Bjoern Hartmann
Urban Studies 405
Professor Andrew T. Lamas
Essay

From It to Thou:
Anthropomorphizing Computation in Human-Computer Interaction and Multiple Agent Systems

The inhumanity of the computer
Is in the fact that once programmed and put at work,
It behaves in a perfectly honest way.
(Isaac Asimov)

It is a widely held belief within the computer science community that a paradigm shift in the way that humans relate to computers is about to occur. Catchphrases such as artificial intelligence and affective computing have caught the general public's attention, but are mostly met by distrust and aversion, mainly because of a lack of straightforward explanatory information and an over-abundance of myths and misrepresentations. Areas of computer science dealing with issues of anthropomorphization and socialization of computation – giving a human face to computers or representing human personal and social characteristics as computational models – have been under particularly close scrutiny, as widely known cinematic stereotypes, both positive and negative (Star Trek's benign Data versus Space Odyssey's tyrannical HAL), shape much of lay opinion. This essay will attempt to analyze two distinct areas of computer science research where recent trends to emulate human processes of thought or action have emerged. First, the area of Human-Computer Interaction (HCI) will be examined. HCI is principally concerned with the dyadic relationship between user and machine, and consequently with the design and implementation of suitable computing systems for human use. Subsequently, an introduction to the theory of Multiple Agent Systems (MAS) will presented. MAS describe

heterogeneous computational environments where more than two actors interrelate. Of those actors, one or more can be human, but they need not necessarily be. In fact, our discussion of MAS will focus on computer-controlled systems where each agent can be represented as a computational process, and how these systems respond to the challenges of competition and cooperation. To gain a deeper understanding how recent projects in HCI and MAS differ from previous approaches in their respective fields, they will be evaluated with respect to the theories of Martin Buber and John Berger about different modes of communication in human-to-human interaction.

To start our discussion, a brief introduction to the philosophies of Martin Buber and John Berger should be given to establish a theoretical framework for later reference. Martin Buber distinguishes between two primary words that govern how the (human) individual relates to the world around her. These primary words are *I/Thou* and *I/It*. Buber argues that since we are social beings, "[t]here is no I taken in itself" (1961, p. 44), and thus either the *It* or the *Thou* are always implied in the subject's actions. The *I/It* establishes a world of experiences whereas *I/Thou* establishes a world of relations. In Buber's terms, the *I/It* experience always has an object – the individual literally objectifies the *It* and treats it as a means to achieve a personal goal. However, no substantive enrichment can occur since by merely experiencing the world, "nothing in the situation is changed." (p. 45) *I/Thou* relationships in contrast are characterized by an absence of object and by a true boundless connection with the counterpart.

In Ways of Seeing (1972), John Berger constructs a similar classification of human relationships. Deducing his terminology from the art historical canon and its tendency to objectify the female body, Berger speaks of the *nude* as a person treated by someone else as

an object. Being *naked*, in contrast, is being oneself and revealing oneself to another person on an equal level. To summarize and interconnect the two theories we can say that when Buber's subject creates and *I/Thou* relationship with another individual, both parties reveal themselves as *naked* in the process. Conversely, when a subject objectifies someone on an *I/It* basis, he treats that person as a *nude*. In the following, we will see how these concepts can be applied to interactions between human and non-human partners.

As an extension of the engineering discipline of Human Factors analysis, Human-Computer Interaction research specifically deals with the question of how to optimize the interfaces that define how interaction between humans and computers can occur. Over the past half century, these methods of interfacing have already passed through several paradigms: after the rotary dials and glowing vacuum tubes of the first prototype computers came the punch-cards of IBM accounting machines which in turn were replaced by monitor-and-keyboard text terminals connected to the mainframe computers of the 1970s. Finally, the foundations of today's desktop-metaphor graphical windowing systems controlled by both keyboard and mouse can be traced back to research completed in the early 1980s at Xerox-PARC (see Myers et al., 2000, p. 4). What has remained constant throughout the history of user interfaces is the conception of the computer as a rigid tool that, given an exact order of what to do and how to do it, will provide utility to the user by alleviating her of strenuous repetitive tasks. In Buber's terms this clearly constitutes an *I/It* relationship, which was to be expected since the computer is merely an inanimate object, a robotic slave, created specifically for the purpose of being 'objectified' by a human master.[1] However, with the

[1] In fact, the computer's position as a 'laborer' can be likened to that of an assembly line worker in a factory run according to Taylor's scientific management principles.

establishment of the computer as a principal component of our everyday working environment, some unexpected problems have arisen.

Inger V. Eriksson (1994) points out that in our present work settings, people find themselves *abused* by their computer systems (p. 86). He exemplifies this claim by referring to problems occurring in the context of task allocation. The division of labor on any given project is often executed by first assigning to the computer what it does best (structured, repetitive numerical analysis), and leaving the human user with the leftover jobs, which often end up being disconnected and meaningless. Hence, "the tool has come to be regarded as more important or priviledged [sic] than its user" (p. 88) and this in turn forces the user to adapt her way of working to the computer's (p. 90). In essence, the *I/It* relationship has been turned on its head and the user suddenly finds herself to be the object of the computer's actions. At the source of the problem lies an ineffective model of communication between user and computer. But how prevalent are such communication problems? Unfortunately they appear to be quite pervasive: a widely-publicized study conducted in 1999 found that "84% of help-desk managers surveyed said that users admitted to engaging in 'violent and abusive' behavior towards computers" as a clear expression of their frustration with those systems (quoted in Picard, 1999, no pagination). This number alone should serve as sufficient motivation to investigate alternative methods of HCI.

The most extensive investigation into ways of improving HCI to date has been conducted at the MIT Media Lab in Cambridge, Massachusetts. Professor Rosalind W. Picard refers to her research efforts as experiments in "affective computing." (1999, no pagination) By expanding human-computer interaction to include emotional communication, Picard seeks to create interfaces in which computers naturally adapt to their users' emotional state and "respond like an empathetic friend." (see Goleman, 1997, no

pagination) While not attempting to fully understand or replicate the range of human emotions, Picard suggests that evaluation of emotional expressions will not only increase the computer's utility by enhancing communication efficiency, but that it will also provide a qualitative, intangible benefit to the user by reducing her frustrations. Empirical data (such as the study cited above) has shown that people naturally express emotion to machines. Similarly, software equipped with a cognitive model and a mind state, concepts now regularly applied in artificial intelligence applications, can relate its 'sentiments' back to the user, for example though a graphical avatar on the screen. But while humans can readily infer the emotional states of their interaction partners, computer systems do not naturally recognize such information. The challenge, then, is to enable computers to understand user emotion. Current approaches range from facilitating deliberate emotional expression by users (think of an "I am angry at you" button in your favorite word-processing software) to automatic emotion recognition by means of facial-expression extraction from video data collected and analyzed in real-time. How can we understand these efforts in light of our framework of Buber and Berger? By exposing her inner state to the computer, the user moves away from treating the machine as an object under her control (her slave) towards a more evenhanded association. The process of revealing herself can be seeing as a deliberate presentation of herself as *naked* so as to foster a mutually beneficial relationship and at the same time end the treatment of the machine as a *nude*. Preliminary studies have shown these methods to be quite effective, although the results are too recent to already pass an authoritative judgment.

Multiple Agent Systems present a second context in which instances of anthropomorphized computational processes based on human social models have recently cropped up. Some background information about agent theory will help to make sense of these environments.

Broadly speaking, an agent is a self-contained application whose purpose is the completion of a set task or the pursuit of an interactive goal within a larger environment. Agents are characterized by the following four necessary conditions of agenthood: *autonomy* – control over their own actions and internal state; *social ability* – the capability to inter act with other agents in order to assist or be assisted in the carrying out of a task; *responsiveness* – the ability to react to external stimuli and *proactiveness* – the ability to initiate action towards achieving a goal. (Wachsmuth, 2000, p. 3; and Kalenka and Jennings, 1999, p. 1) Because an agent is fundamentally goal-oriented, it primarily inhabits the *I/It* world of relations.

In a Distributed Artificial Intelligence (DAI) system, numerous agents interact with each other to collectively and collaboratively solve a given task. (Wachsmuth, p. 4) While the system as a whole is still understood as being goal-directed, its constituent parts are now faced with the problem of having to interact socially with each other to maximize their collective efficiency. The most frequently employed models to govern these social interactions are master-slave protocols and contract networks. From our previous discussion, we should expect contract networks to generate better overall performance, since in the *I/It* model of master-slave relationships "nothing in the situation is changed." (Buber, p. 45) Unfortunately, no experimental data was available to directly compare these two concepts - future research in this area is required.

Multiple Agent Systems are taking the evolution towards socialization one step further: here heterogeneous and self-interested agents have to interact, negotiate and coordinate with each other to achieve their individual objectives or tasks in absence of a unified system goal. (Castelfranchi and Tan, 2001, p. 6) Even though no universal targets exist within MAS, the individual agent still finds itself faced with an imperative to collaborate with other agents since its own success is dependent upon the assistance of those other

agents, which in turn are only likely to offer their help if the agent exhibits cooperative social behavior. In this regard, then, Multiple Agent Systems can be likened to human communities in that their members are pursuing personal goals while being constrained by their mutual interdependence. If the analogy holds, we should expect both more harmonious and more productive systems to result from endowing agents with a level of social awareness. (Kalenka and Jennigs, 1999, p. 2)

Indeed recent studies investigating methods of socially conscious agency have reported results supporting this hypothesis. Kalenka and Jennigs present a MAS where a community of *cooperative agents* that are willing to perform actions which are personally detrimental while at the same time being beneficial to the community fare better than both *helpful agents* (which will cooperate if the tasks are not self-detrimental) and *selfish agents* (which will only cooperate if it is in their own positive interest) at achieving their goals. Glass and Gross (1999) describe another MAS where agents were repeatedly faced with the decision to act in the interest of the larger community or to default and pursue individual goals. A system of "brownie points" rewarded agents for making socially conscious decisions and punished them for defaulting on their obligations to the community. Interestingly, socially conscious action proved to be most beneficial to the system as a whole at an intermediate level – neither completely self-interested nor totally altruistic models produced optimal results. Besides this remarkable idiosyncrasy, though, it still appears as if the lessons of mutually respectful, socially aware interaction on the *naked* or *I/Thou* level promises somewhat related benefits in machine controlled agent systems as it does in human societies.

In conclusion, the preceding discussion sought to demonstrate that issues previously thought to be applicable solely to interpersonal relationships have unexpected relevance to Human-

Computer Interaction as well as computer-computer interaction in Multiple Agent Systems. However, a warning not to misinterpret and overextend these parallels has to be extended. While the frameworks of Buber's *I/Thou-I/It* and Berger's *naked* and *nude* can be applied to computer science contexts, they necessarily take on a distinctly different meaning. Whereas individuals can truly relate to each other with the full breadth of human potential, computer systems will only ever be able to imperfectly imitate such relationships. The cause for this constraint does not necessarily lie in the limits of computability, but rather in our inability to fully understand our own nature. As Arnold Farr put it: "To fully understand myself, I would have to have absolute knowledge of the entire history of humanity." (Lecture presented in class on 04/16/2001) And that is a problem surely beyond the scope of computer science or any other human endeavor.

Bibliography.

Sources related to computer science and artificial intelligence:

Castelfranchi, C. and Tan, Y. 2001. The Role of Trust and Deception in Virtual Societies. Proceedings of the 34th Hawaii International Conference on System Sciences 2001.

Eriksson, Inger V. 1994. Computers As Tools. Ethics in the Computer Age, a publication of the Association of Computing Machinery [ACM].

Glass, A. and Grosz, B. 1999. Socially Conscious Decision-Making. A publication of the ACM.

Goleman, D. 1997. Laugh and Your Computer Will Laugh With You, Someday. New York Times, January 7.

Kalenka, S. and Jennings, N.R. 1999. Socially Responsible Descision Making by Autonomous Agents. Korta, K et al., editors, Cognition, Agency and Rationality. Dordrecht: Kluwer, 79:135-149.

Myers, B., Hudson, S. E. and Pausch, R. 2000. Past, Present, and Future of User Interface Software Tools. ACM Transactions on Computer-Human Interaction, 7: 3-28.

Picard, R.W. 1999. Affective Computing for HCI. In Proceeding of HCI'99, Munich, Germany.

Wachsmut, I. 2000. Methoden der kuenstlichen Intelligenz. A publication of the Universitaet Bielefeld.

Sources taken from or relating to class:

Berger, J. 1972. Ways of Seeing. London: Penguin.

Buber, M. 1961. "I and Thou." The Writings of Martin Buber. Ed. Will Herberg. Cleveland: Meridian Books, 43-62.

Farr, A. 2001. The Location of Thought: Liberation and Epistemic Communities. Lecture presented to class on 04/16/2001.

Lamas, A. 2001. URBS 405 Lecture Notes and hand outs.