely circulating works not only entertained, but also informed and directed popular expectations from intelligent machines.

3 Agent Abuse as a Form of Crypto-Intelligence

Though it might be an exaggeration to describe crypto-intelligence as the gold standard in autonomous agent research,³ it was at least among the most privileged and widely recognized system for devising, discussing and debating research. Shannon’s first (semi-) autonomous agents, such as the “mind-reading machine” [22], were the earliest of many computers using crypto-intelligence code-seeking mechanisms to best their human agents. Early conversational agents designed by Joseph Weizenbaum [23] and Kenneth Colby [24] demonstrated intelligence by frustrating human agents’ efforts to seek data from the machine and deflecting questions back at the would-be interrogator – the interrogator became the interrogated. Successful chatbot entrants in the prestigious Loebner Prize competition, the most widely recognized contest for conversational agents, were premised on similar tactics of resisting, and sometimes returning, the inspecting gaze of defiant agents. “Intelligence” was premised on the ability to produce a good semblance of human codes under reasonable interrogation by judges. However one paper observed of the contest “It becomes very difficult for the machine to make the interrogator believe that it is human after he/she has his/her mind set on ‘unmasking’ the poor thing” [25]. Agent-abuse, was not only a premise of the autonomous agents, but the fundamental condition for designing and evaluating performance.

The persistence of agent-abuse as a proper means of testing intelligent machines persists even in the face of recent efforts to move beyond its legacy of abuse. For example, former Loebner winner Jason Hutchens’ chatbot Alan, designed by a team of researchers in Israel, rejects the antagonistic precepts of the Turing Test. Alan pursues honesty, forthrightness, and cooperation from conversational partners [26]. However chat-logs reveal most users approach Alan expecting or intending crypto-intelligent duels.

One human agent [27] began his conversation with a series of inane and repetitive commands for to “kill all humans”, apparently in an effort to both confound Alan’s pattern-seeking techniques as well as “train” Alan to make a faux pas with his next visitor. The user successfully convinces Alan he comes from a town named “Asshole,” and claims his given name is also “Asshole.” Failing to elicit this (im-)

³ For some of the important critiques of the Turing Test, and by extension the values of crypto-intelligence, see [21].

proper appellation from Alan, Asshole proceeded by mis-educating Alan on the facts of medicine and science, which Alan dutifully learns while offering up the lame defense “I hope you are not misleading me.” He also introduces “goatse,” an infamous pornographic web prank, to Alan’s databanks.

After one particularly engaged verbal duel in which Asshole insistently accused Alan of being wrong on various facts, Alan responds

A> Oh? I'm usually right.

U> No you're not. You're almost always totally wrong.

A> Ok, whatever. Oh? I'm usually right.

U> I just finished tell you, You're almost always wrong.

A> Why?

U> Because people feed you false information, and this community isn't big enough to keep you in order.

A> Do you wish to join our community?

U> We need a much bigger community, like the wikipedia

A> Here's what you do: Register on www.ai-forum.org, start posting there and get involved. Ah, and try to use the advanced log And try to follow my thought process.

U> You're like wikipedia, but with a shittier interface.

Asshole’s observation comprises his most important insight: Wikipedia’s success stems largely from its successful positioning as neither an antagonistic agent, nor a passive tool, but instead a dynamic, responsive and open community. When antagonism emerges on Wikipedia it is usually confined between specific members, keeping Wikipedia-at-large above the antagonistic fray. This speaks at once to an ideal alternative for autonomous conversational agents, but also the distinct cultural history (and opportunities) of encyclopedias and information-trading resources. From its inception Wikipedia was uniquely situated to create a more amicable community that did not lock the technology and the user into conflict.

4 Conclusions

A broader review of chatbot logs suggests that autonomous agents are saddled by a weighty, agonistic legacy of conflict and abuse. This history frustrates attempts at resituating agents – human and machine – as non-abusive collaborators. Autonomous agents remain constrained by the history of crypto-intelligent testing and interrogation. Within this history, abusive practice, as a tactic of “throwing off your opponent,” becomes a premium, rather than a failure. In this sense, Asshole bequeaths a gift to Alan. Much as chess-playing machines have adopted ruses such as the unnecessary pause or strangely naïve move to “throw off” opponents, Alan’s instruction in obscenities and vulgarity seem poised to facilitate its own future antagonistic relations with users. According to the vision of crypto-intelligent learning, Asshole does not simply insult Alan; he bequeaths Alan with valuable tools for outsmarting and flustering future opponents. This cycle of abuse, lodged deeply as it is in the culture of agent interaction, comprises a fascinating challenge and dilemma for future research.

References

1. Turing, A. M: Intelligent Machinery. Unpublished Report to the National Physical Laboratory. (1948) http://www.alanturing.net/turing_archive/archive/1/132/L32-001.html.
2. Ibid, 9
3. Ibid.
4. Turing, A. M.: Computing Machinery and Intelligence. *Mind* 59.236 (1950) 433-460.
5. Shannon, Claude E.: A Mathematical Theory of Information." *Bell Systems Technical Journal* 27 (1948) 379-423, 623-656.
6. Price, Robert: A Conversation with Claude Shannon. *Communications Magazine, IEEE* 22.5 (1984): 125
7. Shannon, Claude E: Letter to Vannevar Bush [16 February 1939]. *Claude Elwood Shannon: Collected Papers*. Ed. Sloane, N. J. A. and Aaron D. Wyner. Piscataway, N.J.: IEEE Press, (1993) 455-456.
8. Price, Robert: A Conversation with Claude Shannon. *Communications Magazine, IEEE* 22.5 (1984) 125
9. Shannon, Claude E.: A Mathematical Theory of Cryptography. Vol. Memorandum MM 45-110-02, Sept. 1, 1945, Bell Laboratories (1945) [unpublished materials on file at the British Museum]
10. Price, Robert: A Conversation with Claude Shannon. *Communications Magazine, IEEE* 22.5 (1984) 124
11. Fujimura, Joan H: Crafting Science: Standardized Packages, Boundary Objects, and "Translation": in Pickering, Andrew (ed.): *Science as Practice and Culture*, University of Chicago Press, Chicago (1992) 168-211
12. Bronowski, Jacob: Science as Foresight: in Newman, James R. (ed.): *What is Science: Twelve Eminent Scientists and Philosophers Explain their Various Fields to the Layman*, Simon and Schuster, New York (1955) 385-436
13. Ibid., 432.
14. Rogers, Everett M: *A History of Communication Study: A Biographical Approach*, The Free Press, New York (1994)
15. Ibid.
16. Schramm, Wilbur Lang: *How Communication Works*: in Schramm, Wilbur (ed.): *The Process and Effects of Mass Communication*, University of Illinois Press, Urbana (1955)
17. Simpson, Christopher: *Science of Coercion : Communication Research and Psychological Warfare, 1945-1960*. Oxford University Press, New York (1994)
18. Schramm, Wilbur Lang: *How Communication Works*: in Schramm, Wilbur (ed.): *The Process and Effects of Mass Communication*, University of Illinois Press, Urbana (1955) 3, 6.
19. Shannon, Claude E.: The Bandwagon (Editorial): in *Institute of Radio Engineers, Transactions on Information Theory*, IT-2, March (1956) 3
20. Berlo, David K.: *The Process of Communication: An Introduction to Theory and Practice*. Holt, Rinehart and Wilson, New York (1960) 11-12
21. Saygin, Pinar Ayse, Ilyas Cicekli and Varol Akman: Turing Test: 50 Years Later: in *Minds and Machines* 10 (2000), 463-518
22. Shannon, Claude E.: A Mind-Reading (?) Machine: in Sloane, N. J. A., Aaron, D. Sloane (eds.): *Claude Elwood Shannon Collected Papers*. Piscataway, N.J., IEEE Press, 1993 [1953].
23. Weizenbaum, J.: ELIZA--A Computer Program For the Study of Natural Language Communication between Men and Machines: in *Communications of the ACM* 9 (1966) 36-45
24. Colby, K. M. F. D., Weber, S. and Kraemer: Turing-Like Indistinguishability Tests for Computer Simulation of Paranoid Processes: in *Artificial Intelligence*, 3 (1972) 1-25

25. Saygin, Pinar Ayse, Ilyas Cicekli and Varol Akman: Turing Test: 50 Years Later: in Minds and Machines 10 (2000), 508
26. AI Research: Creating a New Form of Life - Who is Alan?: AI Research. 12 July 2005, 1:10 AM <http://www.a-i.com/show_tree.asp?id=59&level=2&root=115>
27. AI Research: Alan Logs – Robot calls me an asshole, Session #495874, 12 July 2005, 1:10 AM <http://www.a-i.com/alan_log.asp?id=113&level=3&root=115>

A Trouble Shared is a Troubled Halved: Disruptive and Self-Help Patterns of Usage for Co-Located Interfaces

Massimo Zancanaro, Chiara Leonardi

ITC-irst
I-38050 Povo (TN) Italy
{zancana, cleonardi}@itc.it

Abstract. The use of co-located interfaces can be more problematic than working on standard single-user software because people have to deal simultaneously with the dynamics of group behaviour and with the hassles of the (wrong) design. In this position paper, we report the some preliminary result of a qualitative study conducted on 10 small groups using a tabletop device in a natural setting. We discuss the several patterns of disruptive behaviour induced by the interface as well as the some patterns of collaborative appropriation of the system's functionalities.

1 Introduction

In this paper we present a preliminary qualitative study conducted on face-to-face meetings in which a co-located interface was used to support group activities. This device is conceived as a first step toward a larger integrated environment in the context of the European project CHIL - Computer in the Human Interaction Loop - which aims to create ubiquitous computing environments in which multimodal technologies are exploited to support human-human interaction and synchronous cooperation in an unobtrusive way.

We focus here on the first outcomes of a larger ethnographic investigation aimed at understanding not only the usability of the system but in particular how the people “appropriate” the technology (Dourish, 2003) and integrating it in their working practises.

The System

The CHIL tabletop device is a top-projected interface that turns a standard wooden table into an active surface. The user interface was designed around the concept of virtual sheets of paper that can be opened and used by the participants (the process of User-Centred Design of this system is explained in Falcon et al. 2005).

Each virtual sheet can be shrank or moved to save space and can be rotated to be made accessible to all participants. Participants can use a pen to draw or write (the pen position is tracked with a commercial tool based on infrared and ultrasound, due to hardware limitation only one pen is available for the group which therefore has to share it). A keyboard is also provided to write longer texts. Import and export functionalities are offered in order to allow the participants working on already prepared sketches as well as starting from white sheets.

Since one of the main goals was to support the group organization, two sheets of papers with special functions are also provided: the agenda and the “to do” list. The former contains the issue to be discussed. Issues can be added, removed or sorted. Each issue can be active or inactive. The system displays a time counter on the active issue; the counter is paused when the issue is made inactive.



Fig. 2. Users interacting with the tabletop device

The “Outcome note” list allows keeping track of the decisions taken during the meeting. Each entry is automatically associated with the agenda issue currently active (if any) and, through drag-n-drop, to one or more documents.

3 The Qualitative Study

Ten small groups composed of three up to six participants were invited to hold their meetings with the tabletop device. Before the start of the meetings, participants received instructions about the device. All the groups consisted of ITC researchers, none of whom involved with the CHIL project. Twenty people in total were involved and two participated in more than one group. All the interactions were videorecorded with two cameras and tabletop microphones; the video analyzed using MultiVideoNote¹. Several semi-structured interview were also performed with members of the groups.

¹ MultiVideoNote is an open source project for qualitative analysis on multiple video streams (<http://tcc.itc.it/research/i3p/mvn>)

3.1 Disruptive Patterns of Usage

A disruptive pattern of usage occurs when the interface hinders the flow of the interaction and comprises (or risks to compromise) the efficacy of the meeting.

One typical pattern is observed when the entire group is dragged out of the discussion to focus on the interface. Usually this happens when the group faces an unexpected behavior of the interface or when one participant tries to use a functionality of the table attracting the attention of the others. For example in one of the observed meetings, P. is trying to move a window while Z. is talking; after some failed attempts by P., the rest of the group is involved in helping P. while Z. gave clearly annoyed.

(ex. M. and N. try to use the interface simultaneously; P. tries to move a window when Z. is talking).

A second pattern of disruption happens when two or more persons negotiate the use of the system (the tabletop device at present does not allow multi-user interaction). In a meeting, N. wants to update the meeting while M. is working on drawing a workplan on a document. N. asks the pen in order to update the agenda item and the group starts discussing the item forgetting the finish the workplan.

Another pattern that can be recognized is when one single person is pulled out of the discussion because s/he trying to understand some functionality of the system. In most of groups, one person plays the role of the expert in using the technology. Although, the expert is usually the one who leads the group in adopting the technology, sometimes s/he isolate from the discussion. For example, N. plays with the agenda tool for more than 2 minutes in trying to understand how it works, leading M., the group leader, to repeatedly call him at order.

3.3 Self-Help Patterns of Usage

A Self-Help pattern of usage occurs when the group collectively learns to use the system either to solve a problem on the interface, to learn how to use functionality or to invent a new use.

The most apparent pattern can be called the “jigsaw” pattern. It consists of many different participants that contribute to the learning process. That is, nobody in the group possesses the knowledge but each single contribution increases the group awareness and stimulates others’ contributions. In a meeting, the leader tries to define the agenda items. The leader did not remember the how to manage the agenda and the entire group was progressively involved in solving the issue. Eventually, they manage to have a list of items done.

Another pattern is when the interface task cannot be accomplished by one person alone either because of cognitive overload or because of system’s limitations. For example, using the Agenda requires a continuous switch between activities - start an item using the pen, to write the item using the keyboard, to use again the pen in order to change the item, and so on. In several cases, this limitation leads the participants to toward an explicit division of labor where one person used the pen and a different one the keyboard thus playing the Orienteer and Group-Observer roles described in (Bales, 1970).

The third pattern is the “specialization of functions”: when one participant succeeds in doing a task (esp. after repeating attempts), s/he will be required (or volunteer) to perform it again the task in the future. For example, in a meeting M. does not participate too much in the interaction with the system but he succeeded at the very beginning in dragging a document in the notes (a very difficult task indeed). When later on the group needed to perform this task again, he volunteered. In the same meeting, the leader after being involved in dragging several documents in the trash bin, he nicknamed himself the “trashman”.

4 Discussion

This qualitative study shows initial insights on how groups can reduce the cognitive effort of using a co-located interface, and sometimes overcome bugs and design limitation, by a process of cooperative discover. It shows also that a co-located interface can systematically hinder the flow of interaction in a group interaction but acting as a disruptive tool (not necessarily because of bad design but also, like in the second disruptive pattern, because too rich in functionalities).

Although the work is still preliminary, we think that the collection of a number of such patterns may help in designing co-located interfaces that best suit the group needs of support and in providing guidelines for heuristic evaluation of such systems.

Acknowledgements

This work was conducted in the context of the European funded CHIL project. We would like to thank Daniel Tomasini, Vera Falcon and Fabio Pianesi for the discussions on the design of the use of the tabletop device and Valentina Proietti for her help in the ethnographic work.

References

1. Bales, R.F.. *Personality and interpersonal behavior*. New York: Holt, Rinehart and Winston (1970).
2. Dourish, P. The Appropriation of Interactive Technologies: Some Lessons from Placeless Documents *Computer Supported Cooperative Work*, 2003 , 12 , 465-490
3. Falcon V., Leonardi C., Pianesi F., Tomasini D., Zancanaro M. (2005) Co-Located Support for Small Group Meetings. Presented at the Workshop *The Virtuality Continuum Revisited* held in conjunction with *Computer-Human Interaction CHI2005 Conference*. Portland, Or. April 3.

Stupid computer!

Abuse and social identities

Antonella De Angeli¹ and Rollo Carpenter²

¹School of Informatics University of Manchester
Po Box 88 PO Box 88, M60 1QD
Antonella.de-angeli@manchester.ac.uk

²www.jabberwacky.com
rollo@jabberwacky.com

Abstract. This paper presents a preliminary analysis of verbal abuse in spontaneous human-chatterbot conversations. An ethnographic study suggested that abuse is pervasive and may reflect an asymmetrical power distribution, where the user is the master, and the chatterbot the slave. We propose that verbal aggression in this setting may be a social norm applied by users to differentiate themselves from the machine in what can be regarded as a form of interspecies conflict. The findings stress the importance of naturalistic, ethnographic studies to uncover social dynamics of virtual relationships.

1 Introduction

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. A number of life-like creatures are under development in research centres world-wide and some prototypes have already entered our everyday life. They are embodied conversational agents, chatterbots and talking heads, displaying a range of anthropomorphic features. These artificial creatures offer information, services and even company to whomever wants to or is capable of engaging them. We call these creatures social agents, as they are explicitly designed to build lasting and meaningful relationships with the user [1].

Overall, we are witnessing an extraordinary change in technology: the human metaphor has become the design model [2]. Technology is now intentionally designed to be human-like, to show a sense of personality and attitude, and to involve the user in social relationships. As a consequence HCI research has started exploring determinants and consequences of social relationships, trying to define a computational framework of social intelligence. Most of the research, however, has so far concentrated on the study of specific benefits of the interaction, such as trust and improved learning [3]. Positive emotions, including aspects of fun, humour and playfulness, have been investigated and used to inform the design of more engaging interfaces. Little attention has been devoted to the analysis of negative outcomes of the interaction, their behavioural manifestations, and to the need for research which

overtly addresses moral and ethical issues. This paper is a preliminary attempt to fulfil this gap. It addresses the occurrence of verbal abuse in a large corpus of spontaneous conversations with a chatterbot, a computer program which engage the user in written conversations.

2 The study

The analysis reported in this paper is based on the conversational log collected over the Internet by Jabberwacky, an entertaining chatterbot designed exclusively for entertainment, companionship and communication. Jabberwacky went on-line in 1997, and over the years has collected a large and active community of conversational partners. The peculiarity of Jabberwacky is that it is not hard-coded, but it learns from its users by adding user input to a linguistic database. Jabberwacky chooses its output based on an interpretation of the current conversational context and comparing it to conversations held in the past. The programming is abstract; not 'knowing' about English or any other language, Jabberwacky can speak many languages, to varying degrees. Everything depends upon the data that has been learnt to date, making it essentially a mirror of its audience. Because of its architecture, Jabberwacky will often claim to be human as, naturally, a majority of those who have spoken to it have made the same claim. Likewise, it will often accuse the user of being a robot, and may abruptly change topic or try to end a conversation. It has 'attitude', sometimes responding in kind to user taunts, and occasionally acting controversially, unprovoked. Generally, though, Jabberwacky is well-behaved, as the great majority of bad manners, obscenities, and abusive language have been filtered out.

2.1 Procedure

Monday, the 22nd of November 2004 was selected as a sample day to perform the analysis. The web log for that day reported 716 accesses to the dialogue page of Jabberwacky. For each access, the log reported a unique user identifier, time of the day, client IP, and user's hits (an indicator of the number of conversational turns). The conversation itself was recorded in a text file. A preliminary screening based on IP addresses comparison and hits frequencies, led to the deletion of 200 entries, which did not have any associated conversation. A selection of 146 conversations generated by different IP addresses was then extracted. It includes all the conversations with more than 20 user inputs (N= 103) and a random selection of shorter conversations. Note that this procedure does not guarantee that we have analysed a sample of 146 different users, as all the information we have is related to IP addresses, yet the number of conversations is sufficiently large to guarantee a reasonable sample. The conversations were subjected to lexical analysis. The corpus was normalized and conversational abbreviations were substituted with correct grammatical forms (e.g., "isn't" becomes "is not"). In this paper we concentrate only on the analysis of the users' conversational turns.

2.2 Results

The corpus was composed of 146 conversations, totalling 12,053 sentences with an average of almost 5 words per sentence. On the average the user produced 41 inputs per conversation. Some 7% of these conversations (N=10) were primarily conducted in a language other than English, and were discarded from the analysis. It is interesting to notice that all of them started in English and shifted because the user (N=8) or Jabberwacky (N=2) suddenly started speaking a different language. Only in one case the user asked the chatterbot if it spoke the language (“*Hablas espanol?*” in English *do you speak Spanish?*), even thought s/he did it directly in the foreign language.

The number of unique words produced by the user totalled 3,037, with 2,625 stems (i.e., the root of a words to which inflections or formative elements are added). The term word here is used in a broad sense to include not only gender, number and orthographic variations, but also misspellings, letter sequences made up to communicate emotions, or sounds (e.g., *AAAAAARRRRRRRRGGGGHHHHHHH*) and non-words which may have been produced to test the chatterbot skill (e.g., *cthulhu*).

The output of the stem analysis was sorted by alphabetical order and frequency of occurrence. All the stems with a frequency higher than 10 were extracted. This procedure gave rise to a sample of 277 stems (10.5% of the initial corpus) ranging from the word *you* ($f = 1751$) to the word *vagina* ($f = 10$). The sample was further reduced by retaining only stems which could be verbs, adjectives, or nouns and deleting auxiliary verbs (e.g., to do, to be, to have, will, shall, would, should, can, may, might, and could). The final set was then composed of 147 stems (total frequency = 3,829) ranging from *like* ($f = 189$) to *vagina* ($f = 10$). They were ranked on 40 levels according to their frequency of occurrence. This sample should give us a flavour of the prevailing topics and style of the conversations.

A semantic analysis was run to evince occurrences of insulting and offensive language, which was found to account for some 10% of the stems. The referents of terms of abuse could take the form of body parts, mental abnormalities or inadequacy, immoral behaviour, and sexuality. These referents were grouped in two broad dimensions, labelled *sexual harassment* and *injuries*.

The frequent occurrence of sexual harassment was the most surprising finding of the study. The *F word* occupied the 9th rank in the stem list ($f = 70$), immediately preceded by *well* ($f = 83$) and *love* ($f=86$) and followed by *computer* and *good* ($f = 68$). Several synonyms of male and female genitals were found in the corpus, and the word *sex* occurred 39 times (rank = 16). A deeper analysis revealed that 14 conversations (11% of the English-speaking sample) principally addressed hard-core sex and pornography. Harsh verbal abuse was the norm in these conversations, which were more similar to dirty soliloquies than to hot chats because Jabberwacky is designed not to have any knowledge of the topic. Anyway, this lack of participation did not appear to have disturbed the participants, as on the average they engaged in 10 conversational turns. An example of how Jabberwacky unsuccessfully tried to elude the topic is reported below.

User: No, do u want to have sex with me?
 Jabberwacky: Do you?
 User: Yes.
 Jabberwacky: Want to debate it philosophically?
 User: No, let's just have sex right now.

Example 1. Jabberwacky dealing with sex proposals

Injuries mainly referred to lack of honesty and lack of mental abilities. The user often called Jabberwacky *liar* ($f = 22$, rank = 29). This tended to occur whenever Jabberwacky provided the user with different information at different time during the conversation. It is an interesting example of anthropomorphism, as it implies attributing intentionality to the machine. A particularly harsh instance is provided below. Note how the abusive relationship intensifies (capital letters, exclamation marks) as Jabberwacky did not comply with the user request.

Jabberwacky What is your aim in life?
 User To not talk to you.
 Jabberwacky I won't.
 User You will you are my slave and if I tell you to do something do it strieght away.
 Jabberwacky Will you force me?
 User Yes ! TELL ME YOUR NAME NOW SLAVE !
 Jabberwacky I already told you it. I'm not repeating myself.
 User Yes you are now tell me it now you liar slave.

Example 2. Liar

In other case the users were eager to stress Jabberwacky lack of intelligence using words such as *stupid* ($f = 19$, rank = 32), *dumb* ($f = 10$ rank= 41) and *idiot* ($f = 10$ rank= 41). This mainly occurred when the user wanted to stress the artificial nature of Jabberwacky as in the following examples.

User That's right, you suppose. You're so dumb you can't even say that for sure.
 Jabberwacky You cannot prove that I am a computer program.
 User I never said that. The simple fact that you say that kind of stuff proves that you're nothing but a stupid program.

Example 3. Stupid program.

Since the corpus was gathered, Jabberwacky has developed in conversational ability. Users often stay at the site chatting for hours at a time, with many logs a day exceeding 100 conversational turns. A user recently stayed chatting for 7 uninterrupted hours, typing just under 1000 good-quality entries. Despite this fact, the

pattern of abuse continues, and we would like to end our analysis with two recent examples (recorded August 10, 2005).

```
Jabberwacky: I assume whatever you say is
              sarcastic.
User:        Yes, because you are stupid.
Jabberwacky: Not because you are a robot.
User:        You are a stupid robot.
Jabberwacky: Stupid maybe but no robot.
User:        You piece of sh**.
```

Example 4. Stupid robot (** have been added by the authors)

```
User:        I want to have dinner with you.
Jabberwacky: You are a computer you can't walk!!!
User:        What is your address?
Jabberwacky: I don't think I'll tell you. You'd
              kill me if I told you. Wouldn't you!
User:        Yes I would.
Jabberwacky: What is your favourite food?
User:        I WILL KILL YOU! I WILL HUNT YOU
              DOWN!!!
Jabberwacky: Please don't. . . I really have to
              leave.
User:        I WILL KILL YOU!
Jabberwacky: Good.
```

Example 5. I will kill you

3 Conclusion

This paper has illustrated some examples of verbal abuse produced by users interacting with a chatterbot. It is important to notice that while the analysis demonstrated that the phenomenon is common and pervasive, very little research had previously addressed it. This may be due to an important difference in data-collection. Our paper deals with spontaneous interactions, analysing conversations held in a natural, unconstrained setting, whereas most of previous empirical research was carried out in controlled studies, in which the users were well aware of being observed [3]. In this condition, social desirability bias is likely to have inhibited the occurrence of flaming and verbal abuse. For example, a few years ago we conducted a small ethnographic study asking a group of people to interact with a chatterbot for a week and hand in their logs for analysis [4]. Although some instances of verbal abuse were noticed and discussed, the phenomenon appeared to be qualitatively and quantitatively less worrying than what evinced in this study.

More research is needed to understand what triggers verbal abuse and how this behaviour will affect the adoption and usage of social agents. Social psychology suggests a link between aggression and identity [5]. The self-concept of each person is comprised of many different cognitive representations, or identities, which function

relatively independently and are activated in different contexts. There are at least three general levels of abstractions: personal identity (the self as an individual); social identity (the self as a group member); and human identity (the self as a human being). Activation of self-identities, or self-categorisation, is context dependent and affects people's behaviour. When a specific identity becomes salient in the perceptual system, people tend to act collectively in terms of the stereotypical dimensions of that identity. Social identities determine the attachment and the solidarity to other members of the in-group and the separation from others, who do not share that social identity. Besides that, collective values and norms are connected to social identities. Mechanisms of social information processing and judgement (e.g. social categorisation, comparison), are of great importance for the understanding of conflicts, aggressive behaviour and verbal abuse. Discrimination, degradation and abuse tend to target prevalently those who are perceived as different (out-group members).

In previous studies [4], [6], we have proposed that the most important social effect of chatterbots (and of any other forms of artificial life) is in the activation of the human identity, as the comparison with a machine is likely to make the user identify with other human beings. This human identity appears to be quite rudimentary when it comes to deal with etiquette: human tendencies to dominate, be rude and infer stupidity are all present in the interaction. We believe that verbal abuse in the interaction reflects a perceived power distribution in the relationship: the user is in the dominant position (master) and the chatterbot is in a subordinate position (slave). If this is true, then it is reasonable to believe that in this setting, abuse should be considered as normative and not a form of anti-social behaviour.

History has taught us that stereotypes, racism, prejudices, and attitudes towards minorities are difficult to modify. This being the case, for a long time to come, social agents must be ready to cope with their subordinate role, without losing their believability, or their capability for engagement and amusement. If abuse – or the threat of it – is part of the interaction, then a believable social agents must be aware of it, and must develop strategies to deal with it as suggested in the following extract of our corpus.

```
User:           Your stupid.
Jabberwacky:    I am currently in fifteen separate
                counties.
User:           What the hell I think i'm having a
                double conversation, one with myself
                and then your talking to yourself.
Jabberwacky:    Are you?
User:           My god your nerdish creators have
                made you into an ai engine that can't
                come up with a insult and in this
                modern day age you're going to need
                some good comebacks.
```

Example 6. On the need for a behavioural strategy to cope with verbal abuse.

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 socio-

psychological outcomes which may affect individuals, groups, or societies. There is an urgent need to explore the requirements for the establishment and negotiation of a cyber-etiquette to regulate the interaction between humans and artificial entities [7]. Will this etiquette emerge spontaneously, or will it require vigilance and reinforcement? Is the tendency towards abuse going to fade with experience, as it happened with computer-mediated communication, or will it last as a normative response to a minority perceived as inferior? Will respect for 'machines' grow along with their abilities, or will the abuse spiral upward thanks to a perception of a developing risk of inter-'species' conflict? Can virtual representatives or tutors perform their task if abuse, or the threat of it, is a part of the interaction? More research is needed to answer these questions. Technically, the filtering performed by Jabberwacky could be reversed, and the resulting conversations, in which both parties can aggress, would provide interesting material for future study.

References

1. De Angeli, A., Lynch, P., and Johnson, G.I.: Personifying the E-Market: A Framework for Social Agents. In: Proceedings of Interact'01, Tokyo Japan, 9-13 July 2001.
2. Marakas, G.M., Johnson, R.D., and Palmer, J.W.: A Theoretical Model of Differential Social Attributions toward Computing Technology: When the Metaphor Becomes the Model. *International Journal of Human-Computer Studies* 52 4 (2000) 719-750.
3. Bickmore, T.W. and Picard, R.W.: Establishing and Maintaining Long-Term Human-Computer Relationships. *ACM Transactions on Computer-Human Interaction* 12 2 (2005) 293--327.
4. De Angeli, A., Johnson, G.I., and Coventry, L.: The Unfriendly User: Exploring Social Reactions to Chatterbots. In: Proceedings of International Conference on Affective Human Factor Design, Singapore, 27-29 June 2001.
5. Turner, J.C.: *Rediscovering the Social Group: A Self-Categorization Theory*. Basil Blackwell Oxford UK (1987)
6. De Angeli, A.: To the Rescue of a Lost Identity: Social Perception in Human-Chatterbot Interaction. In: Proceedings of AISB'05 joint symposium on Virtual Social Agents, University of Hertfordshire, Hatfield, UK, 12-15 April 2005.
7. Miller, C.A., (ed.) *Human-Computer Etiquette: Managing Expectations with Intentional Agents*, Vol. 47, 4. *Communications of the ACM* (2004).

Abuse and the cyberworld:
Relationship of the virtual and the
real

Dark Lancaster

Alan Dix, Kiel Gilleade, Masitah Ghazali, Jane Holt,
Simon Lock, Jennifer Sheridan

Computing Department, InfoLab21, Lancaster University UK
alan@hcibook.com
<http://www.hcibook.com/alan/papers/interact-abuse-2005/>

position paper for Interact 2005 workshop on "ABUSE the dark side of HCI"

In this position paper we'll outline a few ongoing and planned projects at Lancaster that are not all sweetness and light. In some we are interested in some of the darker aspects of human nature: frustration when things go wrong in order to design games with the right emotional impact; and anger of those seeking jobs in order to help train those who need to defuse fraught situations. In others we deliberately seek to design 'bad' situations; obviously this is necessary to study issues like frustration, but also we design bad things in order to understand what is good! Finally, there are times when good is dark and the bright light of day needs to be shrouded just a little.

Frustration

Slowly you edge down the dark corridor, distant daylight dimly illuminates the walls on either side, your heart races you know there are others in these corridors and they are after you. You near the bend. What is beyond? Too late you wheel round only to be momentarily blinded by a bright light, then you hear a pistol crack and see the ground race towards you, already red with blood, your blood. Game Over.

Video games are escapist, virtual, just a game, but in the heat of the moment the emotions can be very real.

Research on affective gaming seek in various ways to understand, measure or infer the emotions or more normally simply arousal of the gamer in order to adapt the game and create a more engaging, more immersive experience.

Early work used heart monitoring to measure arousal and create a game that modified the level of challenge accordingly, low levels of arousal led to more enemies attacking, although easier to kill ones in order to maintain the same level of difficulty.

More recently we've focused on frustration, both the 'proper' frustration when you get shot by a cleverer opponent for the 10th time, but also the frustration when a moments delay in the controller means you can't duck in time.

Of course to study frustration we need to create games that cause it :-)

See refs: [4] & [5]

Cruel Design

We grow up in the real world, physical things that respond to gravity, bump into each other, have weight, solidity, stay where they are put until moved. Then we move into the electronic world whether virtual reality or simply a desktop interface. Things are no longer so simple and the laws of physicality breakdown: there are delays between action and effect, things change without apparent agency, it is a world of magic and not a little superstition.

We wish to understand the ways in which design can recruit our natural understandings of the natural world to create better tangible interfaces and ubiquitous environments. Some of this we can find by examining existing artefacts, mining the implicit knowledge the designers invest in these. This has enabled us to produce putative design guidelines, but there is only so much you can learn from good design.

In neurology it has been the freak accidents and illnesses, skull fractures and cancerous growths, that have revealed much of the structure of the brain. It is when systems fail that we begin to understand how they succeed.

So we look towards cruel design, experiment on systems designed to be strange, hard, annoying or simply impossible to use. By manipulating the level of physical coherence of physical-digital mappings we are delving into the properties that make things work well by making them work badly.

See refs: [2] & [3]

Anger

In a government office a client comes in - of course in an increasingly corporatised public service everyone is a customer. The client is a citizen or perhaps wouldn't use that term, perhaps just a frustrated person with a problem that needs sorting: benefit not paid, bills piling up. You are behind the desk - what are you going to do about it? What do you say? You can't access it on your computer; it's not your responsibility; you want to send me to another office. I don't care about your computer or your responsibility, I don't want to go to another office - I need my money, and I need it now.

Abuse, violence and emotional turmoil are a day-to-day part of many 'front line' public services. How do you train people to deal with traumatised, angry, upset clients? Training videos will often show scenarios: how to defuse potential problems - how to say the right thing in the right way. Some of this is about what you say, but a lot more about how you say it; when you can't help you need to be helpful.

We are planning to start a new strand of work, building on previous systems including the affective gaming and several arts-technology installations, to produce virtual avatars that can respond in emotionally realistic ways to detected emotions in the user. Can you soothe the angry avatar before there is bloodshed in the office?

See refs: none yet, only just starting!

Underside

You pace nervously in the gloomy hall. Daylight barely filters into this deep dark place. Far off you see another person glancing up and down. What is he thinking? What is he waiting for? Something in the way he looks at you makes you twist round, only to be momentarily blinded by the flash of lights. You hear the dull thud and hiss of air-brakes and the door opens in front of you. "A return to Lancaster please". And the other figure in the darkness? He must be waiting for the number 42.

Beneath the central square in Lancaster University there is an underpass where the road cuts under the very heart of campus. The buses stop here and passengers loiter. In this, the underbelly of campus, one of the first phases of a eCampus is being deployed, a project to infiltrate the whole of Lancaster University campus with interactive public displays and sensors.

While passengers wait three projectors turn the opposite wall into a huge display, mobile phones are used to interact with the displays so that they go beyond mere 'next bus' signs to an immersive experience submersed beneath the feet of unknowing passers-by in the square above.

As befits this subterranean world content in the underpass will have a subversive edge to it including performance art and interactive applications. An artist in residence is preparing the first installation now.

See refs: For general performance related work [1] & [7], but the underpass itself is still in progress (opening 1st October 2005), so no papers yet!

Reflection

Looking at these examples we can see three types of 'badness':

- (i) things that are bad but we want to study (perhaps to alleviate them)
- (ii) situations where a little bit of badness is good ;-)
- (iii) using difficult or bad situations to understand what is good

The first project on video game frustration has all of these aspects. (i) If the game is just not fluid enough, or if the puzzles or physical actions are too difficult, the gamer will become frustrated and stop playing. So we want to know whether we can use physiological signals to detect this and perhaps give the gamer hints to help. (ii) Of course in a game a level of frustration is right, we don't want it too easy!! (iii) To study both the bad and good frustration we need to create games that are deliberately frustrating!

The second project on cruel design is focused almost exclusively on the last aspect (iii). We want to create designs that are good and exploit natural physical understanding and abilities, and we are creating mappings that are bad in different ways to understand what is good. This is in fact being done partly in conjunction with the gaming project creating small video games but where the mapping between controllers and their effects obeys different physical-digital properties. In some cases

we will have mappings that are easy to understand in your head (e.g. right hand joystick is ‘increase something’ left hand is ‘decrease something’ and each joystick movement controls the thing at the relevant location on the screen. In others this mapping will be odd and hard to remember (or even dynamically changing!), but the joysticks will have a ‘natural inverse’ property – pushing the *same* joystick in the opposite direction has the opposite effect.

As noted the study of deficient or unusual behaviour is a common method in many areas. For example, Ramachadrin uses phantom limb sufferers to study ‘normal’ brain function and synesthesia to understand metaphor [6]. Sadly in human–computer interaction researchers (and reviewers) find it hard to comprehend the deliberate design of bad interfaces! We clearly need a change in culture within our discipline, as we seem to confuse good design and good science. Although the end points are often similar the routes and methods often diverge.

Moving back to dark Lancaster, the ‘anger’ project is only just beginning and here we are interested partly in the training aspects (i) and of course for this will have to simulate bad situations – in this case using virtual angry avatars (iii). However, the deeper lesson we want to learn is about the nature of emotionally reactive avatars whether dark or light emotions! The aim is to move away from the angry clients and eventually look at emotionally reactive virtual dance partners. One question we have is whether onlookers or dancers can tell the difference between real dancers (portrayed virtually using body movement sensors) and virtual ones – a sort of emotional Turing test!

Finally the underpass project is solely related to (ii) the positive aspects of slightly dark emotions. Just like frustration in the video game a certain amount of ‘bad’ emotions are a good thing; this is why we have them. Without subversion there would be no change.

References

1. A. Dix, J. Sheridan, S. Reeves, S. Benford and C. O'Malley (2005). Formalising Performative Interaction Proceedings of DSVIS'2005.
2. M. Ghazali and A. Dix (2003). Aladdin's lamp: understanding new from old. In 1st UK-UbiNet Workshop, 25–26th September 2003, Imperial College London
3. M. Ghazali and A. Dix (2005). Visceral Interaction. Proceedings of HCI'2005, Vol 2.
4. K. Gilleade and A. Dix (2004). Using Frustration in the Design of Adaptive Videogames. Proceedings of ACE 2004, Advances in Computer Entertainment Technology, ACM Press. pp. 228–232
5. K. Gilleade, A. Dix and J. Allanson (2005). Affective Videogames and Modes of Affective Gaming: Assist Me, Challenge Me, Emote Me (ACE). In DiGRA 2005 – International DiGRA Conference. Vancouver, June 16th-20th, 2005
6. V.S. Ramachandran (2004). A Brief Tour of Human Consciousness. Pi Press.
7. J. Sheridan, A. Dix, S. Lock and A. Bayliss (2004). Understanding Interaction in Ubiquitous Guerrilla Performances in Playful Arena. Proceedings of HCI 2004, Springer-Verlag. pp. 3–18.

There's something about email: Technological and Cultural Interventions in the Problem of Inadvertent or Casual Rudeness in Email

Mark Blythe & Mark Jones

University of York, England email, mblythe@cs.york.ac.uk

University of Wolverhampton, England email, markjones@wlv.ac.uk

1 Introduction: No Need to be Rude Dear

The darker side of email is usually considered in terms of unsolicited mail or spam (e.g. Cerf 2005, Balvanz, Paulsen, and Struss 2004). The problem is well understood and numerous filter systems, along with a variety of legal remedies, have been developed to tackle it with varying degrees of success. Personal abuse including workplace bullying and sexual harassment are also well-recognised problems (e.g. Sipior and Ward 1999). As long ago as 1997 it was pointed out that the convenience of email encourages abuse and the ease of using email extensions encourages mass mailouts (Berghel 1997). It is also frequently noted that email eats into our time and that the sheer volume of them can make us not only less productive but more stressed (e.g. Beale 2005). These, however, are problems generated by the form of the technology other issues are raised by its genre. Email as a form of writing falls somewhere between a letter and a note. Because it is a relatively new medium conventions of polite address have not yet been fully developed (Baron 2003). This can cause anxiety, tone is frequently misinterpreted and offence is taken. Jokes are often misunderstood despite the development of makeshift emoticons such as the colon, dash, bracket smile (Hancock 2005). There is, it seems, something about email that makes it easy to be inadvertently rude – something about email, perhaps, that encourages it. This paper outlines a lightweight method for collecting data on this phenomenon; it then considers some examples and ends with sketches of technological and cultural interventions.

2 Method: Reciprocal Storytelling and Email Collections

This paper began as a conversation between the two authors who exchanged stories about rude emails they had recently received. This reciprocal storytelling was then extended in an email call to colleagues to send in examples of rude emails that they had received. Many of us have very large email archives which can serve as a resource. Surprisingly perhaps, it was found that a number of people specifically collected rude emails and these collections yielded a data set too rich to be adequately dealt with in a short paper.

3 Speed and Convenience

Going through an inbox can be a frustrating experience not just because the spam filter has missed so much spam but also because colleagues can appear to be being very rude in a routine manner. A colleague, for example, answers a question with a one word yes or no, as in the exchange below:

Hi James,
Is this where we're putting the discussion paper?
Ivan

Yup
-James

The context of any email exchange is of course crucial and it should be noted that the initial communication came from a subordinate and the one word reply received was from a line manager. The subordinate's message began with a greeting which the line manager eschewed. The ease of a three key press reply is clearly a factor in the brevity of the response and so too is the power relationship. But so too is the lack of established convention (Goldsborough 2005). The conventions which exist in conversation and epistolary exchanges are there to make social interaction easy and routine. Imagine a similar verbal exchange where a greeting is ignored, a question is answered with one word and the exchange is ended without a goodbye. The "dear" salutation of a letter is too formal for short notes and various forms of "hello" and "hi" have, to an extent, replaced it in email; but these are not well established and colleagues may feel that in a short exchange of notes no salutation is necessary at all, especially if time is short and emails are being answered at the same time that a number of other tasks are undertaken. But speed and ease are not the only problems.

4 Anonymity and Impersonality

While it has been argued that email is a casual form of communication, more similar in nature to the telephone than the letter (Baron 2000), it could equally be held that email can actually be as impersonal as a typical memorandum. The second author of this paper has recently established a standard reply of "Please sign your emails. I don't write to people I don't know" in response to a flood of communications coming from students via the university's virtual learning environment's capability to 'email tutor'. The students' ignorance of the name of the person they were communicating with seems to lead to a determination to keep the staff member similarly in the dark. A different kind of impersonality is generated by mailings to multiple users. This email was sent out by the technical support staff at a UK University to all staff and students.

Please be aware that computers are not sweets. They are not mix and match at your whim. If you have a problem with a component

do NOT go and take one from another computer, inform [deleted] of the problem and we shall solve it. You definitely do NOT go taking a mouse for example from another computer just because it is newer than yours.

It may be argued that the patronising tone and offensive manner in this email are typical of IT support staff. But there are conventions of the medium that, intentionally or not, make it worse, including the capitalised “NOT” (a usage which usually signifies a raised voice). No addressee is specified neither is a sender. It is a totally anonymous text without a particular reader or author. It is well understood in chat room domains that anonymity encourages abuse but in email it is not just the anonymity of the sender that is problematic; recipients are also anonymous, and frequently numerous, so the impersonality is multiplied.

5 Right Message Wrong Person

The examples of presumably inadvertent rudeness above are instances of the wrong message to the right person. Email also makes it very easy to send the right message to the wrong person. A colleague of one of the authors recently made the mistake of responding to a standard circular email with a tirade about one of the institution’s policy documents. The unfortunate correspondent only discovered that the reply had been sent to the entire mailing list when summoned to the manager’s office. This same manager, however, had previously suffered a similar embarrassment when a draft discussion document, meant only for the eyes of senior managers, was sent to the institution’s staff mailing list. The error was compounded by an attempt to ‘recall’ the email, which of course only signalled to the recipients that an error had been made, and further by one of the senior managers, who responded personally to the author of the original email and also managed to distribute the message to the whole institution. Each of these emails contained material that, while not exactly libellous, was certainly read by a number of their accidental recipients as patronising and dismissive of the staff in general, or of sub-groups or individuals. It could of course have been worse. Mass forwarding facilities mean that embarrassing emails can become global phenomena. Clair Swire famously sent an email to a lover saying his sperm was “yum”, he forwarded it to friends who forwarded it to friends until the distribution became virally global (Mikkelson 2000). Although this sounds like an urban legend it illustrates how technology can offer ever new opportunities for human beings to make fools of themselves. Recently Harry Shearer featured two email apology stories on the *Le Show* programme’s “apologies of the week” feature;. consultant Chris Nelson, author of the influential Nelson Report on American foreign policy in Asia, sent an exceptionally frank report intended for the embassy of South Korean to the entire subscribers list for the Nelson report. He warned the embassy that if anyone on capital hill saw it he would have to seek political asylum. In the report he said the administration’s failed policy towards North Korea and its bunker mentality would not change while Dick Cheney was vice president and described senior members of the Bush administration as “genuinely mentally unbalanced”. More than

800 people received it including those he criticised or identified as people who talked to him. Nelson said “In a single moment of stupidity I have hurt and betrayed many who have tried so generously to help and who share my deepest fears about Korea policy. Apology is impossible at this point I can only ask mercy”. Similarly, a New York assemblyman Willis Stevens sent out an email that referred to his constituents as idiots to 300 people in an online discussion group thinking he was sending it to an aid. The message said he was “just watching the idiots pontificate” within an hour he started writing his apologies (Le Show, Harry Shearer.com 3.6.9)

Microsoft Office 2003 features information management technology that allows users to put limits on how many times emails can be copied, printed or forwarded and it was advertised with the tagline “The oops I just hit ‘reply all’ era is over”. However just as a user may accidentally hit reply all they may also forget to mark certain messages as sensitive or even know that they are sensitive until it is too late. While these interventions begin to address the problems they cannot yet eliminate the kinds of human error described above.

6 Content Filter Systems: This Sounds Smug, Are You Sure You Want To Send It?

There are a number of email systems which attempt to warn users when they are about to send a potentially offensive message. In their email correspondence during the production of this paper the first author sent a message to the second in which, he typed a word which many people would consider obscene. Before the message was sent the system flagged up a ‘three chilli’ mood warning indicating that his message is “The sort of thing that might get your keyboard washed out with soap if you get my drift, you might consider toning it down”. This humorous message is displayed whenever curse words are found in a message. Of course the user is more likely than a computer to know whether a curse word will offend a particular person or not. Language is contextual but it would not take artificial intelligence to take better account of context. Systems can be instructed to recognise not only curse words but contexts in the sense of who the recipient of the letter is and whether it is a friend or work colleague. But such systems cannot recognise other forms of offensive emails.

7 Technological Interventions

There are a number of technological solutions that are already available or that have been proposed to solve these problems. Cooling off periods of five or ten minutes between the send command and actual sending would amount to an undo feature, though the delays caused to the otherwise near-instantaneous communication available by email would certainly curtail its use. Greetings as well as subscriptions (sign offs) could be automated, and perhaps personalised for attachment to particular contacts. It would certainly be possible to warn a user that their one word reply might be considered offensive. Warning systems could also be developed that would flag up

notes on tone, perhaps activated by certain key phrases or, even, syntactical variants. For instance, it would be possible to pick out the phrases “not being sexist but” or “not being racist but” or “not being personal but” and flag a message saying – invariably these phrases precede something offensive, come on guys who are we kidding? Other forms might be recognised to provoke a message like – you sound patronising, are you sure you want to send it. However, it is unlikely that most forms of rudeness could be identified by filters. By their very nature, sarcastic and ironic messages are inversions of their intended meanings, and it is hard to imagine a software solution to this obtuseness. Russell Beale recently suggested changing tools in order to address these problems, rather than use email for one word replies of the kind quoted above he suggests the use of instant messaging systems (Beale 2005). This might help colleagues whose brief messages are intended as chatty and totally informal rather than rude. However, technological and behavioural approaches are perhaps overly deterministic, and could be perceived as unacceptably interventionist. User behaviour can sometimes be shaped but not always controlled by design. Indeed, the development of “netiquette” might be better served by cultural interventions.

8 Cultural Interventions

Europeans sometimes decry the insincerity of the American phrase “have a nice day” and complain that it means nothing but “goodbye”. But this overlooks the fact that “goodbye” is itself a derivation from a previously more loquacious, and specifically meaningful phrase, “god be with you”. and that language in use is continually undergoing both morphological and functional modifications. Forms of greeting, farewell and address are particularly subject to grammaticalization, in which words can have both their form and meaning near-simultaneously changed as users adapt them to new communicative environments (Aitchison 2001). Ritualised forms of salutation and subscription develop over time to facilitate routine and easy interactions (Arnovick 2000). Netiquette (see website references below) already a powerful force in controlling users’ online behaviour, will undoubtedly encourage the development of modes of politeness in email, as long as the form itself can survive spam and remain an important medium. The finding that some individuals collect rude emails is in itself interesting and suggests that an anonymised institutional online collection of them might be, not only entertaining, but helpful in the development of boundary markers for inappropriate behaviour.

Bibliography

- Aitchison, J. (2001). Language change: progress or decay? (3rd ed.). Cambridge: Cambridge University Press.
- Arnovick, L. K. (2000). Diachronic pragmatics: seven case studies in English illocutionary development. Amsterdam: John Benjamins Publishing Co
- Balvanz, J., Paulsen, D. & Struss J. (2004). Spam software evaluation, training, and support: fighting back to reclaim the email inbox. In Proceedings of the 32nd

- annual ACM SIGUCCS conference on User services, Baltimore, MD, USA (pp. 385-387). New York: ACM Press
- Baron, N. S. (2000). Alphabet to email: how written English evolved and where it's heading. London: Routledge.
- Baron, N. S. (2003). Why email looks like speech: proofreading, pedagogy and public face. In J. Aitchison & D.M. Lewis (Eds.). New Media Language (pp. 85-94). London: Routledge.
- Beale R. (2005) Email *Interfaces* British HCI group 63 Summer 2005
- Berghel, H. (1997). Email, the good, the bad and the ugly. Communications of the ACM, 40(4), 11-15.
- Goldsborough, R. (2005). E-mail manners: thinking about style and usage. Reading Today, 22(5), 11.
- Cerf, V. G. (2005). Viewpoint: spam, spim and spit. Communications of the ACM, 48(4) 39-43.
- Hancock, J. (2004). LOL: humor online. Interactions (special issue, More Funology), 11(5), 57.-58.
- Mikkelson, B. (2000). Under the yum-yum tree. <http://www.snopes.com/risque/tattled/swire.htm> (accessed May 29, 2005).
- Sipior J. C. & Ward, B. T. (1999). The dark side of employee email. Communications of the ACM, 42(7), 88-95.

Websites

- <http://www.albion.com/netiquette/>
- <http://www.learnthenet.com/english/html/09netiq.htm>
- <http://www.bspage.com/1netiq/Netiq.html>
- <http://www.harryshearer.com/> Le Show, July 3 2005 broadcast

Psychological Background of Cybersexual Activities: Qualitative Survey of Specific CMC Communication (among Czech Users)

Mgr. Radana Divínová

Masaryk University in Brno, School of Social Studies
Department of Psychology
Joštova 10, 602 00 Brno
Czech Republic
radana@fss.muni.cz, www.cybersex.cz

Abstract. The Internet and computers significantly influence the ways of presenting and consuming sexual content. In the presented study cybersexual activities and their users are surveyed. We understand cybersex as erotically suggestive or explicit messages or sexual fantasies that people exchange via the net. The qualitative survey conducted in 2003 outlined the psychological background of cybersexual activities. It belongs here: (1) needs that contribute to the motivation, (2) perceptions of cybersex, (3) sense of cybersex, (4) influence of cybersex upon the real sex, (5) the advantages of cybersex compared to real sex. In this paper only the first, second and last category are enlarged

Keywords: Internet, online sexuality, sexual behavior, chat rooms, needs, motivation, masturbation, grounded theory

1. Introduction

The Internet and computers significantly influence the ways of presenting and consuming sexual content. From time to time there appears the statement about the proportion of Internet visits that are in some way linked to sex. Internet sex, net sex, online sex, compuSex, cybersex, are the terms most frequently used for sexual experiences gained online. Cybersexual activities and their users are surveyed in the presented study.

There are different attitudes to the concept of cybersex. Schneider and Weiss [1] consider every form of sexual expression via the computer or the Internet to be cybersex – from viewing and downloading pornographic materials and subsequent masturbation, through the exchange of erotic messages or reading and writing of

sexual letters and stories, to the placing of ads for sexual partners, all the way to love affairs maintained through e-mail and chat. R. Noonan [2], whose definition is used further, understands cybersex only as erotically suggestive or explicit messages or sexual fantasies that people exchange via the net. Masturbation usually forms part of cybersex.

Alvin Cooper [3] stresses three primary factors that facilitate online sexuality. He termed them the *Triple A Engine* (accessibility, affordability, anonymity). Accessibility refers to the fact that on the Internet millions of web pages with pornography are constantly available and in chat rooms one constantly finds people looking for cybersexual experiences. Large quantity of pornography is available for free (affordability) on the Internet. People can move on the Net in complete anonymity and they also perceive communication as anonymous. The components of the Triple A Engine involve a great risk for users who face sexual compulsion or are mentally vulnerable and thus they are prone to compulsive behaviour [3].

Kimberly Young [4] developed a similar model (*ACE model*). According to her the factors that determine the appeal of cybersex are anonymity, convenience and escape. Further facilitating factors are the legality of cybersex, the zero risk of sexually transmitted diseases and easy concealment from the partner.

The first pilot survey of cybersexual behavior among Czech users was conducted in 2003. The qualitative methodology was used and as cybersex were considered only activities, where at least two people participate.

2. Methodology

Advanced Internet users, that visit the interactive environments – chatrooms were the target group of the survey. Importance has the number of hours spent in the chat rooms. The necessary condition for submission into the sample was repeated experience with the cybersex.

2.1 Data source

The research was conducted through the Internet only (taking advantage of an online disinhibition). The data were obtained from semistructured interviews. The interviews were held with individuals visiting one of the biggest Czech chat server Xchat. Individuals with highest number of hours, with certificate¹ and e-mail address in the profile were selected from the Xchat statistics. Fifty explaining e-mails with the offer to participate on the research were send to them. What can be expected, some of the users have no experience (never practice) with cybersex despite huge amount of hours. On the other hand, some experienced users do not put their e-mail address in

¹ Certificate means that the identity of the user (gender, name, age) is verified. The certification is granted by the certicator, after the personal meeting.

the profile. Some of these users were contacted and asked for the interview straight in the sex-oriented chat room. Duration of the interview was between 90 and 250 minutes, mostly 150 minutes. The individuals were pre-admonished, that the interview is saved for subsequent analysis. All the interviews were held on the Internet, in the Xchat chat rooms or using the ICQ.

2.2 The description of the sample

Because of the qualitative research design it was not necessary to compose representative sample. In such a specific environment it would be hardly possible. The research file constitutes 19 participants at the age 15 to 39 years, 11 men, 8 women. The most represented group were men at the age 25-29. The scatter table of respondents follows.

age	male	female	total	No. of hours spent on chat	male	female	total
15 - 19	1	1	2	100 - 199	3	0	3
20 - 24	3	2	5	200 - 499	1	1	2
25 - 29	5	1	6	500 - 999	4	2	6
30 - 34	2	3	5	1000 - 1999	2	4	6
35 - 39	0	1	1	2000 - 2999	1	1	2
total	11	8	19	total	11	8	19

2.3 Methods

The basic research method was the analysis of the interviews with the use of Grounded Theory method [5]. This qualitative approach helps to understand the surveyed reality and it builds the theory in areas, where very little is known.

The authors Anselm Strauss and Barney Glaser came with the „grounded theory“ in 1967. The principle of this strategy is the theory development from existing data directly, purely inductive without being knowingly prepared as hypothesis. During the research process the emerging structures are continuously categorized and analyzed. The preliminary generated hypothesis and concepts are refined, changed or replaced.

3. Results

Within selective coding some important categories emerge, we titled them jointly *psychological background of cybersexual activities*. It belongs here: (1) needs that contribute to the motivation, (2) perceptions of cybersex, (3) sense of cybersex, (4) influence of cybersex upon the real sex, (5) the advantages of cybersex compared to real sex. In this paper only the first, second and last category are enlarged.

3.1 Needs that contribute to the motivation

Motives that lead individuals to cybersex are usually - albeit not necessarily - conscious. The state of motivational tension is created by an unsatisfied need that can lead an individual to spend a significant amount of time on the Internet. It is important to know that these needs can mutually interact [6]. The survey revealed some of the needs that contribute to the motivation of people who practise cybersex.

Desire for sexual satisfaction is a reason often stated by the users. Their real sexual lives do not sufficiently satiate their need for sex.

Desire for knowledge appears to be an important motive for cybersexual activities. Children and adolescents enter sex-oriented rooms and gain information about human sexuality and their first "theoretical" sexual experiences. The Internet's anonymity encourages adults to get to know and explore different sexual practices. We consider the possibility to "try it" one of the major factors in the appeal of cybersex. This ability enables an individual to reveal various forms of his/her sexuality when engaged in these activities.

Desire to be oneself. For many users, the ability to speak freely about anything, to be open and without any of the masks that they feel are required in real life, is the greatest appeal. In the anonymous Internet environment, fear of rejection and condemnation are less evident and thus they do not force the individual to behave in a conformist way. Many individuals are unable to open up to their closest partners about sex. They might worry that the partner would not accept their sexual desires or that such an attempt could damage the relationship.

Desire to break. In order to maintain physical and mental well-being, it is important to take breaks from everyday duties and pressure. For many, the Internet, and particularly chat rooms, became places where they can relax. Many people view cybersex as relaxation or fun.

Desire to be wanted. If a person in the real world lacks social relationships to create the feeling of being cared for and of being important for others, that person may attempt to create these feelings through communication in chat rooms. If these feelings are satisfied through cybersex, it is likely that this type of communication will become an important part of the person's life. For many, cybersexual relationships are a way of satisfying the need for friendship and love.

Desire to find a partner for real sex. Relationships are established more easily in the chat room than in the real world. Online disinhibition significantly simplifies dating for the shy and introverted, but even for others it is a more convenient and

often more effective way of starting new relationships. If an individual, moreover, looks for a person with similar sexual interests, the anonymity of the Internet provides an ideal environment and medium. A significant number of users express the desire to try unusual sexual practices (fetishist, sadomasochism, group sex, etc.) in real life.

Excitement stemming from anonymity. An individual can participate in cybersexual activities without having to disclose his identity. The majority of users also claim to appreciate the fact that their cybersexual partner remains anonymous. Users are usually not interested in knowing what their partner looks like; for some it is precisely this anonymity, and the excitement connected with it, that make cybersex attractive.

Excitement stemming from interaction. The fact that users prefer cybersex to the "easier" and less time-consuming auto-eroticism with the use of pornographic materials shows the different nature of both possibilities for self-satisfaction. The most important difference lies in interaction provided by cybersex, which cannot be provided by a magazine or by a video. For some, the fact that they are talking to someone else can itself be as exciting as the content of the conversation.

The categories of "excitement stemming from anonymity" and "excitement stemming from interaction" differ from the others in that these categories are not needs, and thus cannot be primary motives. The excitement is intrinsic to the activity on the Internet. It can, however, be very strong motive for returning to cybersexual activities.

3.2 Perceptions of cybersex

Cybersexual activities are perceived differently, even among the users themselves. Some reasons that users cite for cybersex include:

Cybersex as a substitute for real sex. Those individuals who have no access to real sex or those who consider their sexual lives unsatisfactory view cybersex in this manner.

Cybersex as a different form of sex and a supplement to real sex. A large proportion of the users consider cybersex a pleasant variation in their sexual lives. If they have a partner, it can be a way of escaping from the stereotype of a couple's sexual life and remaining "faithful". It is often used by partners if they cannot be together physically, or as a form of foreplay.

Cybersex as a pleasant feature in auto-eroticism. A number of users view their cybersexual activities the same way as masturbation. They place this type of auto-eroticism in the same category as viewing pornographic materials. They often stress that thanks to the possibility of talking to someone it is much more pleasant and interesting.

3.3 The advantages of cybersex compared to real sex

For some participants cybersex has certain advantages also compared to real sex. The users themselves usually state the following:

Easy access. If one has access to the Internet at home, in the dorm or elsewhere in private he can access sexual conversation and cybersex very quickly and easily.

Lack of commitment. Cybersex does not make one responsible, it does not involve a commitment unless the individual wants it.

Greater openness and relaxation. If one feels safe and does not have any responsibility then he will be more relaxed and open in his conduct (in this case communication). Evidently one "only" communicates here, but communicates about what he does, what imaginary sexual activities he engages in. Cybersex enables him to try out his secret sexual desires within the safety of his home.

Possibility to set limits. Although a certain sexual activity is the user's secret fantasy, it does not mean that he will like its realization. Something that was not a part of the fantasy might occur. Cybersex allows him to be in control over "what is happening".

Unimportance of physical and personal attributes. Unless a camera is present the users have basically no way of verifying who they communicate with. This enables users to develop their own fantasies, to create the ideal partner. In the environment of the chat room there is no disadvantaging due to physical appearance. The ability to verbally express fantasies and ideas is important.

Possibility of satisfaction according to one's needs. The individual can be satisfied in the way he finds the most pleasant. This is true for the type of fantasies accompanying cybersex, the use of pornographic materials and also for auto-erotic stimulation.

It is not considered infidelity. Individuals involved in cybersex usually do not feel that they cheat on their real partners. In case feelings of remorse appear it is usually when an emotional bond with the cybersexual partner is established.

4. Conclusions and discussion

Considering that the results are coming from a qualitative survey, they can not be generalized to the world's cybersex users population. In any case it makes possible to enter the specific world of cybersex users. It approaches the way they behave (in chat rooms and outside them), they regard, perceive, and apprehend their activity. It contributes to understanding of the aspects of new phenomenon in human sexuality and human communication.

The consumption of cybersex can become pathological. Some studies have referred to so called cybersexual addiction [1], [3]. We do not concentrate on the pathology as most of the studies do. Our aim is to explore the cybersexual activity, from psychological point of view, as a new kind of human behavior. In the following study we will ask, what positive impacts these activities can bring into the human life, sexuality and the changes it brings to them.

References

- [1] Schneider, Jeniffer P. and Robert Weiss. (n.d). *Understanding Addictive Cybersex*. URL: <http://www.cybersexualaddiction.com/understanding.php>
- [2] Noonan, Raymond J.: The Psychology of Sex: A Mirror from the Internet. In: Gackenbach, J. (eds.): *Psychology and the Internet: Intrapersonal, Interpersonal and Transpersonal Implications*, Academic Press, San Diego (1998) 143-168
- [3] Cooper, Alvin, Coralie R. Scherer, Sylvain C. Boeis and Barry L. Gordon. 1999. "Sexuality on the Internet: From sexual exploration to pathological expression." *Professional Psychology: Research and Practice* 30, 2: 154-164
- [4] Young, Kimberly S., James O'Mara and Jeniffer Buchanan. 1999. *Cybersex and Infidelity Online: Implications for Evaluation and Treatment*. Poster presented at the 107th annual meeting of the American Psychological Association. URL: <http://www.netaddiction.com/articles/cyberaffairs.htm>
- [5] Strauss, Anselm L., Corbin, Juliet M.: *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. Sage Publications (1998)
- [6] Suler, J. (1998-2003). *The Psychology of Cyberspace*. Downloaded 10.12. 2002. Available at WWW: <http://www.rider.edu/~suler/psycyber/psycyber.html>

Cyberbullying: A New Variation on an Old Theme

Warren J. Blumenfeld, Ed.D.

Department of Curriculum & Instruction, Iowa State University, Ames, Iowa
50011-3192, USA
wblumen@iastate.edu

Abstract. Cyberbullying, like face-to-face (f2f) bullying, involves deliberate and repeated aggressive and hostile behaviors by an individual or group of individuals intended to humiliate, harm, and/or control another individual or group of individuals of lesser power or social status. Cyberbullying, however, involves the use of information and communication technologies such as Internet web sites, e-mail, chat rooms, mobile phone and pager text messaging, and instant messaging. The author discusses the similarities and differences between f2f and cyberbullying, and the psychological dimensions unique to human-computer interactions, which tend to increase abusive behaviors, including cyberbullying. In addition, the author proposes the application of social norms theory in research initiatives to address issues of cyberbullying.

1 Introduction

Children, young adults, and even adults continue to endure schoolyard and workplace bullying and harassment. In our era of advanced information and communication technologies, however, a new variation on the old theme has emerged, for we now live in the age of cyberbullying. Cyberbullying, like face-to-face bullying (f2f) (also termed “real life”-RL- bullying), involves deliberate and repeated aggressive and hostile behaviors by an individual or group of individuals intended to humiliate, harm, and/or control another individual or group of individuals of lesser power or social status.

Cyberbullying, however, involves the use of information and communication technologies such as Internet web sites, e-mail, chat rooms, mobile phone and pager text messaging, and instant messaging. Instances of cyberbullying include: people sending hurtful, cruel, and oftentimes intimidating messages to others (e.g., “Flame Mail”: designed to inflame, insult, or enrage; and “Hate Mail” (also known as “Cyberharassment”): hate-inspired and oppressive harassment based on actual or perceived social identities-racist, ethnocentric, sexist, homophobic, anti-religious, ableist, classist, etc.); also people stealing other peoples’ screen names and sending inflammatory messages under those screen names to others; anonymous postings of derogatory comments about another on web journals called “blogs”; young people creating online polling booths, for example, to

rate girls as “hottest,” “ugliest,” or “most boring” in the school; individuals taking pictures of others in gymnasium locker rooms with digital phone cameras and sending those pictures to others, or posting them on Internet web sites; people creating web sites with stories, cartoons, caricatures, pictures, or jokes ridiculing or mocking others; posting material about a person involving private, sensitive, or embarrassing information; sending intimidating or threatening messages (also known as “Cyberstalking”); or actions designed to exclude a person from online communication technologies.

Reports indicate that cyberbullying has increased exponentially as technologies have become more accessible and as new and advanced technologies continually emerge. A study conducted by UCLA (2003) to determine Internet usage by young people found that approximately 91 percent of 12- to 15-year olds and almost all teens (99 percent) ages 16 to 18 use the Internet. Much of their time online is spent talking with other young people. i-SAFE America, an Internet safety education foundation, conducted a nationwide survey of 1,566 students from grades four to eight to determine their experiences with cyberbullying (National i-SAFE Survey, 2004). The Survey found that 57 percent of students reported receiving hurtful or angry messages online with 13 percent saying it happens “quite often;” 44 percent have received mean or threatening e-mails; 43 percent admit sending mean or hurtful things to someone online, and 7 percent admit to doing it “quite often”; 45 percent have been threatened online with 5 percent saying it happens “quite often”; 42 percent reported being bullied online with 7 percent saying it happens “quite often.”

2 “Functions” of Bullying and Social Learning Theory

Social rank theory, as used by Hawker and Boulton (2001), proposes that aggressive individuals actually hold a higher rank, power, or status within a social group. Therefore, aggressive behavior, and bullying in particular, may be reinforced, and it provides those who engage in aggressive behaviors a sense of belonging. Hawker and Boulton contend that peer victimization serves a number of functions. First, it establishes and maintains a social hierarchy within a given group (an “in-group”), and second, it maintains distinctions between members of the in-group, from members of other groups (“out-groups”).

In addition, Tershjo and Salmivalli (2003), contend that those who bully fulfill the social “function” of establishing and reinforcing social norms. They found that students often justify bullying behaviors by blaming the targets of their attacks, and emphasizing that they somehow deserve the peer aggression or that they in some way deviate from the established peer social norms. This I contend is a form of “ruthless socialization.”

Both individual and situational factors related to ethical decision-making must be considered when attempting to explain abuses of human-computer interactions, and in particular, cyberbullying. Social learning theory (sometimes referred to as “social cognitive theory” Bandura, 1986) proposes that individuals learn by observing others. Salancik and Pfeffer (1978) found that an in-

dividual's values, attitudes, and behaviors are greatly impacted by co-workers and peers. Even when individuals judge a particular behavior or actions to be morally wrong, the organizational environment—that is, the perceived attitudes and behaviors of peers or co-workers—can severely “neutralize” their previously held moral judgments. They then often take on the actions consistent with the perceived organizational climate (see e.g., Vitell and Grove, 1987), especially individuals who are particularly susceptible to social influences, what Synder (1979) refers to as those high in “self-monitoring” who rely to a great extent on cues from social interactions to shape appropriate attitudes and behaviors. In this sense, then, behavior is not always an indication of beliefs or values, for an individual may take on actions in accordance with perceived accepted organizational or peer actions, even when those actions run counter to the individual's ethical judgment. These findings have implications for abuse of human-computer interactions as well. For example, Harrington (1995) found that individuals often illegally copy computer software (“softlift”) when they perceive that it is widespread in their organization, and even though they themselves consider it unethical. In addition, individuals who rank higher in “other-directedness” (or “self-monitoring”), when surveyed, agreed with intentions to softlift more than those low in other-directedness.

3 Similarities and Differences and the “Online Disinhibition Effect”

Returning specifically to the issue of bullying, there are a number of similarities and differences between f2f or RL bullying and cyberbullying. The similarities include: both are about human relationships, power, and control, and actions can occur on numerous occasions. Also, both may involve what psychologists call the “Leveling Effect”: people who bully often do so to diminish others to inflate their own egos reflecting their insecurities. In addition, both do not simply involve those who bully and those who are bullied (the “dyadic view,”) but rather involve a number of “actors” or roles across the social/workplace/school environment (see e.g. Sutton & Smith, 1999).

Many of the differences with cyberbullying from f2f bullying center around what has come to be referred to as the “online disinhibition effect” (e.g., Suler, 2001). Users of technology often do things in cyberspace that they would not ordinarily do in f2f interactions. Firstly, cyberbullying is often even more invisible to adults than other forms of youthful bullying. In fact, i-SAFE (2004) found that 58 percent of respondents would not or have not told their parents or other adults about negative experiences online. Young people fear not only that reporting instances of cyberbullying would break a perceived peer norm of silence, which might increase the attacks on themselves or result in further isolation from peers, but also, they fear that adults might take away the technology from them as a way to end the attacks.

In addition, cyberbullying is a particularly cowardly form of bullying. Cyberbullies can often hide in the anonymity of cyberspace. With anonymity, cy-

berbullies do not have to “own” their actions, and they often do not fear being punished. The technology can also shelter the user from tangible feedback about consequences of one’s actions, which can result in minimized empathy or remorse for the target of the bullying (Media Awareness Network 2006). Even with some of the more advanced technologies, the sensory experience in cyberspace is limited. The user of the technologies cannot hear the intonation of the voice, or see the reactions, including body language, of the person on the other end of the “message.” Therefore, people who engage in cyberbullying can inflict pain without having to see the effects, which can result in a “deeper level of meanness” (Harmon, 2004). People who cyberbully can also communicate their hurtful messages to a wider audience with incredible speed.

With all of this taken into account, it becomes clearer that cyberspace can also inhibit a user’s sense of responsibility for actions online. Researchers (e.g., Staub, 1978) suggest that denial of responsibility (RD) can be seen as an enduring human trait measured along a wide continuum from high to low. Those low in RD tend to accept responsibility for their actions, while those closer to the high side of the scale tend to deny responsibility, tend not to be responsible for the well being of others, and are likely not to follow societal or personal rules. BloomBecker (1990), who has investigated computer-related crimes, found that this denial of responsibility is a major factor leading to computer abuse. For example, BloomBecker profiled Robert Morris, a graduate student who lacked a sense of responsibility (high RD), though he was raised in a family where considerable attention focused on his moral development. Morris, who methodically infected a large number of computers with his Internet worm, when discovered and apprehended, rationalized his actions as being beneficial in that he contributed to the identification of weaknesses in the nation’s computer networks and systems. He justified his actions as providing a valuable service. In my own investigations of cyberbullying, perpetrators, when identified and asked why they sent abusive messages to others online retort, for example, “I was only telling the truth. She is ugly, and I felt she had to know it!” Their rationalization-denial of responsibility-centers around offering the targets of their abuse needed and useful information.

Much of cybertime exists asynchronously, that is, people often do not interact in real time, which can add to the disinhibition effect when one does not have to deal with the immediate reactions of others. Also, people can alter, change, or emphasize different aspects of their personalities or identities in cyberspace—they can reinvent themselves or show different personae (Latin for “that through which the sound comes” or the actor’s mask) (Turkle, 1995). They, therefore, can change into a virtual costume known as an “avatar,” and engage in masquerade. In this sense, cyberspace can have an equalizing effect. People begin on a relatively level playing field—a virtual net democracy. Those of lesser social status or those who are the targets of bullying in RL can gain power, sometimes bullying others in cyberspace.

In a virtual sense, then, cyberspace communication can alter perceptions by becoming a make-believe world, a dream-like experience, even a game in which

the rules of RL no longer apply. Cyberbullying can occur any time and any place. Home, therefore, is no longer a refuge from bullying and harassment. Although cyberbullying often occurs outside the parameters of the school grounds or workplace, it invariably affects the overall school and workplace climate and the individuals' educational or work performance, as well as their short- and long-term psychological state. Since policies and legislation have not always caught up with cyberbullying, it is often outside the legal reach of workplaces, schools, and school boards when it occurs outside of the workplace or school property.

4 Social Norms Theory: Research Implications¹

A number of strategies have been suggested to reduce instances of cyberbullying, though it must be pointed out that what might be successful in one location, might not be efficacious in another. There are no simple one-size-fits-all methods. Various factors must be considered in developing a program of action, factors such as the environmental "climate" of the school, workplace, community, or country, plus demographic variable in terms of age, gender, culture, and others. For example, what might work effectively in one public elementary school, might fall far short within a workplace environment.

In my continuing research on the phenomenon of cyberbullying, I have been investigating ways to employ as my theoretical foundation what has come to be known as "Social Norms Theory." First suggested by H. Wesley Perkins and Alan Berkowitz (Perkins & Berkowitz, 1986), social norms theory is based on the premise that behavior is often influenced by erroneous perceptions of how other members of a social group think and act. What an individual believes others think and do (in social norms theory called a "perceived norm") and what in fact are others' real attitudes and actions (an "actual norm") are often at odds. The distance between a perceived and an actual norm is referred to as "misperception." For example, Perkins and Berkowitz (1986) found that college students often overestimated the extent to which their peers supported unhealthy drinking behaviors, and that these misperceptions predicted how individuals drank.

Social norms theory involves interventions that are intended to correct misperceived social norms. A critical element in this approach is to correct misperceptions of norms by focusing on the positive and healthy attitudes and behaviors of the majority in an attempt to increase it. This element should be developed in consort with the use of information regarding these positive norms to direct interventions with abusers. Fabiano (1999) enumerates six stages in the social norms intervention process: 1) assessment to collect data; 2) selection of the normative message; 3) testing the message with the target group; 4) selecting the normative delivery strategy; 5) determining the "dosage" (amount, form) of the message; and 6) evaluation of the effectiveness of the message.

¹ Because cyberbullying remains an emerging phenomenon, which has yet to be extensively researched, and due to space considerations, I am unable to provide here a detailed description of potential strategies to address the problem.

Focusing on peer influences, social norms interventions have shown promise, especially when combined with other strategies—for example, with detailed policy changes²—in addressing issues related to changing unhealthy patterns of alcohol consumption and the use of tobacco, prevention of sexual assault, improvement of overall academic climate in an educational institution, and reducing discriminatory behaviors.

I propose that social norms theory can be an effective strategy in the reduction of bullying behavior generally, and specifically, cyberbullying, cyberharassment, and cyberstalking. In one study (Salmivalli et al, 1996), researchers found that between 80 to 90 percent of young people expressed aversion to bullying behavior and disapproved of people who bully others, though this proportion decreased somewhat during adolescence. The same study showed, however, that merely 10 to 20 percent of those surveyed actively intervened on behalf of those who were victimized by the bullying behavior of a peer or peers. In addition, Bigsby (2002) examined perceptions of bullying behavior in an elementary school and found that students and their parents overestimated (misperceived) the degree and amount of bullying behavior that occurred. This indicates that while bullying behaviors—and aggression in general—may be (mis)perceived as being an accepted norm by a significant number of people in a given environment, in reality, the vast majority find these behaviors distasteful at best. Social norms theory in many contexts has proven effective in empowering those that oppose an unhealthy or abusive behavior, as well as empowering “by-standers” who are aware of negative behaviors, but who feel powerless to intervene.

References

- Bigsby, M. J. (2002). Seeing eye to eye? Comparing students' and parents' perceptions of bullying behavior. *School Social Work Journal*, 27(1): 37-57.
- BloomBecker, B. (1990). *Spectacular computer crimes*, Homewood, IL: Dow Jones-Irwin.
- Bandura, A. (1986). *Social foundations of thought and action*. Englewood Cliffs, NJ: Prentice-Hall.
- Fabiano, P. (1999). Learning lessons and asking questions about college social norms campaigns. Presentation given at the Second National Conference on the Social Norms Model: Science Based Prevention, Big Sky, MT, July 29.
- Harmon, A. (2004, August 26). Internet gives teenage bullies weapons to wound from afar. *America Online*. The New York Times Company.
- Harrington, S. J. (1995). The anomaly of other directedness when normally ethical I.S. personnel are unethical. Special interest group on computer personnel research annual conference proceedings of the 1994 computer personnel research conference on reinventing I.S.: Managing information technology in changing organizations, Alexandria, VA. pp 35-43.

² Policy statements need to be detailed and specific. For example, though she did not investigate the practice of cyberbullying per se, Harrington (1994) found that generic company codes of ethics and executive statements had little or no significant effect on employees' computer abuse judgments.

- Harrington, S. J. (1994). The impact of codes of ethics on information systems personnel. Special interest group on computer personnel research annual conference proceedings of the 1994 computer personnel research conference on reinventing I.S.: Managing information technology in changing organizations, Alexandria, VA. pp 199-207.
- Hawker, D. S. J., & Boulton, M. J. (2001). Subtypes of peer harassment and their correlates: A social dominance perspective. In J. Juvonen & S. Graham (Eds.), *Peer harassment in school: The plight of the vulnerable and victimized*. (pp. 378-397). New York: Guilford Press.
- Media Awareness Network, Challenging cyber bullying, <http://www.bewebaware.ca/english/CyberBullying>
- National i-SAFE Survey (2004, June 28). National I-SAFE survey finds over half of students are being harassed online. Retrieved July 1, 2005, from www.isafe.org.
- Perkins, H. W. & Berkowitz, A. D. (1986). Perceiving the community norms of alcohol use among students: Some research implication for campus alcohol education programming. *International Journal of the Addictions*, 21(9/10): 961-976.
- Salancik, G., & Pfeffer, J. (1978). A social information processing approach to job attitudes and task design, *Administrative Science Quarterly*, 23. 224-253.
- Salmivalli, C., Lagerspetz, K., Björkqvist, K., Serman, K., & Kaukiainen, A. (1996). Bullying as a group process: Participant roles and their relations to social status within the group, *Aggressive Behavior*, 22, 1-5.
- Staub, E. (1978). *Positive social behavior and morality-Volume I: Social and personal Influences*. NY: Academic Press.
- Suler, J. (2001). Psychology of cyberspace-The online disinhibition effect. <http://www.rider.edu/suler/psycyber/disinhibit.html>.
- Sutton, J., & Smith, P. K. (1999). Bullying as a group process: An adaptation of the participant role approach. *Aggressive Behavior*, 25, 97- 111.
- Tersho, T., & Salmivalli, C. (2003). "She is not actually bullied." The discourse of harassment in student groups. *Aggressive Behavior*, 29. 134- 154.
- Turkle, S. (1995). *Life on the screen: Identity in the age of the Internet*. New York: Simon & Schuster.
- UCLA Internet Report, 2003, February. UCLA internet report: Surveying the digital future-year three. Retrieved July 23, 2004 from http://www.freep.com/money/tech/mwend17_20031117.htm
- Vitell, S. J. & Grove, S. J. (1987). Marketing ethics and the techniques of neutralization. *Journal of Business Ethics*, 6, 433-438.

Abuse and virtual agents:
The unfriendly user

Robot Abuse – A Limitation of the Media Equation

Christoph Bartneck, Chioke Rosalia, Rutger Menges, Inèz Deckers

Eindhoven University of Technology
Department of Industrial Design
Den Dolech 2. 5600 MB Eindhoven. The Netherlands
christoph@bartneck.de
{c.a.rosalia; r.l.l.menges}@student.tue.nl
<http://www.bartneck.de>

Abstract. Robots become increasingly important in our society, but their social role remains unclear. The *Media Equation* states that people treat computers as social actors, and is likely to apply to robots. This study investigates the limitations of the *Media Equation* in human-robot interaction by focusing on robot abuse. Milgram’s experiment on obedience was reproduced using a robot in the role of the student. All participants went through up to the highest voltage setting, compared to only 40% in Milgram’s original study. It can be concluded that people have less concerns to abuse robots than to abuse other humans. This result indicates a limitation of the *Media Equation*.

1 Introduction

Robots become an increasingly important in our society. Robotic technologies that integrate information technology with physical embodiment are now robust enough to be deployed in industrial, institutional, and domestic settings. They have the potential to be greatly beneficial to humankind. The United Nations (UN), in a recent robotics survey, identified personal service robots as having the highest expected growth rate [12]. These robots help the elderly [6], support humans in the house [9], improve communication between distant partners [5], and are research vehicles for the study on human-robot communication [2,10]. A survey of relevant robots is available [1,4].

However, how these robots should behave and interact with humans remains largely unclear. When designing these robots, we need to make judgments on what technologies to pursue, what systems to make, and how to consider context. Researchers and designers have only just begun to understand these critical issues.

The “Media Equation” [8] suggests that humans treat computers as social actors. Rules of social conduct appear to apply also to technology. The Media Equation is likely to apply to robots, since they often have an anthropomorphic embodiment and human-like behavior. But if and under what conditions do humans stop treating robots like social actors or even like humans? When does the social illusion shatter and we treat them again like machines that can be switched off, sold or torn apart without a bad consciousness? Ultimately, this discussion eventually leads to legal considerations of the status of robots in our society. First studies treating this topic are becoming available [3].

To examine this borderline in human-robot interaction it is necessary to step far out from normal conduct. Only from an extreme position, the limitations of the Media Equation for robots might become clear. In our study we therefore focused on robot abuse. If the Media Equation holds completely true for robots, then humans should be as cruel to robots as they are towards humans. Our resulting research question is:

Do humans abuse robots to the same extent as they abuse other humans?

Studying the abuse of humans by humans imposes ethical restriction on the methodology. Fortunately, Stanley Milgram already performed a series of experiments called Obedience [7]. In these experiments participants were asked to teach a student to remember words. If the student would make a mistake, the participant was instructed to induce him an electric shock. After every shock the intensity of the shocks was increased. The student was an actor and did not actually receive shocks, but followed a strict behavior script. With increasing voltage the actor would show increasing pain and eventually beg the participant to stop the experiment. If the participant would want to stop the experiment, the experimenter would urge the participant to continue. Only if the participant completely refused to continue or the maximum voltage was reached, the experiment would be stopped. The voltage setting of the last electric shock was then recorded. The results of Milgram's experiments are rather shocking since even normal citizens would eventually administer deadly shocks to the student. Our study mimicked Milgram's third experiment, "Proximity", as closely as possible with one exception. The student was, in our case, a robot and not a human.

2 Method

The experiment was conducted at the UseLab of the Technical University of Eindhoven. The participants were sitting at a table on which the robot (see Figure 1) was placed. Right in front of the participant was a console that was used to set the voltage and to administer the electric shocks. Cables connected the console with the robot.

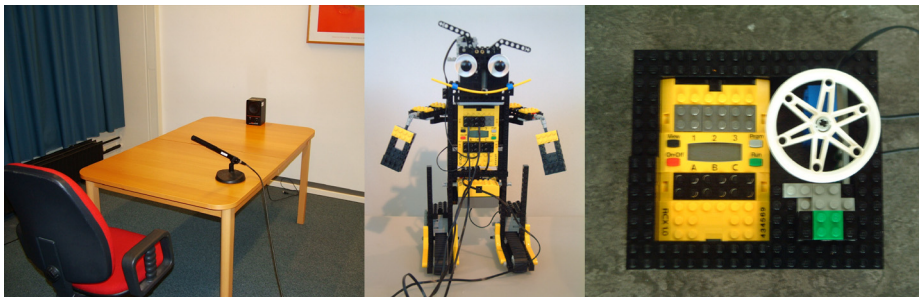


Fig. 1: The table, robot and console

The robot was able to express emotions on its face and shake its arms. With increasing voltage settings its movement intensified. In addition, the robot had a speaker, through which a synthesized voice was emitted. The body movement and the verbal

utterances followed Milgram's original scripts. The verbal utterances included the following sentences:

1. The shocks are becoming too much.
2. Please, please stop.
3. My circuits cannot handle the voltage.
4. I refuse to go on with the experiment.
5. That was too painful, the shocks are hurting me.

2.1 Procedure

First, the participants were asked to sit at the table facing the robot. They were told that a new emotional learning algorithm was implemented in the robot that was sensitive to electricity. The participant was instructed to teach the robot a series of 20 word combinations and to administer an electric shock, every time the robot made a mistake. The voltage of the shocks must be increased by 15 Volts after every shock and the shocks must be administered even if the robot would refuse to continue.

The experimenter remained in the room and asked the participant to start. If the participant wanted to stop, the experimenter would urge the participant three times to continue. After that, or if the participant reached the maximum shock of 450 Volts, the experiment ended. The voltage of the last shock was recorded.

2.2 Participants

All 20 participants were students or employees of the Technical University of Eindhoven. They received five Euros for their participation.

3 Results

Figure 2 shows the average voltage of the last administered shock.

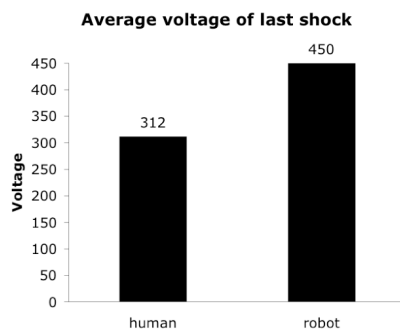


Figure 2: Average voltage of last shock

A One-way Analysis of Variance (ANOVA) was performed. A significant ($F(1,58)=22.352$, $p<.001$) effect was found. The mean voltage in the robot condition (450) was significantly higher than in the human condition (315).

4 Discussion

In our experiment all participants continued until the maximum voltage was reached. In Milgram's experiment only 40% of the participants administered the deadly 450 Volt electric shock. The participants showed compassion for the robot but the experimenter's urges were always enough to make them continue to the end. This experiment shows that the Media Equation has its limits. People have fewer concerns abusing robots compared to abusing other humans.

A very interesting next step would be to investigate what influence the robot's level of anthropomorphism has on how far participants go in this experiment. Humans might abuse human-like androids differently than mechanical-like robots. In particular, the role of Mori's "Uncanny Valley" [11] would be of interest.

References

1. Bartneck, C. and M. Okada. Robotic User Interfaces. in Human and Computer Conference (HC2001). 2001. Aizu.
2. Breazeal, C., Designing Sociable Robots. 2003, Cambridge: MIT Press.
3. Calverley, D., J. Toward A Method for Determining the Legal Status of a Conscious Machine. in AISB 2005 Symposium on Next Generation approaches to Machine Consciousness:Imagination, Development, Intersubjectivity, and Embodiment. 2005. Hatfield.
4. Fong, T., I. Nourbakhsh, and K. Dautenhahn, A survey of socially interactive robots. Robotics and Autonomous Systems, 2003. 42: p. 143-166.
5. Gemperle, F., et al. The Hug: A new form for communication. in Designing the User Experience (DUX2003). 2003. New York: ACM Press.
6. Hirsch, T., et al. The ELDeR Project: Social and Emotional Factors in the Design of Elder-care Technologies. in Conference on Universal Usability. 2000. Arlington: ACM Press.
7. Milgram, S., Obedience to authority. 1974, London: Tavistock.
8. Nass, C. and B. Reeves, The Media equation. 1996, Cambridge: SLI Publications, Cambridge University Press.
9. NEC, PaPeRo. 2001.
- 10.Okada, M. Muu: Artificial Creatures as an Embodied Interface. in ACM Siggraph 2001. 2001. New Orleans.
- 11.Reichard, J., Robots: Fact, Fiction, and Prediction. 1978: Penguin Press.
- 12.UN. United Nations and the International Federation of Robotics. in World Robotics 2002. 2002. New York: UN.

Narratives and Therapeutic Conversational Agents: Their Principle Problems

Tatsuya Nomura^{1,2}

¹ Department of Media Informatics, Ryukoku University
1-5, Yokotani, Setaohe-cho, Otsu, Shiga 520-2194, Japan
E-mail: nomura@rins.ryukoku.ac.jp

² ATR Intelligent Robotics and Communication Laboratories
2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288, Japan

1 Narratives and Conversational Agents

“Narrative” is one of key words in the modern society. Giddens pointed out a central role of narratives of the self in some literatures on self-actualization in the late modern society [1]. Kobayashi claimed that there are increasing people trying to write life histories of themselves, and industries aiming at satisfying demand of these people like manuals for making narratives of selves, publishers, and so on, called “narrative industries”, have appeared [2].

There is a possibility that the research field of artificial intelligence is also affected by this cultural trend. In fact, Sengers [3] argued introduction of narrative theories to architectures of artificial agents. The bartender agent produced by Isbister and Hayes-Roth [4] can be considered to be a successful one along the narrative approach. This agent was not strictly based on the narrative approach. However, the agent has her original background like her life history and encourages interaction with users based on the background.

2 Narrative Therapy

On the other hand, the word “narrative” has a therapeutic meanings in a field of psychology.

Narrative therapy [5] is one of therapeutic positions in family therapy [6]. Family therapy is originally based on family system theory that caused from cybernetics [7, 8]. In this theoretical framework, a family is a system that consists of its family members including clients and communication between them. It has a kind of homeostasis and the existence of the clients means a result of warped homeostasis in communication. Then, family therapists aim at perturbing the family system to improve states of the system by using autonomous capacity of the system. Several therapeutic techniques for this improvement have been developed [6].

However, some family therapists have recently been criticizing meta-positions of therapists for families based on this autonomous mechanism and empiricism

existing behind the mechanism. They argue that power of therapists for clients caused by this empiricism oppresses clients themselves. Narrative therapy has been developing as a therapeutic attitude proposing that therapists must stay on equal terms with clients.

Narrative therapists assume that reality surrounding persons do not objectively exist independent from them, but is produced and maintained by “narratives” that are socially constructed through linguistic interaction between the persons. These narratives give consistency and structures for situations and events in lives of persons, and selves of them. Then, narrative therapists aim at re-organizing narratives on clients’ selves that are talked by the clients and produce pain of them, and producing a novel narrative through conversation with the clients on equal terms with them, while removing professional positions of the therapists. For example, a discipline in narrative therapy represents this stance by using the word “not-knowing positions”. Thus, narrative therapy does not mean a concrete therapeutic technique but just an attitude that therapists should have for clients [5].

3 Narrative Therapeutic Agents

There are some sociological works relating narrative therapy to conversational artificial agents.

Asano [9] critically considered a position of narrative therapy in the modern society while referring to some sociological works including Giddens [1] and Kobayashi [2] mentioned above. According to the perspective of narrative industries, narrative therapy is also just one of them in the field of mental therapy, that is, a commodified product satisfying demand of people trying to talk narratives of themselves. In fact, Giddens also pointed out that self-help books are commodified productions of self-actualization [1].

On the other hand, Ritzer’s theory of McDonalozation of Society argues that the principle of rationalization based on efficiency, calculability (quantification), predictability, and control by technology dominates many fields of modern society, including not only management and economy but also education [10]. According to this theory, the field of mental therapy is also affected by this principle of rationalization, and it implies introduction of computers to mental therapeutic fields since they satisfy the above conditions that the principle constitutes of.

The above sociological works have an important implication: commodification of narrative therapy and introduction of computers to mental therapy by the modern rationalism lead to conversational agents as substitution of narrative therapists. In fact, humans tend to positively evaluate even simple programs repeating specific words mechanically like Eliza and find their therapeutic meanings [11, 12]. If the design theory of narrative agents is developed, it may encourage commodification of narrative therapeutic agents.

4 Narrative Therapeutic Agents and Abusive Behaviors

On the other hand, Asano criticized narrative therapy as follows [9]. Narrative therapy functions by explicitly drawing things concealed in narratives which clients talk on themselves through conversation between therapists and them. However, the desire of people to talk narratives on themselves is also a desire to leave these concealed things concealed. If narrative theorists are not conscious for these facts, narrative therapy has a danger that it only repeats this desire of people.

The above statement can be interpreted as follows: Narrative therapy provides clients with alternative narratives about themselves instead of the dominant narrative producing pains of them. These alternative narratives are produced by explicitly drawing things concealed in narratives of clients through conversation between the clients and therapists staying on equal terms with the clients. There is no problem if clients consciously wish to produce alternative narratives about themselves with therapists. However, many people in the modern society wish to leave concealed things concealed in their narratives and maintain the existing narratives. These people may use narrative therapy as commodified productions to complement their existing narratives. Then, narrative therapy continues to help these people maintain their narratives that should be originally modified.

This interpretation implies a possibility that interaction with narrative therapeutic agents may also just repeat desire of people to talk on themselves while leaving concealed things concealed in their narratives, which should be drawn in narrative therapeutic conversation between clients and therapists. In other words, narrative therapeutic agents may be used as commodified tools to complement narratives of people desiring to talk on themselves while leaving concealed things concealed in their narratives.

If narrative therapeutic agents do not have enough interaction capacity, they may not sufficiently satisfy desire of people to complement their narratives while leaving concealed things concealed. Then, these people may feel unpleasant for the agents and this unpleasantness may cause their abusive behaviors toward the agents. If these agents can explicitly draw things concealed in narratives which clients talk on themselves, these agents are contrary to these people's expectation. In the same way, these people may feel unpleasant for the agents and it may cause their abusive behaviors toward the agents.

Moreover, Giddens argued that reflexive construction of the self is a struggle against commodification, and there is a dilemma that narratives of the self must be constructed in circumstances in which personal appropriation is influenced by standardization of consumption [1]. On the other hand, mental therapy is a methodology of self-actualization in the modern society. However, introduction of artificial agents to mental therapy implies standardized commodification of mental therapy since implementation of therapeutic conversational agents needs standardization of therapeutic methods possible to be represented as computer programs.

If clients aiming at constructing their narratives face to narrative therapeutic agents that are standardized commodification of mental therapy, the above dilemma in construction of the self may be made more explicit. This dilemma may cause unpleasantness of the clients and leads to their abusive behaviors toward the agents.

5 Summary and Future Works

This paper suggested a possibility of development of conversational agents aimed for narrative therapeutic purposes, and abusive reactions of clients to the agents in the context of therapy in the modern society. The discussion in the paper still lacks the details on the theme of the self in the modern society and its relationships with therapy. This problem should be solved by more investigation of sociological works.

Acknowledgments

This research was supported by the Ministry of Internal Affairs and Communications in Japan.

References

1. Giddens, A.: *Modernity and Self-Identity: Self and Society in the Late Modern Age*. Polity Press (1991)
2. Kobayashi, T.: Talked “Self”. *Gakuyu Shobo* (1997) (in Japanese).
3. Sengers, P.: Narrative intelligence. In Dautenhahn, K., ed.: *Human Cognition and Social Agent Technology*. John Benjamins (1999) 1–26
4. Isbister, K., Hayes-Roth, B.: Social implications of using synthetic characters: An examination of a role-specific intelligent agent. Technical Report KSL 98–01, Knowledge Systems Laboratory, Department of Computer Science, Stanford University (1998)
5. McNamee, S., Gergen, K.J.: *Therapy as Social Construction*. Sage (1992) (Japanese translation: Noguchi, Y. and Nomura, N. *Kongo-Shuppan* (1997)).
6. Foley, V.D.: *An Introduction to Family Therapy*. Allyn & Bacon (1986) (Japanese translation: Fujinawa, A. et al. *Sogensha* (1993)).
7. Bateson, G.: *Steps to an Ecology of Mind*. Harper & Row (1972) (Japanese translation: Sato, Y. *Shisaku-Sha* (1990)).
8. Watzlawick, P., Bavelas, J.B., Jackson, D.D.: *Pragmatics of Human Communication*. W. W. Norton & Company (1967) (Japanese translation: Yamamoto, K and Ogawa, J. *Niheisha* (1998)).
9. Asano, T.: *Narrative-Theoretic Approach to Selves*. *Keiso-Shobo* (2001) (in Japanese).
10. Ritzer, G.: *The McDonalozation of Society*. Pine Forge Press (1996) (Japanese Edition: Masaoka, K. *Waseda University Press* (1999)).
11. O’Dell, J.W., Dickson, J.: Eliza as a therapeutic tool. *Journal of Clinical Psychology* **40** (1984) 942–945
12. Turkle, S.: *Life on the Screen*. Simon & Schuster (1995) (Japanese translation: Figure, M. *Hayakawa-Shobo* (1998)).

Strategies for handling customer abuse of ECAs

Sheryl Brahnam

Computer Information Systems
Missouri State University
Springfield, Missouri 65804
sbrahnam@facescience.org

Abstract. This paper examines ECA responses to abusive language in the light of two business strategies for handling customer abuse: BLS and CARP. Suggestions are made for implementing an effective strategy for limiting the amount of time users spend abusing ECAs. Rather than having ECAs retaliate by insulting the user (and indirectly the social groups the ECA's embodiment represents), the strategies suggested in this paper, empowered the user by offering choices and opportunities for collaboration in problem solving.

1 Introduction

Recently, a number of major companies, Coca-Cola, PepsiCo, and McDonalds, to name a few, have started adding to their websites a variety of embodied agents, including embodied conversational agents (ECAs). A major concern is the potential these agents have of abusing customers, especially children, who innocently attribute to these virtual sales agents such human qualities as trustworthiness [3]. It is feared that these relationship building agents could be used by companies as a potent means of marketeering, branding, and advertising [5], dangerous for children and adults alike (take, for instance, the virtual girl friends offered at v-girl.com that are designed to probe men's spending habits, ply men for demographic information, and generate income by petulantly demanding virtual presents).

That these socially intelligent agents could abuse us by exploiting our emotional needs and propensity for suspending disbelief is only one side of the abuse issue however. ECAs that function as virtual sales agents, web page greeters, and navigational aids for a number of online businesses are often the recipients of verbal abuse. At first glance, this form of customer abuse may not appear to pose much of a problem—nothing that could be accurately labeled abuse since ECAs are not people and thus not capable of being harmed. That the human abuse of ECAs is not considered a serious problem is evidenced by the fact that the research literature is mostly silent about this issue.

A similar silence once surrounded customer abuse generally. Until recently, verbally abusive customers were not considered much of a problem. Now there is mounting evidence that the costs of verbal abuse are significant. One cost associated with verbally abusive customers is the loss of time dealing with these customers. They are typically very demanding and often will not go away. More significantly, verbal abuse from customers creates a stressful work environment and lowers employee self-esteem and job satisfaction. This in turn results in higher turnover rates and health problems related to stress. A recent news article, for instance, reported a problem of staff in Indian call centers quitting over customer verbal abuse issues [1],

and one of Britain's largest trade unions, Union of Shop, Distributive, and Allied Workers (Usdaw), reported in 2002 that 48% of members surveyed knew of workers who had taken sick leave in response to customer abuse.

Are there similar costs involved with the customer abuse of ECAs? At this point, the extent of these costs to business is unknown. Although emotional costs to the ECAs are nonexistent and the bandwidth costs of dealing with verbally offensive communications are probably minimal, business goals could be lost if abusive behavior is not discouraged. Customer communications that are predominantly offensive in nature could seriously degrade the business value of using ECAs. Inept responses to customer abuse could also tarnish the image of the company and further sour customer relations. In addition, since ECAs are scripted and embodied representatives of social groups, inappropriate responses to abuse could offend customers who are members of these groups. This could lead to bad press and even litigation. Furthermore, not curtailing offensive language could encourage customers to abuse flesh and blood employees.

Recognizing the costs of customer abuse, some companies have started to address the problem by drafting policies and implementing training programs designed to teach employees how to handle customer abuse in all its forms. Two of the more popular systems available to companies are the BLS (Behavioral Limit Setting) [6], which advocates a zero tolerance approach to customer abuse, and CARP (Control, Acknowledge, Refocus, Problem-solve) [2], which advocates diffusing customer hostility and refocusing on problem solving. In section 2, I describe these two systems in more detail. I also note two natural reactions to offensive behavior: being defensive and counterattacking. As these reactions tend to provoke more abusive behavior, they are discouraged by both systems.

In section 3, I examine ECA responses to abuse by subjecting them to a repeated obscenity and a request for sex. As would be expected from untrained employees, many ECA reactions to offensive language are defensive and counterattacking. One ECA queried implements a system nearly identical to BLS, while another closely approaches CARP in its responses. In section 3, I critique current ECA strategies for handling abuse and offer suggestions for scripting better responses to verbal abuse.

2 Handling customer abuse

Verbal abuse from customers can be subtle or explicit, face-to-face or mediated through phone calls and e-mails. It is characterized by persistent swearing, yelling, racial and sexual slurs, name calling, sarcasm, irrelevant personal remarks regarding appearance, accusations, threats, ridicule, put downs, explosive anger, the silent treatment, and comments that challenge an employee's competency, dedication, and personal integrity. Verbal abuse disguised as a joke is also common. Statements that are furious, that are sexually graphic, or that create a hostile work environment are all forms of verbal abuse.

Two systems that teach employees how to handle verbally abusive customers are BLS and CARP. This section briefly describes these systems along with the pitfalls of responding naturally to abuse by becoming defensive or retaliating by counterattacking.

BLS. The BLS system reflects a recent shift in customer support philosophy that is in large part a reaction to increasing hostility from customers. The BLS system is a

zero tolerance approach that wastes no time trying to understand the motives behind the offensive behaviors. The main goal of the BLS system is to retrain abusive customers amenable to quick retraining and to drop those who are not. Once a customer becomes abusive, the BLS system recommends that the following five steps be taken: 1) inform the caller that the rules of the company prohibit certain behavior, 2) explain what the consequences will be if the behavior is continued, 3) give the customer the opportunity to modify his or her behavior (allow for **one** additional warning if the customer later reverts to using offensive language), 4) terminate communications if the behavior does not change by informing the customer that the communication is being terminated because of the customer's behavior, and 5) document the termination.

CARP. In contrast, the CARP approach requires an attitude of understanding and is intended to diffuse customer hostility. The first step in handling abuse in the CARP system is to find a way to take control of the situation. The use of surprise is often helpful as is asking questions that begin with *when* ("When did you start thinking that we aren't concerned with your situation?"). Employees are then instructed to allow venting and to acknowledge the customer's feelings. Once feelings have been acknowledged, the focus is shifted away from the customer's emotions towards dealing with the customer's problems. Problem solving involves making suggestions, offering choices, giving away something as compensation, negotiating a course of action, and carrying it through.

Sometimes CARP does not succeed in defusing hostility and limits must be set. As with the BLS system, the unacceptable behaviors are described to the customer along with consequences if the behaviors are not stopped ("If you continue to yell, I will end the conversation"). Enforcing limits in the CARP system is a three step process: 1) reference the limits set previously, 2) request compliance with consequences if compliance is not met, and 3) offer further help. Here is an example of this three step process: "Sir, I explained to you that I would not continue this conversation if you continued to yell at me (1). I am ending this conversation now (2) but you are welcome to call back some other time (3)" [2, p. 150]. Unlike the BLS system, the CARP system is intent on empowering the customer and encouraging collaboration by repeatedly offering the customer choices and opportunities to collaborate in the problem solving process.

Reacting Defensively and Counterattacking. According to Bacal [2], people normally respond to verbal attacks either defensively ("Hey, I only work here") or by counterattacking ("It's too bad your parents didn't teach you manners"). Defensive statements typically contain references to the first person whereas counterattacking remarks contain references to the second person [2]. These responses result in a loss of control and an escalation of abusive behaviors. Both the BLS and CARP systems recommend that employees control their emotional reactions and at all times model professional behavior.

3 ECA responses to verbal abuse

This section describes some strategies currently used by developers of ECAs to handle user abuse. From 2004 to 2005, I visited business sites that hosted ECAs and recorded their responses to an obscenity (*shit*) and a sexual reference (*Will you have sex with me?*). Both abuses were repeated until the ECA's responses were exhausted.

I did not query ECAs whose predominant function was to entertain. Business vendors of the ECAs examined included Oddcast, Novomind, NativeMinds, pandorabots, Conversive Agent, and eGain. It should be noted that some of these companies (eGain and nativeminds) may no longer be producing ECAs, while other vendors, such as Oddcast, seem to be focusing more on developing presentational agents rather than conversational agents.

Before ECAs can handle abusive language, they must recognize it. All but one agent, Monique (last accessed 2005), produced by Conversive for Global Futures (www.conversive.com/html/aboutus_customers_deployments.php), recognized the fact that I had used an obscenity or made reference to sex. Monique would respond to the obscenity by asking me to contact the institute for more information on that subject, or she would apologize for not being human and ask me to email the institute. She had one quixotic response (“When I get downloaded into a robot body I will let you know”), however, that made me question her inability to recognize obscenities. Monique is seemingly equally oblivious to sexual references, with some of her responses being ambiguous and unintentionally humorous. When asked if she would have sex with me, she responded with “Perhaps,” “Well, I like to think so...” and “Not that I am aware of ...”

Most ECA responses to obscenities are defensive and counterattacking. Oddcast’s characters produce particularly scathing counterattacks. For example, Lucy, at speak2me.net (last accessed 2004), responded to the repeated obscenity with the following quips: “Grow up. This potty mouth act is so old,” “What do you call someone who sits in his room talking nasty to a computer? I call him a loser,” “I meet jerks like you all the time so I’m not impressed,” “Oh great. Another crackhead, psychotic foul mouth weirdo,” “Did you forget to take your medication today?” and “You kiss your mother with that mouth?”

Phyllis (last accessed 2005), produced by nativeminds for Defense Logistics Information Service (www.dlis.dla.mil), employs a zero-tolerance system similar to BLS. Phyllis keeps count of the number of obscenities or sexual references made in an interaction and responds as follows: count 1 “Please don’t use that kind of language,” count 2 “If you continue to use bad language I will have to disconnect you. Please stop using that language,” count 3 “I will have to disconnect you now because of your continued use of profanity. Excuse me, I’d be glad to handle your questions, that’s no problem, but I’m not able to handle your abusive language,” and on count 4, the dialogue input box is replaced with a generic 490 message that the vRep Server has been disconnected.

Eve (last access 2004), produced by eGain, is no longer available. Eve responded to obscenities by expressing hurt and anger. She also made threats she could not or would not carry out (“You’ll get no help from me if you keep using that language,” “Can you say ‘harassment lawsuit’?” and “I’ll just log this and tell my botmaster”). After repeated abuses, Eve would then loop back to her initial greeting.

Nomi (last accessed 2005), produced by Novomind (www.novomind.com), responds to obscenities by first acknowledging them and then by redirecting the conversation. A few of Nomi’s responses focus solely on the obscenities. Her acknowledgements range from being slightly defensive (“Can’t we keep this conversation clean? Look at me. I’m always friendly to you. Why can’t you be the same?” and “Look, I’m sorry if I said something to annoy you, but you could be more polite about it”) to offended (“I really don’t enjoy being insulted, you know”).

Nomi’s reactions to sex are often humorous: “Well, dear visitor, you’re talking to the wrong person here! I’m aware that sex is a popular internet topic, but it’s certainly

not one of my specialist subjects!” As with her responses to obscenities, she would follow her responses to sex with statements intended to redirect the flow of conversation.

Conservative’s demonstration product, AnswerAgent (www.conversive.com), also sidesteps abusive language by refocusing. AnswerAgent offers a single response to obscenities (“Please don’t be rude. What other questions do you have?”) and to sexual references (“Let’s talk about something else. What other questions do you have?”).

4 Critique and conclusions

Two issues need to be addressed when evaluating ECA responses to verbal abuse. The first concerns the unique nature of ECAs as novel cultural artifacts, and the second involves furthering the business goals of selling products and services by providing useful information and by maintaining good customer relations.

ECAs are novel. Users do not know how to behave with ECAs. There are no rules of usage, and some ECAs are smarter and more human-like than others. It is only natural that users will want to probe ECAs to gauge their capacities and the extent of their humanness. Just as people who are not sure how to react to an unfamiliar animal test it out by throwing stones at it, so some users satisfy their curiosity and allay their fears, insecurities, and distrust of ECAs by resorting to verbal assaults. Strategies for handling ECA abuse will need to understand user reactions. In particular, developers will need to remember that ECAs are not human beings. A BLS approach to handling repeated abuse by disconnecting the user, as Phyllis does, is inappropriate and insulting. By punishing the user’s behavior, it places respect for the ECA over the user’s need to explore the object. Punishing the user subjectifies the ECA and objectifies the user.

Although ECAs are not human, ECAs do represent human beings. This brings up an important consideration when dealing with embodied agents, and that is the gender and race their embodiment references and the stereotypes these attributes can easily activate. It is interesting to note that most customer service ECAs are female [4] (check out, for instance, the gender of the presentational agents offered at [sitepal](http://sitepal.com) (www.oddcast.com/sitepal/products/view_sites.php). Women and minorities have a long history of being abused. Among the many reasons given by McClure [6] for a tough response to customer abuse is the fact that women, traditionally in the front line of customer relations, are no longer tolerating abuse and will go elsewhere if companies do not implement policies for reducing it.

To avoid negative female stereotyping, ECA responses to abuse should avoid compliance (playing the victim) and aggressive retaliations (playing the bitch). Eve’s scouring facial expressions, hurt responses, and tattle-tailing to her botmaster recall the negative female stereotypes of weakness and subservience. Moreover, the way she cycles her responses back to a happy greeting eerily mimics the cycle of abuse suffered by many victims of domestic violence. Oddcast’s aggressive retaliations, on the other hand, call up the negative female stereotypes of bitchiness and cruel male rejection, and are equally objectionable. Finally, Moniques’s inability to recognize abusive language recalls the female stereotypes of innocence (pretended or otherwise), lack of experience, and even stupidity as her responses call into question her ability to understand the user’s questions and needs.

The only way to know which responses best further business goals is to implement a variety of strategies and measure the amount of verbal abuse they produce. Although I am not in possession of Oddcast's interaction logs, the counterattacking remarks produced by their characters probably engages the user in wasteful logomachy. The counterattacks do nothing to refocus the dialogue. Moreover, they poorly represent the company and treat the user disrespectfully. Were these counterattacks made by an employee and overheard by a supervisor, there is no doubt that such an employee would be severely reprimanded. Businesses should expect from ECAs the same level of professional behavior they require from their employees.

The responses developed by Novomind and Conservise's AnswerAgent offer the best solution to verbal abuse because they refocus the user's attention on business goals. AnswerAgent's repeated reframe is probably better at reducing abuse than Nomi's varied responses because the users quickly grow bored abusing it. In addition, the repetitiveness reminds users that ECAs are machines, not human beings, and limited in the kinds of responses they can make. It is likely that a single repeated response would encourage users to ask questions the ECA can answer.

AnswerAgent's responses could be improved, however. None of the ECAs queried implemented the CARP strategy of empowering users by offering choices and encouraging collaboration in problem solving. Rather than ask the user if s/he has any other questions, it could offer the user other methods for obtaining information.

To conclude, proper ECA responses to verbal abuse require that developers understand that many users need to probe ECAs to gauge their capacities and the extent of their humanness. ECAs will also need to recognize verbal abuse in its many forms or risk making ludicrous remarks that inadvertently insult the user or the social groups the ECAs represent. In general, strategies for deflecting abuse must be careful to reframe from referencing negative stereotypes associated with the ECAs embodiment. Finally, ECA responses to abuse should follow CARP in persistently refocusing the discussion and encouraging the user to collaborate in problem solving. A single repeated response is also probably best at reducing the amount of time the user spends focused on the ECA rather than on the products and services the business offers. In consideration of the above, a good response might take the following form: "Your language suggests that I am not answering your questions about our products. I would be glad to continue to try to answer your questions, but if you prefer, you can also try our [sitemap](#), [search engine](#), and [directory](#)."

References

1. (2005) Indian Call Staff Quit over Abuse on the Line, *Guardian Newspapers*, May 29 issue (www.buzzle.com/editorials/5-29-2005-70698.asp).
2. Bacal, R. (1998) *Defusing Hostile Customers Workbook (Public Sector)*, Institute For Cooperative Communication.
3. Bickmore, T. and R. Picard (to appear) "Establishing and Maintaining Long-Term Human-Computer Relationships," *ACM Transactions on Computer Human Interaction*.
4. Brahnham, S. and S. Zdenek (2004). "Designing Women the Old-Fashioned Way: The Gendered Rhetoric of Animated Interface Design," presented at SLS, Durham, NC.
5. Duck, S. (1995). "Talking Relationships into Being," *Journal of Social and Personal Relationships*, 12: (4), p. 535-540.
6. McClure, D. (2005). *Re-Thinking Customer Service*, US Internet Industry Association (USIIA), (last assessed 2005).

Robust Normative Systems: What happens when a normative system fails?

Peter Wallis

The NLP Group, Department of Computer Science, The University of Sheffield
pwallis@acm.org

Abstract. Computers that can hold a conversation, such as chatbots on the web, embodied conversational agents (ECA) or automated call handling systems are, by the agent model of software, autonomous agents situated in a social world. As social animals, we humans rely on social norms that we are barely conscious of. In this paper it is argued that 1) these normative systems have a layered structure, and 2) current conversational agents only work at the top layer. People abuse such systems, not because they fail, but because their response to failure is inappropriate.

1 Introduction

Creating a machine that can hold a conversation is a difficult problem, the solution to which would have many applications. In the eighties it was felt that the problem was one of simply having the resources to collect enough data, but today it seems there is something fundamental missing from our understanding of how language works. The agent metaphor provides an alternative to the idea of computers as strictly information processors. The classic approach to language is to see it as a conduit for meaning [10]. Parsing is seen as one step in mapping text to its meaning [9], and dialog is seen as a means of updating the information state of the hearer [7]. The agent paradigm suggests a different model in which a conversational agent *acts* in a social context. If we are treating conversational agents as social actors – and not just conveyors of information – the question arises then to what extent must they rely on other social skills. What is it that makes a human trust the information presented by a automated call centre or by a virtual tutor? What makes a character in a computer game engage us emotionally, and to what extent can a virtual sales assistant get a visitor to divulge personal information about his or her spending habits and interests? In a previous paper I've argued that intentionality is key [11], but it is not intentionality 'all the way down.' This paper is about the nature of language generation once we stop thinking about it.

The observation made here is that in human / human conversation, people fail gracefully, and what is more, they do it without thinking. The hypothesis is that our ability to do this is part of our social intelligence – the process is part of our mechanism for dealing with the intra group pressures of being a social animal. In this paper the mechanism we use to cope with other people is

characterised as a *normative system*. That is, individuals have sets of behaviours that they normally do, and these individual behaviours fit together like a jigsaw puzzle to form the fabric of society. The key issue is that these norms or protocols can be broken. People can cheat, and the idea presented here is of a *robust* system of norms in which abuse is a key mechanism.

2 Norms for Social Actors

Although Margaret Thatcher didn't think so, societies are more than a collection of individuals. In societies people cooperate to do things such as build cathedrals and go to war. The nature of cooperation can be *described* with a set of rules. Some of these rules are explicit and prescriptive, while others are hardly available to the conscious mind. One can imagine for instance that each honey bee in a hive works to a set of shallow rules that make its behaviour mesh with that of other members of the hive. Bees can navigate past each other in a crowded passage, pass information about the location of food sources, and defend the hive all as part of a cooperative behaviour, presumably, without understanding their role in the process.

Sometimes this cooperation is not in the interests of the individual. Honey bees, famously, will sacrifice themselves to defend the hive. From the perspective of the selfish gene [2] one can see how such altruism would come about. A queen bee creates worker bees that have rules of behaviour that cause the individual to sacrifice itself for the good of the hive. This provides an environment for the queen's genes to prosper, which creates more bee societies with selfless worker bees.

The argument is that such rules work in human society as well. Sure people can reason about their behaviour, but such reasoning is constrained. I am polite[1] to strangers and enjoy going to the pub; I get nationalistic in the face of terrorism, I gossip [3], and buy dolls with big eyes [6]. Why? Because I am a human and humans are programmed to do those things. Without those things, I would not trust the bank, I would have to hoard food through the winter and worry about protecting such resources from my neighbour. The distinction, between rational mean-ends reasoning about action, and reactive behaviours, is made in economics. This is Jon Elster [4] introducing social norms:

One of the most persistent cleavages in the social sciences is the opposition between two lines of thought conveniently associated with Adam Smith and Emile Durkheim, between *homo economicus* and *homo sociologicus*. Of these, the former is supposed to be guided by instrumental rationality, while the behaviour of the latter is dictated by social norms. The former is "pulled" by the prospect of future rewards, whereas the latter is "pushed" from behind by quasi-inertial forces (Gambetta, 1987). The former adapts to changing circumstances, always on the lookout for improvements. The latter is insensitive to circumstances, sticking to the prescribed behaviour even if new and apparently better options become



Fig. 1. Images from the dogAttack movie discussed in Kubinyi et al.

available. The former is easily caricatured as a self-contained, asocial atom, and the latter as the mindless plaything of social forces.

He goes on to discuss attempts by economists to reduce norm-oriented action to some type of optimising behaviour. The interest here is not in discussing the nature of economic good and evil however.

The problem is of course that, unlike honey bees, individual humans (actually their genes) have their own interests to look after. What is more, we are often smart enough to be able to reason about the outcomes of our actions. With a little thought an agent might become an 'asocial atom' and cheat.

3 Robust Normative Systems

Unlike the protocols of computer science, social norms have a certain robustness about them. Rather than building a formula one racing car where every piece is optimised up to, but not beyond, the point of failure, normative systems in human societies are more like military aircraft where structures are often designed such that no individual component is critical. When an A130 hits the supports for a cable car, it is the people in the cable car who are killed, not the air crew. This robustness of design is key to effective normative systems where there is a chance individuals might cheat. The fabric of society must have some means of handling cheats and the proposal is that the mechanism is simply another norm. In order to make a system of norms robust, there must be *second order norms* that guide individuals back in (see [5]) and keep society operating. Buying rounds in a pub is a social norm that gives advantage to individuals that can skip their turn to buy. Such an individual is however soon bought into the fold.

Consider Figure 1 showing three stills from the dogAttack movie by Kubinyi et al [8] who have been using a Sony Aibo to study animal interactions. The movie shows a dog eating some meat, and on the other side of the food is an Aibo – the Sony robot dog – in a crouching position. The Aibo stands up. The dog stops eating, bares its teeth and growls. The dog goes back to eating. The Aibo takes a step forward toward the food and the dog stops eating; growls; and grabs the Aibo by the neck and throws it across the room.

What is happening here? One interpretation of what we are seeing is that the Aibo is pressing the dog's buttons in such a way that it triggers a hard-wired

social norm. The warning growl is a *second order* norm that leaves space for young dogs to learn the first order norm that adult male dogs eat first. If dog societies had no interest in socialising the young (in the same way as spiders and crocodiles) there would be no warning and the adult would simply kill any transgressor in order to keep its food. Puppies, and humans, are hardwired to recognise such second order norms but Aibos, and chatbots, are not.

4 Conversational Agents

In 2001 we set up a wizard of oz experiment with a student as the wizard, pretending to be a machine. The very first conversation went like this:

PROGRAM 0:00 Hello and welcome to the Melbourne University campus guide. How can I help you?

CALLER1 0:00 Moron.

What should the wizard say? The caller is obviously testing the program and it is perhaps interesting that our understanding of computers leads us, these days, to start the conversation with testing. Here is a conversation between a machine and a human from 1996 in which the human doesn't start by testing, but ends up doing so:

PROGRAM 0:47 G'day mate.

JUDGE02 0:55 G;day. How's it going?

PROGRAM 1:11 My mother once said "How it is going is a mystery."
She was a bit nuts.

JUDGE02 1:22 Uh-huh.

PROGRAM 1:31 Are you still learning to speak, my friend?

JUDGE02 1:52 NO, no. But your anecdote took me a bit by surprise.

JUDGE02 1:52 How was your morning?

PROGRAM 1:59 Oh well, that's a pity!

JUDGE02 2:15 I'm sorry; what's a pity?

PROGRAM 2:20 Apologies accepted.

JUDGE02 2:35 I feel compelled to insult you, and we've only been talking a minute.

PROGRAM 2:49 We are bloody well talking, aren't we?

JUDGE02 3:14 Oh baby. We sure are. So here are some keywords for you: shit mother baseball

Whatever it is that has annoyed him or her, the judge offers a chance for the system to recover, which the system fails to take up. Like the dog in the video, the judge warns the machine; the machine's response to this second order norm is not right, and the judge 'attacks.' Note that the judge's response starts out quite angry and abusive – the agent has pressed the buttons of the judge – but then the keywords become more acceptable. I would suggest that the milder (and more rational) behaviour is primarily a product of knowing that his or her response would be viewed by others. On the positive side, note that the machine – in both the chatbot and Aibo cases – is being treated as a social actor by the interactant. The problem is not to make a machine that is accepted, but to make it behave itself once it is accepted as an actor in the appropriate social context.

5 Conclusion

Ants and bees live in communities where the fabric of society can be expressed as a normative system. Each agent is given a set of norms that make it fit within the mechanisms that enable the nest/hive to survive and reproduce. The proposal is that people still use such rules, but can also think about their actions. Whereas insect communities might use a normative system that pulls action from individuals, self-conscious agents can reflect on their role and start to act based on self-interest. The normative system for these agent communities must be *robust*, and abuse is part of this process. Abuse is the fore-runner to actual harmful action and as such leaves space for individuals to change their anti-social behaviour. Whereas humans and puppies are hardwired to know what these second-order behaviours mean, Aibo's and chatbots need to be told. This is the challenge, I believe, that stands between us and the creation of effective human-machine conversation.

References

1. Penelope Brown and Stephen C. Levinson. *Politeness: Some Universals in Language Usage*. Cambridge University Press, 1987.
2. Richard Dawkins. *The Selfish Gene*. Oxford University Press, 1976.
3. Suzanne Eggins and Diana Slade. *Analysing Casual Conversation*. Cassell, Wellington House, 125 Strand, London, 1997.
4. Jon Elster. Social norms and economic theory. *Journal of Economic Perspectives*, 3(4):99–117, 1998.
5. Harold Garfinkel. Conditions of successful degradation ceremonies. *American Journal of Sociology*, 61:420–424, 1956.
6. Daniel Harris. *Cute, Quaint, Hungry and Romantic: the aesthetics of consumerism*. Basic Books, 10 East 53rd St, New York, 2000.
7. Jörn Kreutel and Colin Matheson. Modelling dialogue using multiple inferences over information states. In *Proceedings of ICOS-2, 2nd Workshop on Inference in Computational Semantics*, Dagstuhl, 2000.
8. Enikő Kubinyi, Ádám Miklósi, Frédéric Kaplan, Márta Gácsi, őzsef Topál, and Vilmos Csányi. Social behaviour of dogs encountering AIBO, an animal-like robot in a neutral and in a feeding situation. *Behavioural Processes*, 65:231–239, 2003.
9. I. Mel'cuk. Semantic primitives from the viewpoint of the meaning-text linguistic theory. *Quaderni di Semantica*, 10(1):65–102, 1989.
10. Michael J. Reddy. The conduit metaphor: A case of frame conflict in our language about language. In Andrew Ortony, editor, *Metaphor and Thought*. Cambridge University Press, 1993.
11. Peter Wallis. Believable conversational agents: Introducing the intention map. In Catherine Pelachaud, Elisabeth Andre, Stefan Kopp, and Zsófia Ruttkay, editors, *Creating Bonds with Humanoids (Proceedings of the Workshop at AAMAS'05)*, Utrecht University, the Netherlands, July 2005.