Are Agents an Answer or a Question?

Joseph A. Goguen

Department of Computer Science & Engineering

University of California at San Diego

1. Introduction

Agents are increasingly upon us. Although opposition is rare, "intelligent agents" have been attacked for user interface problems, and on larger social issues. Agent supporters have countered these arguments and raised doubts about alternative technologies. We place this in historical, social, and ethical contexts, noting the cyclic nature of such debates. One conclusion is that many problems with artificial agents arise from a poor understanding of social aspects of human agents.

2. Historical Perspectives

The history of technology has seen many movements call for human-like systems, and use anthropomorphic terminology to generate understanding and support. Such movements often make excessive claims, perhaps misled by their own rhetoric or their (sometimes impressive) partial success. This raises unrealistic expectations, which often leads to disappointment, which is surprisingly often followed by rebirth with similar goals, and somewhat improved terminology and technology.

One example is artificial intelligence (AI), which produced many claims and predictions that turned out false; early AI research stressed the logical representation of mental states, reaped huge publicity and funding (often based on projected military applications), and attempted (with some success) to colonize adjacent fields such as cognitive psychology and neuro-science; it then crashed when numerous projects terminated because their demos didn't scale up to genuine applications, precipitating the "AI Winter." Yet many modest successes were scattered among spectacular failures, and there was a rebirth using new

technologies like neural nets and machine learning, again based on analogies with human agents.

Other projects that followed a similar path are machine translation, perceptrons, compiler generators, and the Japanese Fifth Generation. Each failure can be largely attributed to inadequately understanding cognitive and social factors. For example, early machine translation assumed that a good lexicon and grammar would suffice, whereas we now know the key importance of background knowledge and context. The silver lining on these clouds is a deeper understanding of technological limitations, and (to a sadly lesser extent) an enhanced humility.

3. Critiques of Intelligent Agents

Ben Shneiderman, a leader in user interface design, has criticized agent interfaces for their lack of predictability, disenfranchisement of the user, and unclarity of moral attribution [11]; he claims agents are by nature difficult to understand and predict, whereas users want a sense of control, to know that agents will do what they want them to do. He also worries that, if an agent does something the user does not intend, such as destroying a crucial file, or unleashing a virus on the internet, it is unclear who should be held accountable, the user, the manufacturer of the agent, the installer of the agent, the creator of the last patch, or .....

Jaron Lanier [8] says "the idea of 'intelligent agents' is both wrong and evil." It is wrong because it will not solve the problems that it is supposed to solve, and it is evil because it will "make people redefine themselves into lesser beings." A common application of agents is to help users find what they want on the web. But Lanier argues that agents will inevitably trivialize the interests of users, and "deliver an overdose of kitsch," due to taking the same "lowest-common-denominator approach to content that plagues TV." Unfortunately, his view is supported by even a glance at commercial web search engines, though there are exceptions. Lanier also argues that people will come to think of agents as human-like and actually intelligent, and therefore will reduce their own autonomy, by having to force their behaviors into the restricted patterns allowed by and effective with agents; even worse, they may come to think of themselves as being like computers.

Agent terminology and properties have roots in the philosophy of Immanuel Kant, who assumed rationality and autonomy of human agents in his project to construct a non-theistic foundation for Western morality[1]. His assumptions are disputed by many modern philosophers and psychologists, but this doesn't mean they can't be useful properties of artificial agents. The adjective "intelligent" also has a connotation of Kantian rationality, but seems to be slowly going out of fashion, presumably from its vagueness and overuse, as well as its association with discredited aspects of AI. Perhaps the word "agent" will someday suffer a similar fate.

4. The Advocates Respond

The most basic argument of agent boosters is that the increasing complexity of both systems and their environments requires users to abandon some control to automatic systems, whether or not one calls them agents; in short, agents are inevitable. An often cited example is filtering agents to stem the flood of unwanted email. Shneiderman concedes the need for computational assistance, but insists that users must understand and control any software to which they delegate [12]. He does not oppose automation, but rather unpredictability, incomprehension, and disenfranchisement.

A prominent defender of agent technology is Pattie Maes, former head of the Software Agent group in the MIT Media Lab. In [12], which records a 1997 debate between Maes and Shneiderman, Maes accepts Shneiderman's criteria (predictability, comprehension and control), and says these are part of her research agenda. But Shneiderman and Lanier both say they don't know any agents that perform significant tasks and also adequately satisfy the criteria, and they cite prominent failures, such as Microsoft's infamous "BOB," and its recently demoted "Clippy."

Shneiderman suggests that direct manipulation and information visualization should be used instead of agents. But Maes and others say that many domains are too complex for either direct manipulation or visualization, e.g., the world wide web [12], and also say they want to exploit the best interface technology they can get, including direct manipulation and information visualization.

[1] The situation is quite different for many Eastern moral system, such as Mahayana Buddhism [10].

Regarding responsibility for the actions of an agent, Maes says unequivocally that users should be held accountable [12]. But in view of examples like a badly designed agent unleashing a virus, this does not seem reasonable. My impression is that Maes and Shneiderman followed pre-determined rhetorical strategies, and hence often talked past each other. This is reinforced by more recent responses to agent criticisms, including guidelines emphasizing that successful agents must obtain user's *trust*. Unfortunately, the issue of whether an agent deserves that trust is not addressed, although it seems seriously unethical to encourage trust in an agent that does not do what its user really wants.

It seems we are on the cusp of a new cycle, where agent research reinvents itself with nearly the same goals (more emphasis on user satisfaction), and an enhanced technology embracing insights from interface design. The suggestion of Maes and others to drop the word "intelligent" [12] supports this view, and might suggest that the debate is mainly about terminology: should we use the word "agent" for the resulting technology? Or more precisely, where should we draw the line, such that only software with sufficient autonomy is called an agent? Another conclusion is that ethical issues are not being adequately addressed, especially by the advocates of agent technology.

5. Socially Intelligent Agents

The most important issues about agents do not concern terminology. Raising the stakes to "socially intelligent agents" makes this clearer by moving the debate to questions that are obscured when the primary focus is technology, as in the Shneiderman-Maes debate. The following are some deeper questions:

1. How human can agents appear? How human *should* they appear? What are the ethical issues with anthropomorphism?
2. Can agents have emotions? Should they?
3. What effects can an agent's (simulated?) emotions have on users? How can this be exploited, e.g., in advertising? Is this ethical?
4. Is it necessarily the case that adaptive agents are less predictable? Is predictability always good?

5. Do socially intelligent agents need good models of their users? What needs to be modeled? How detailed must the models be? Are models of the user's model of the agent needed? How far should this recursion on models go?
6. Should models of individual users be public? Or should some information be private? If so, how can this be enforced on remote agents?
7. Can agents produce the appearance of qualities like individuality, intelligence, and confidence? Should they? If so, when?
8. What are the inherent limitations of agent technology? What are its greatest strengths?

Such questions can help us think more deeply about artificial agents, and about what it means to be human. It seems difficult to give satisfactory answers, and most questions lead to even deeper questions. For example, recent research on the physiological basis of memory shows that it is connected to the limbic system, and thus to the emotions, and indeed, to the whole body. Therefore agents without emotions and bodies won't have the same associative capabilities as humans (or at least, could only have them in very different ways). But what could an artificial limbic system be like?

We have mainly discussed cognitive rather than social capabilities of agents. We know very little about technical, social and ethical issues for communities containing both artificial agents and humans. Here are a few questions:

1. What would it mean for a software agent to be part of a community? Is there a good operational definition?
2. Should agents lie to other agents? If so, when?
3. Should agents lie to humans? If so, when?
4. Can agents hurt each other's feelings? Should they? What about hurting human feelings?

Some of these may seem like science fiction – indeed, some may *be* science fiction – but they can help us get deeper into this difficult territory.

The social sciences have accumulated enormous information about how human societies and human communication work. For example, research in conversation analysis (CA) emphasizes "recipient design,"

that speech is carefully crafted to match its recipients, taking account of factors like shared background and values, language ability, and attention span. Being able to do this is part of what it means to "know" someone; it is also reciprocal, i.e., you expect speech directed at you to be recipient designed to the extent that you know that the speaker knows you. Work in CA also highlights the incredible accuracy of timing in ordinary conversation, e.g., turn transitions are accurate down to a millisecond, in contrast to the approximate 500 milliseconds required for conscious intervention.

Socially sensitive requirements elicitation could help improve agents. Good systems are much more likely with a clear vision of what and who they are for. Another lesson is that good requirements require careful work with real users; designer's guesses are almost never adequate [3].

Distributed cognition also has rich implications; for example, work of Edwin Hutchins and Lucy Suchman carefully describe important kinds of interaction that are rarely considered for artificial agents, such as mentoring, story telling, and using plans as resources for coordination rather than action; the ways that replanning, monitoring other agents, and delegating tasks are handled by humans are often far from anything we can program today.

The branch of the sociology of technology and science called actor-network theory (ANT) is also relevant. Founded by Michel Callon and Bruno Latour [9], it focuses on *networks* of *actants*, connected by *links*, rather than on autonomous agents; actants may be non-humans, such as PCs, programming languages, and transmission media. Work, consisting of chains of translations along links among actants, must be done to hold the network together, "recruiting" actants to contribute by translating into their languages. This is the mechanism by which socio-technical compromise is achieved and projects are coordinated. It is also the heart of technology transfer.

Star and Bowker [1] add to classical ANT an emphasis on the infrastructure required to support networks, on the technical standards and classification systems that allow complex constructions to be accomplished, and on what gets left out in the stories that are told about projects. To this, my own recent research would add an emphasis on the *values* that are embodied in actants, and that are translated along

the links: social and ethical concerns are inseparable, and are ubiquitous in all socio-technical systems, even though they may be hidden.

Much is known about narrative structure from socio-linguistics [6]. Following an optional *orientation section*, a typical narrative has a sequence of *narrative clauses* interleaved with *evaluative material*, which places the narrative material in its social context by appeal to shared values. The *narrative presupposition* says that the order of narrative clauses is the order of the events they report.

Although more cognitive than social, work by Lakoff [7] and others on metaphor has significant implications. Far from being an esoteric literary device, metaphor is pervasive in ordinary human interaction, and indeed, is the most significant tool we have for understanding and communicating in many key areas of life. *Blending* multiple concepts into a single conceptual space [2], is also important and pervasive, as well as unconscious and almost instantaneous, so that we hardly notice it. Without an ability to process complex metaphors and blends, agents can only have very superficial understandings of human communication. Computational aspects of blending have been little explored, but [4] gives a formalization now being implemented at UCSD.

Unfortunately, agent research is often technology-driven, paying little attention to social issues of any kind, let alone the rather sophisticated results surveyed above. This helps explain why it has accumulated so much criticism.

6. An Alternative

The near amicable resolution of the Shneiderman-Maes debate suggests that the dichotomy between delegation to agents and direct manipulation is artificial. First, both are metaphors, which can be realized in many ways, and which typically occur blended with other metaphors, such as information visualization, window, menu, button, etc. Second, these are not the only alternatives; in particular, navigation is important, and can also be blended with other metaphors.

My laboratory is experimenting with *narrative-driven navigation*. Kumo is an agent that assists users with proofs, and generates websites to display those proofs [5]. It can be used stand-alone, or in a dis-

tributed mode where multiple agents each assist one user. The second mode uses a protocol to maintain proof consistency over the distributed database, so that users know (modulo internet delays) what other users have done. Kumo output is viewed on a web browser, and proof websites have narrative structure, to improve proof comprehension using the temporal and evaluative possibilities of narrative, including dramatic tension. The human prover tells Kumo what to try and how to display the result. Our experience suggests using a carefully chosen blend of interaction metaphors for interfaces to complex systems, based on an empirical analysis of the social context of system use.

7. Conclusions

In exploring the agent debate, we found that we knew little about agents, either human or artificial; we found more questions than answers, and the unanswered questions seemed more important than the current answers. Still, it seems safe to predict that agent technology will continue to reinvent itself, and that substantial new moral problems will arise. It is unlikely that technical progress will bring agents close to being human, but we will continue to try, in the process learning more about why it is so difficult; much of this will be social. I hope that this journey will inspire more humility about our technology, and more awe of what it means to be human.

References

[1] Geoffrey Bowker and Susan Leigh Star. *Sorting Things Out*. MIT, 1999.

[2] Gilles Fauconnier and Mark Turner. *The Way We Think*. Basic, 2002.

[3] Joseph Goguen. Requirements engineering as the reconciliation of social and technical issues. In Marina Jirotka and Joseph Goguen, editors, *Requirements Engineering: Social and Technical Issues*, pages 165–200. Academic, 1994.

[4] Joseph Goguen. An introduction to algebraic semiotics, with applications to user interface design. In Chrystopher Nehaniv, editor, *Computation for Metaphors, Analogy and Agents*, pages 242–291. Springer, 1999. Lecture Notes in Artificial Intelligence, Volume 1562.

[5] Joseph Goguen and Kai Lin. Web-based support for cooperative software engineering. *Annals of Software Engineering*, 12:25–32, 2001. Special issue for papers from a conference in Taipei, December 2000.

[6] William Labov. The transformation of experience in narrative syntax. In *Language in the Inner City*, pages 354–396. University of Pennsylvania, 1972.

[7] George Lakoff. *Women, Fire and Other Dangerous Things: What categories reveal about the mind*. Chicago, 1987.

[8] Jaron Lanier. Agents of alientation. *Journal of Consciousness Studies*, 2:76–81, 1995.

[9] Bruno Latour. *Science in Action*. Open, 1987.

[10] Keiji Nishitani. *Religion and Nothingness*. University of California, 1982.

[11] Ben Shneiderman. Humans unite! *Scientific American*, March 1999. Interview by Tim Beardsley.

[12] Ben Shneiderman and Pattie Maes. Direct manipulation vs. agents. *Interactions*, 4(6):43–61, 1997.