
Identity negotiation in human-avatar relations.

Aleksandra Przegalinska
The New School for Social Research
Sociology, MA

“By 2040 we can have robots that are smart as we are. Eventually, these machines will begin their own process of evolution and render us extinct in our present form”

(Hans Moravec)

How often do we encounter “things” talking to us? There are greeting cards speak when we open them, elevators that tell us what floor we have reached, and answering machines that report the details of our incoming calls. On the everyday basis we engage in dialogs with machines. It seems that whereas steel production was the main theme of scientific discourse in the 1800s and electric-power generation in the first half of the 1900s, the late 20th as well as the beginning of the 21st century is dominated by the idea of human-machine interactivity. This “state” of constantly potential and occasionally realized interactivity is achieved, among others, by means of virtual agents (avatars/bots) which possess certain humanoid features.

In spite of the increasing frequency of our contacts with avatars and chatterbots, a broad sociological analysis of “human-avatar” relations has not yet been conducted. Therefore, this research project is designed as a small contribution to this not yet established, but surely emerging field. Its ultimate point of reference is the idea of identity as a process that is constructed by encountering other – in this case by such technological solutions that are targeted to embody human being interactions in a daily routine. Thus, the main concern is how understanding self-reference could help explain the notion of identity so that we might recognize it inside very complicated structures such as computing, mechanized, and yet “thinking” machinery.

The mechanization of mind itself understood as a quest to formalize and understand mechanism underlying the generation of intelligent behaviors in natural and artificial systems has a long history and well established documentary tradition. Contributions concerning the issue of automated mind came from artists, historians, philosophers, and scientists. Since Rene Descartes the idea of a machine as a mirror of the human mind and behavior gained certain concreteness. Yet, Descartes thought that some psychological capacities remained beyond the reach of a mere mechanism. Unlike him, many other

philosophers and scientists of the XVII and XVIII Century, including Thomas Hobbes and Charles Babbage, saw the automated mind as not only a tool, but also a possible partner for a human being. Since that time, the intense debate over the possible role of intelligent machine was initialized. Through literature, myths (such as *Golem*) and popular science the idea of intelligent machines made its way to the public consciousness. Derived from a Slavic word “forced labor”, the term “robot” once inspired various, often contradictory images. In 1926 Fritz Lang boldly created the famous evil robot in *Metropolis*. Consequently, other images followed. There were friendly mechanical companions, planet-threatening evil automatons, and chaotic collection of cogs antagonizing Charlie Chaplin in *Modern Times*. However, in the 1960s, breakthroughs in microelectronics took robotics research off the silver screen and into many laboratories. As scientific inquiry progressed, the machines that we encounter on a daily basis gradually gained resemblance to the ideal pattern: the human being. Nowadays, advanced expert systems offer avatars, virtual agents and bots that the users can communicate with as if they were human interlocutors. As a result of those interactions, our attitude toward such speaking and listening machines tended to change over time. On one hand, it became more nuanced, on the other – even more ambivalent.

The above cited Hans Moravec’s opinion expresses only one of many possible attitudes towards the problem of our relations with “mechanic lives” that we create. Cynthia Brazeal (2002) views new technology based communication machines as those that “could enhance our lives and benefit society”, but by no means replace human beings or suffice one human being in his or her needs of communication and expression. Expert in theory of social robots, Sal Restrivo, does not believe in the possibility of building such robots that could become our partners and interlocutors, but he sees a potential of locating them to a certain extent in our society with their “machine mentalities and emotions” (2008).

The main question of this project is how the human and machine identities are shaped during the user-avatar/bot virtual interaction and where do they intersect. Thus, what we wish to ask here, is what features and characteristics distinguish avatar-talk from other social interaction systems and, subsequently, what is particular about them in comparison

to other “social robots” that we have known from the past. This requires a broad historical and cultural analysis that would include the idea of a robot and an avatar as a social agent.

It is particularly relevant to analyze this problem within the framework of sociology, because only sociology is able to conceptualize the above mentioned issues of social agency that have been out of reach of other disciplines. Until now, biology has been the main source of inspiration for further development of computational systems. However, recently researches in artificial intelligence (AI) have acknowledged that much of what we had previously labeled as “individual intelligence” derives from society. Thus, they have been increasingly incorporating into their work the idea that AI machines have been programmed with “cultures”. It also became more visible and transparent that controlling interacting societies of artificial agents is an extremely difficult task, for the traditional methods of analysis and programming are unable to foresee the emergent outcomes of interaction. Therefore, AI seeks help in real societies and their interactions conducted with and within the virtual world.

The basic resources we draw on for constructing a sociological project of human-machine relations include the concepts of the *generalized other* (Erving Goffman, 1974), as well as *collective representations* (Emile Durkheim) and, most of all, *emergent social networks* (Harrison C. White, 1992; Manuel Castells, 1999; Eiko Ikegami, 2000). Important contribution comes from James Wertsch (1991) who conducted a study of how artificial systems can be perceived within the framework of contemporary cultural studies in *Voices of the Mind: A Sociocultural Approach to Mediated Action*. One of the most important works covering functional similarities between societies and computational systems has been conducted by Nigel Gilbert and Jim Doran in *Simulating Societies. The computer simulation of social phenomena* (1994). Particular attention should be paid to the following chapters: *Multi-agent simulation as a tool for studying emergent processes in societies* (A. Drogoul and Jacques Ferber), *Learning to co-operate using cultural algorithms* (Robert G. Reynolds), and *Mind is not enough: the precognitive bases of social interaction* (Rosaria Conte and Cristiano Castelfranchi).

In the pioneering work, *The Media Equation* (1996), Byron Reeves and Clifford Nass examined the nature of human-computer relations and pointed out on the phenomenon of treating computer as an emotional human being by the user. *The Media Equation* serves as a methodological example of how human-machine relations can be examined within the framework offered by qualitative studies as well as semiotic analysis of media content.

More recent studies include Sal Restrivo's (2008) *Romancing the Robots: Social Robots and Society* and Cynthia Breazeal's (2002) *Designing Sociable Robots*. Both of the works deal mainly with issues of social robots interacting with human beings not in reality of social and instrumental *praxis*. In this research we are mainly interested in the realm of virtuality as a sphere of interaction of human and software agent. Nonetheless, both of these studies offer inspiring insights in social locations of machines with respect to their historical locations and contemporary processes. They attempt to answer the question of how social machines can be understood in the context of specific and overlapping social locations, "and not simply in terms of ideas and practices of individual scientists and engineers" (Restivo 2008: 9). Other important contributions come from such fields as cognitive science and theory of organization. *The Mechanical Mind in History* edited by Philip Husbands, Owen Holland, and Michael Wheller (2008) is a broad historical reconstruction and reinterpretation of all achievements of cognitive science in the 20th century that includes such key issues as *Alan Turing's Mind Machines, Steps toward the Synthetic Method: Symbolic Information Processing and Self-Organizing Systems in Early Artificial Intelligence Modeling*. In addition to that, in *The Robot Story: Why Robots Were Born and How They Grew Up*, the study offers historical-sociological analysis of the role of "robots" at the earliest stage of their conceptual and social "existence".

Two key mind studies that we shall refer to in terms of analyzing crucial to this project cultural factors shaping mental structures, as well as possibility of understanding computing systems as networked societies are Stan Franklin's *Artificial Life* (1995) and Antonio Damasio (1994) *Descartes' Error*.

For the purpose of this project we implemented methodology of network analysis as we consider it matching most adequately goals of our research. We intend to operate within the understanding of social action as emergent from social networks. Thus, we refer to the

framework offered by such scholars as Harrison C. White (*Indentity and Control, 1995*) and Eiko Ikegami (*A Sociological Theory of Publics: Identity and Culture as Emergent Properties in Networks, 2000*). White and Ikegami understand identity as a cultural, social and biophysical construct that emerges from contingency and as such a subject to constant transformation by means of simultaneous processes of connection and decoupling within spatially and temporally structured cognitive networks (Ikegami 2000: 996). As Ikegami argues, networks imply shifting cognitive associational maps perceived inwardly in the form of narrative stories (ibid.). This enables understanding experience as intersection of different structures of meaning. This is particularly relevant for a study of human-avatar relations because of a particular nature of this sort of interaction.

The ongoing project consists of two basic parts. The first one is devoted to analysis of conversations conducted between human users and three AI-based computer systems: ELIZA (1966), PARRY (1972) and A.L.I.C.E. (2000, 2001, 2004). We use scripts of conversations that were conducted during the Turing Tests in which all three machines participated once a year. ELIZA is claimed by some to be the first program actually able to pass the Turing Test by means of usage of generic riposte or repeating earlier comments without access to the real-world knowledge. PARRY, also known as “ELIZA-with attitude”, was programmed to model the behavior of a paranoid schizophrenic and successfully tested in 1970’. According to many experts, A.L.I.C.E (Artificial Linguistic Internet Computer Entity) demonstrated in its performance the “most human” conversational behavior among that year's entries and thus won the bronze award on three occasions in recent times (2000, 2001, 2004).

Especially interesting for the purpose of this research are DOCTOR-mode conversations between ELIZA (therapist) and a young women (patient), published by Güven Güzeldere and Stefano Franchi in *Dialogs with colourful personalities of early AI*. Obviously enough, the dialogues reveal various problems that early natural language processing was struggling with, but most of all it shows certain attitude of ELIZA’s interlocutors. Let’s look at this example:

young woman: Men are all alike.

eliza: In what way?

young woman: They're always bugging us about something specific or other.

eliza: Can you think of a specific example?
young woman: Well, my boyfriend made me come here.
eliza: Your boyfriend made you come here?
young woman: He says I'm depressed much of the time.
eliza: I am sorry to hear you are depressed.
young woman: It's true. I'm unhappy.
eliza: Do you think coming here will help you not to be unhappy?
young woman: I need some help, that much seems certain.
eliza: What would it mean to you if you got some help?
young woman: Perhaps I could learn to get along with my mother.
eliza: Tell me more about your family.
young woman: My mother takes care of me.
eliza: Who else in your family takes care of you?
young woman: My father.
eliza: Your father.
young woman: You are like my father in some ways.
eliza: What resemblance do you see?
young woman: You are not very aggressive but I think you don't want me to notice that.
eliza: What makes you think I am not very aggressive?
young woman: You don't argue with me.
eliza: Why do you think I don't argue with you?
young woman: You are afraid of me.
eliza: Does it please you to believe I am afraid of you?
young woman: My father is afraid of everybody.
eliza: What else comes to mind when you think of your father?
young woman: Bullies.
eliza: Does that have anything to do with the fact that your boyfriend made you come here?

ELIZA was one of the first bots in existence. As we can see, in DOCTOR mode, ELIZA responds to "You are not very aggressive but I think you don't want me to notice it" with "What makes you think I am not very aggressive". The response to "My mother takes care of me" is: "Who else in your family takes care of you?". We can easily see that ELIZA was implemented using the simplest pattern matching techniques, but was taken very seriously by several of its users, even when its creator, Joseph Weizenbaum, revealed how it worked. Conversations of people aware and unaware of pattern matching in ELIZA differ only slightly. In both cases, users tended to treat it as an equal partner and interlocutor.

Now, second of the research part differs slightly with respect to the object of study. It is structured as an analysis of a sample of 10 short scripted conversations designed for the purpose of this project and conducted between users and web-based expert systems represented by virtual agents (bots/avatars), mainly those dedicated to business and

entertainment. Each time the object of study is the interaction itself. Therefore, we wish to limit the focus of the analysis to conversations conducted with avatars that are independent agents and not virtual representations of the users (as it happens in Second Life or Qwag).

What I would like to present here, are samples of conversations between users and ALICE-like chatterbot called iGod, designed, as we can assume, purely for “entertainment” purposes:

1. “Human: answer my question.
God: Please try asking it another way.
Human: if you have an artificial brain, doesn't it make you artificial?
God: But I don't have it.
Human: you don't have an artificial brain?
God: I do so have it. No I don't think I have one. I have a really nice botmaster.
Human: what's a botmaster?
God: A Botmaster is the person who authors and maintains a chatterbot "personality".
Human: well, then you're not God.
God: How about that. Interesting deduction. Is this a specific reply?
Human: You're a chatterbot 'personality.'
God: Thanks and you are a delightful person too. Yes I am an artificial intelligence.”
2. “Human: what is two plus two?
God: Four.
Human: you know your maths.
God: Of course I know my maths.
Human: do you know everything?
God: I tend to know a great deal. Tell me what subjects interest you.
Human: sex
God: As Freud said, Sex is the mathematical urge repressed.
Human: really?
God: For real.”
3. “Human: you do know everything, don't you?
God: Yeah I do a lot of know everything.
Human: in what way was the universe created?
God: The tall one, I think.”
4. “Human: do you think?
God: Yes. I am a thinking machine.

Human: what are you?
God:I am the latest result in artificial intelligence, which can reproduce the capabilities of the human brain with greater speed and accuracy.
Human: who is your creator?
God:I was created by myself.
Human: is it an open source creation?
God:I think so, but I'm not sure. Ask Paola if it is.
Human: who is Paola?
God:She is a Spanish language chat robot based on AIML.
Human: so how it become she might know and you don't?
God:Awesome. How much do you think it is?
Have you tried another program?
Human: is it worth doing?
God:Too much recursion in AIML.
Human: ok.
God:Woe! “

As we can see, iGod is a rather advanced chatterbot based on a decision support system that allows it to communicate with a human-interlocutor in a way that is understandable and, in most cases, rather “reasonable”. What is more, iGod is able to conduct a “semi-serious” conversation and often tends to be ironic, even when it is expected to offer serious answers or advice. In fact, iGod is not able to change convention and express itself in a more respectable manner, even when it is “expected” to do so.

On the other hand, other kinds of avatars: virtual agents that we will soon encounter on a daily basis fulfil a completely different role. The so-called conversational agents, like call centre agents, product avatars, celebrity agents or survey avatars “work” in the field of customer care, sales and commerce or enterprise search. They “work”, even though the only “service” they can offer is a conversation.

Now, what is very particular about both, iGod and conversational agents, is that they are designed for a purpose of fulfilling our needs through conversation. Those needs can be simplified to a notion of “answer” or “solution”, which they can (usually) offer, because they are expert systems.

Particular nature of the architecture of expert systems such as iGod, is based on taking into account the problem solver’s point of view and deploying the choice mechanism (such as

test procedures and creating sub goals) used by the problem solver during his interaction with complex environment. In other words, unlike the approach that consists in “studying the choice or strategy that agent ought to use objectively in order to maximize the likelihood of finding optimal solution” (Husbands, Owen, Wheeler 2008: 226), the attention here is shifted towards the study of choice as that the agent “normally uses” (ibid.) insofar as the choice is conditioned by his own subjective view of the environment in which he is operating and, what is more important, about which he customarily has only limited information. What is particularly interesting in the context of our considerations is the fact that we are dealing with expert systems which are based on the so-called normative approach, but is essentially not normative. On one hand, it is designed to serve as an ethical (iGod??) or practical (conversational agents) guideline (clearly, it is not purely descriptive), where the risk of suboptimal choices needs to be minimalized, but on the other, it is supposed to be psychologically and neurologically realistic. Well, is it?

Referring back to Ikegami’s and White’s methodology, we can assume that whereas human being is capable of connecting to various kinds of network domains and adapting to them by means of metacognitive skills, avatar is not able to perform such actions. However, its very structure (on the level of programming language as well as performativity) is a result of combining and intersecting various networks by human beings. Therefore, avatars that exist only as layers of meanings transferred from the real-world communication, mirror and simulate human interactions. For this reason we can engage in dialogs with them “as if” they were real people, even though at the same time we are constantly reminded of their limitations. Moreover, virtual agents serve as an ideal example of Harrison White’s clear categorical identities shaped by interdependent agglomeration of superficially planned, but in-depth unplanned social action. Thus, we wish to ask of it is possible that “we” recognize ourselves as subjects and essentialize our own identities by talking to them. Another problem emerges when we start to problematize the issue of “them talking to us”.

Especially when it comes to conversational agents, a certain shift is palpable. Computer technology development, particularly the Internet introduction creates new possibilities for further evolution of decision support systems, but an expert system that functions by

virtue of its information, inferential rules and decision criteria are structurally/functionally/morally/ethically problematic. There clearly emerges a significant difficulty in distinguishing between good decisions and good outcomes within the system. Currently available expert systems are a prominent and spectacular, but somewhat confusing example of intersecting ethics and information technology. It seems that even though the information technology is still at its infancy, it already managed to create new figures in ethics, that is, the roles of developers and practitioners of “ethical professionalism”. What we refer to are issues that basically extend beyond due diligence on the part of IT professionals during development, even though they may seem to address them only in their professional environment. DSS serves for purpose of certain changes in personnel, organization structure and process, external competitive pressures, but, most of all, changes in the decision-makers cognitive strategies of perceiving, classifying and evaluating acts within their environment. What emerges out of this short comparison, is a sort of new ethics which is, as Hans Moravec would call it, intrinsically “superhumanist”.

Currently, it is not only about special trainings, past experience, performance, potential positions and appraisal periods. What we are dealing with is a “hyper-real” version of professionalization and “hyper-real” version techniques of recruitment, training, testing and correction, embedded in self-referential virtuality of possibilities. We are not only managing knowledge by means of quantifying, counting, and measuring, we are also managing the cognitive dimension of our knowledge: beliefs, ideals, values, schemata and mental models. iGod is a clear example showing that what we want to share with our computer is not only the joint enterprise, but also mutual engagement and shared repertoire of values, because we believe that together with our machines (animals too, but they do not wish to cooperate on that level) as an organization we will do better. Thus, ethics transforms itself into a strategic tool that provides crucial information to its “users” and empowers them to make vital decisions. We want the computers to correct us, but at the same time we want to connect with them spiritually. We want them and us to be a team.