Machine agency

As an aspect of the early twenty-first-century technoscientific imaginary the sociality of machines is well established. Beginning with early work on machine intelligence in the 1950s, our conception of machines has expanded from the instrumentality assigned them in craft and industrial contexts to include a discourse of machine as acting and interacting other. Within the project of designing intelligent machines in the 1970s and 1980s, at the time when I first encountered this project directly, the premise was being established that computational artifacts just are interactive, in roughly the same way that we are albeit with some more and less obvious limitations. However ambitious, the problem in this view was a fairly straightforward task of overcoming the limitations of machines by encoding more and more of the cognitive abilities attributed to humans into them.

The less visible and somewhat restrained AI projects of the 1990s played down the personification of machines in favor of cybernetically-inspired computer science and engineering initiatives aimed at recreating artificial life forms, via the technologies of neural networks, genetic algorithms, situated robotics, and the like. These developments shifted the project of machine intelligence away from what is now referred to as "good old fashioned symbolic information processing" AI, toward a technical practice based in more foundational metaphors of biocultural evolution and, in Sarah Franklin’s phrase, "life itself." Nonetheless, attributions of human-like machine agency seem alive as ever in both professional and popular discourse. The growth of the Internet in particular has brought with it a renewed interest in the idea of personified computational artifacts attributed with a capacity for intelligent, interactive behavior. The dominant form of this project today is the promotion of computational agents that will serve as a kind of personal representative or assistant to their human users. The idea of personal agents was imagined perhaps most vividly in the form of "Phil," the bow-tied agent in Apple Corporation’s video "The Knowledge Navigator." But more reduced implementations of this fantasy now abound in the form of "knowbots" and other software agents.

One of the best known proponents of such artifacts is Pattie Maes of the MIT Media Lab. In a 1995 talk titled "Interacting with Virtual Pets and other Software Agents," Maes assures us that "the home of the future" will be "half real, half virtual," and that:

the virtual half of our home won’t just be a passive data landscape waiting to be explored by us. There will be active entities there that can sense the environment ... and interact with us. We call these entities software agents.

Agents are personified in Maes’ interface as cartoon faces, attributed with capacities of alertness, thinking, surprise, gratification, confusion and the like. As Maes describes them:

Just like real creatures, some agents will act as pets and others will be more like free agents. Some agents will belong to a
user, will be maintained by a user, and will live mostly in that user's computer. Others will be free agents that don't really belong to anyone. And just like real creatures, the agents will be born, die and reproduce... I'm convinced that we need [these agents] because the digital world is too overwhelming for people to deal with, no matter how good the interfaces we design... Basically, we're trying to change the nature of human-computer interaction... Users will not only be manipulating things personally, but will also manage some agents that work on their behalf.

Maes' text begs for a thorough rhetorical analysis. For my present purposes, however, I'll simply note the kind of rediscovery of the interaction metaphor (still, interestingly, in rather tenuous form) evident here. Moreover, thanks to the distributive powers of the internet, the computer other with whom we are to interact has proliferated into populations of specialist providers -- agents, assistants, pets -- whose reasons for being are to serve and comfort us, to keep us from being overwhelmed in the future workplace/homeplace of cyberspace.

What appears constant is the value placed on the artifact's capacity to be autonomous, on the one hand, and just what we want, on the other. We want to be surprised by our machine progeny, but not displeased.

My interest here is not to argue the question of machine agency from first principles, but rather to take as my starting point the project of tracing how the effect of machines-as-agents is being generated. This includes the translations that render former objects as emergent subjects, shifting interests and concerns across the human/artifact boundary. We can then turn to questions of what is at stake in these particular productions-in-progress, and why we might want to resist and refugiate them.

My original analysis/critique

Let me return for a moment to my earlier efforts to explore these relations of human and machine. I arrived at Xerox Palo Alto Research Center (PARC) in 1979, as a student of anthropology interested in relations of technology and work, and with a background in ethnomethodology and interaction analysis. My colleagues at PARC in the early 1980's were leading proponents of "knowledge representation," which included efforts to encode "goals" and "plans" as computational control structures. When executed these control structures would lead an artificially intelligent machine imbued with the requisite condition/action rules to take appropriate courses of action. Around 1981 they took on the project of designing an "intelligent, interactive" computer-based interface that would serve as a kind of expert advisor in the use of a photocopier characterized by its intended users as "too complex." Their strategy was to take the planning model of human action and communication then prevalent within the AI research community as a basis for the design. My project became a close study of a series of videotaped encounters, by diverse people including eminent computer scientists, attempting to operate the copier with the help of the prototype interactive interface. While the troubles that encountered in trying to operate the machine shifted with the use of this "expert system" as an interface, the task seemed as problematic as ever.

To try to understand those troubles better I developed a simple transcription device for the videotapes, based in the observation that in watching them I often found myself in the position of being able to "see" the difficulties that people were encountering, which suggested in turn ideas of how they might be helped. If I were in the room beside them, in other words, I could see how I might have intervened. At the same time I could see that the machine appeared quite oblivious to these seemingly obvious difficulties. My question then became: What resources was I, as in effect a fully-fledged intelligent observer, making use of in my analyses? And how did they compare to the resources to which the machine was allowed to have access to this question, I quickly realized, was at least in part that the machine was allowed to have access to a very small subset of the observable actions of its users. Even setting aside for the moment the question of what it means to "observe," and how observable action is rendered intelligible, the machine could quite literally only "perceive" that small subset of the users' actions that actually changed its state. This included doors being opened and closed, buttons being pushed, paper trays being filled or emptied, and the like. But in addition to those actions, I found myself making use of a very large range of others including talk and various other activities taking place around and in relation to the machine, which did not actually change its state. It was as if the machine were watching the user's actions through a very small keyhole, then mapping what it saw back onto a preconceived template of possible interpretations. Along with limitations on users' access to the design "script." In other words, I could see clearly the serious limitations on the machine's access to its users.

My analysis, in sum, located the problem of human-machine communication in continued and deep asymmetries between person and machine. I argued that so-called interactive programs like the expert system devised by my colleagues exploit certain characteristics of human conversation in ways that encourage attributions of interactivity to machines by their human interlocutors. At the same time, these attributions belie the profoundly different relations of person and machine to the unfolding situation, and their associated capacities to interact within and through it. So the instructions offered by an expert help system will be read by the machine's users as comments on the activity underway that should be intelligible, a strategy that proves extremely powerful for moving things along. Human interaction succeeds to the extent that it does, however, due not simply to the abilities of any one participant to construct meaningfulness but to the possibility of mutually constituting intelligibility, in and through the interaction. This includes, crucially, the detection and repair of misunderstanding. And the latter in particular, I argued, requires a kind of presence to the unfolding situation of interaction not available to the machine.

I want to be clear that my purpose here is not to argue from some a priori assumptions about essential aspects of "human nature" (pace Sack 1997, p. 62). As I hope will become clear, I take the boundaries between persons and machines to be discernibly rather than naturally effected, and to be always available for refiguring. It is precisely because the distinction between person and machine rests upon the traffic back and forth among the two terms that these questions matter. With that said, my observation in the book that follows is that while the language of interactivity and the dynamics of computational artifacts obscure enduring asymmetries of person and machine, people inevitably rediscover those differences in practice. This has led me to question the aptness of the metaphor of interactive technologies, and to wonder what other metaphors we might usefully employ.

Excluded middles

Lee and Brown (1994) point out in that framing the question as one of whether machines have agency like humans we presuppose a (non-protematized) Western view of what human agency could be. In particular, we accept that "to be human is to possess agency," and then proceed with the question of to whom or to what such attributions should be extended (p. 772, my emphasis). Instead, they argue, we should begin with the question of just what constitutes agency in any case, for humans or nonhumans. Anthropological efforts to establish criteria of humanness (for example tool use, language ability, symbolic representation) have always been contentious, challenged principally in terms of the capacities of other animals, particularly the nonhuman. I am interested in how these various cognitive behaviors. More recently the same kinds of critical arguments have been made in support of the human-like capabilities of artificially intelligent machines. Whether the concern is animals or machines, debates within this competitive frame inevitably turn on contests over just what counts as the behaviors in question, and who or what can be properly said to demonstrate them.
I believe that it was AI’s twin moves of writing a cognitivist rationality onto machines and then claiming their equivalence to humans that concerned me ten years ago in questioning the notion of interactive machines. My concern then, and now, has less to do with the implications of this move for our notion of what machines are than with the ways in which such a project reproduces traditional cognitivist assumptions about human beings. In setting up my critique, however, I now think that I fell into a classical humanist trap, defending against what I saw as AI’s appropriation of (a certain version of) intelligence and interactivity by effectively reclaiming (a different version of) those qualities for persons. Since then I have struggled with questions of how to preserve the sense of human-machine asymmetry that I developed in my analysis, while taking to heart the correctives offered in recent science and technology studies regarding the distributive and performative character of agency and the fallacy of essentialist human-nonhuman divides.

I think that part of the answer to this puzzle lies in what Latour (1993) has named the “Middle Kingdom” with respect to attributions of agency; that is, the space between translations on the one hand, and purification on the other. With respect to machine agency, translation references particular practices through which agency is shifted between humans and machines. Purification, on the other hand, indicates those practices through which boundaries are maintained as clear and distinct. In resisting the particular form of translations of intelligence and interactivity recommended by AI in the 70s and 80s, I fell back onto a kind of exercise of purification, attempting to maintain those qualities as exclusively human. I now understand that what we need is to, in Latour’s words, “direct our attention simultaneously to the work of purification and the work of hybridization” (ibid., p. 11), to develop a discourse that recognizes the deep mutual constitution of humans and artifacts without losing their particularities. Recognizing the interrelations of humans and machines, in other words, does not mean that there are no differences. The problem rather is how to understand our differences differently.

Agency reconsidered

I want to wander about a bit in Latour’s Middle Kingdom, then, in considering the question of agency. For those like Latour writing in the Actor-Network tradition, agency is understood as a material-semiotic attribute not locatable in either humans or nonhumans. Agency on this view is rather “a relational effect that is generated in different configurations of [human and nonhuman] materials” (Mulkay, p. 58). In the words of Callon, the network of interest is:

not a network connecting entities which are already there, but a network which configures ontologies. The agents, their dimensions, and what they are and do, all depend on the morphology of the relations in which they are involved (1997).

John Law (1996) reminds us with respect to practices of representation that signs and their referents are generated together, in a process that includes their simultaneous creation and separation from one another. That process of reflexive constitution is then deleted in the resulting sign/referent pair, each member of the pair being treated as independent of the other. Science studies, similarly, has provided compelling demonstration of the ways in which apparatuses of observation actively constitute both subjects and objects. Physicist Karen Barad, for example, emphasizes that apparatuses themselves are materially and discursively reiterated and/or rearranged through disciplinary practice (1998). Elaborating on this process empirically, Charles Goodwin’s analyses of what he terms “professional vision” (1994) focus on the social/material interactions through which practitioners learn to see those phenomena that constitute the objects of their profession, at the same time that the phenomena are constituted as disciplinarily relevant objects through occasioned performances of competent seeing.

Goodwin analyses archeological knowledge, for example, as relations between particular, culturally and historically constituted practices and their associated materials and tools. It is out of that relation, quite literally, that the objects of archeological knowledge and the identity of competent archeologists are co-constructed. As a simple case, Goodwin describes how archeologists in the field use a particular artifact, the Munsell chart of universal color categories, to code color features of dirt in a particular site (see also Latour 1999, pp. 58-61). They do this by taking a trowel of dirt, wetting it, and holding it behind holes cut in the chart. From this we see both the artfulness of the color chart’s design and how its use requires the embodied juxtaposition and skilled reading of dirt in relation to chart. Goodwin also describes the problems of actually using the chart (the colors of real dirt, of course, never quite match these ideal types) and its role in bringing archeologists’ perception into the service of a particular organizational/bureaucratic endeavor, that is, professional archaeology.

While not concerned specifically with interactive machines, Goodwin’s example provides a baseline argument against attributions of agency to machines or to artifacts. In the case of machines that are created and used in a field by humans, there is no question of what Goodwin calls the “chart-in-use embodies archeological knowledge.” This gives us a different way of understanding the problem of attributions of knowledge and agency to interactive machines. Unquestionably, the distinctive dynamics of the latter invite such attributions in ways that non-computational technologies like the chart do not: this is at least one reason that discussion of “artificial agents” is now so prevalent in our discourse. The problem is less that we attribute agency to computational artifacts, than that our language for talking about agency, whether for persons or artifacts, presupposes a field of discrete, self-standing individuals.

Modernist epistemologies, in other words, treat agency as something both independent of and contained within singular entities, whether humans or artifacts. In this respect, the language of intelligent artifacts remains consistent with a tradition that treats separation and autonomy, rather than relatedness, as the mark of humanity. Somewhat paradoxically, these analyses suggest that in this respect it is precisely the persistence of the modernist human/machine divide within the discourse of intelligent artifacts that makes machine agency so compelling. That is, having systematically erased the originary connection of humans and machines, many technologists seem to be worried that machines once separated from us are now rendered “lifeless,” and by implication less. They need to be restored to humanness, in other words, to be made like us, in order that we can be reunited with them.

In this respect, also, the interactive machine might be seen as the next logical step after the neutral instrument described by Shapin and Schaffer (1985) in the emergence of the observational sciences. The instrument was taken to speak on behalf of the natural object, albeit that its words were still in need of translation by the scientist. Now the artifact speaks for itself, while similarly erasing, rendering invisible, its creator. As Shapin and Schaffer describe the autonomy of scientific facts:

The matter of fact can serve as the foundation of knowledge and secure assent insofar as it is not regarded as man-made. Each of Boyle’s three technologies worked to achieve the appearance of matters of fact as given items. That is to say, each technology functioned as an objectifying resource ... The world of subjects and objects was in place, and scientists were on the side of objects (1985, p. 77)

It may be obvious why an observational science would be interested in erasing the place of human practice in the emergence of its naturalized objects. But why, in creating computational technologies, do designers increasingly evidence a desire to naturalize them, to obscure their artifactuality? I would suggest in part that it is a kindred desire to that which inspired the development of objectivist science; that is, the desire to disappear and put in our places something transcendent, existing independently of our actions. Richard Doyle (1994) has proposed that the vitality and autonomy of computational artifacts emerge through the camouflage of the networks that

http://www.comp.lancs.ac.uk/sociology/soc040ls.html
support them. By "camouflage" he means encoding in ways that depend crucially on obscuring the embodied activities that are the conditions of possibility for artificial life; that is, "machines, bodies, desires, and other practices on all sides of the screen" (p. 7). In contrast, Doyle argues that the vitality of artifacts comes from "a massive assemblage of machines, users and rhetorics that semiotically and materially distribute their "vitality effect"" (p. 17).

We catch a glimpse of those hidden constituencies in an article by Rosalind Pickard on what is named "affectionate computing" (1997). In a footnote Pickard reports that in lab experiments with students playing the computer game DOOM signs of stress came less with the appearance of a new deadly enemy than during times when students were experiencing difficulty configuring the software. What I find so interesting about this is that it suggests that genuine stress occurs when troubles are "real life," affecting lived accountabilities (in this case, for compliance and competence) and consequences (for getting on with the game/task). For the researchers, this result is reported only as an amusing anecdote, however, noted en passant.

Cultural analysts like Doyle and Julian Bleecker (1994) analyze the compulsion of the virtual as evidenced in widespread fascination with the Maxis Corporation's Sim™™ games and with ALife research. Both point to the opportunities these technologies afford their creators/users for a kind of simultaneous safety with risk, a transcendence over the "world" in question at the same time that one is somehow inscribed within it, engaged with an autonomous and therefore not fully predictable other. This produces a simultaneous sense of control over the virtual from "outside" while being "inside," controlled by larger and more powerful forces. The result is a controlled simulation of the experience of not being in control; hence, the best of both worlds.

**Mutual constitutions**

Having looked briefly at the problem of how agency is attributed to humans or machines, I want now to look more closely at the premise that humans and artifacts are mutually constituted. This premise of technoscience studies has been tremendously valuable as a corrective to the modernist view of humans and machines as autonomous, integral entities that must somehow be brought back together and made to "interact." But at this point I think the sense of "mutuality" warrants a closer look.10

First, as technoscience studies have shown, mutual constitutions of humans and artifacts do not occur in any singular time and place, nor do they create fixed human/artifact relations. Rather, artifacts are produced, reproduced and transformed through highly specialized "labours of division," in Law’s phrase (1996), that involve continuous work across particular occasions and multiple sites of use. This work of production and reproduction across time and space results in very diverse assemblages, involving participants with different histories, relations of familiarity or strangeness, and the like. As Diane Mulcahy points out (1997), it is the increasingly extensive distribution of technologies and the range of variations across user/machine pairings that render protocols, standards, instructions and the like necessary to the successful production and reliable reproduction of human/artifact interactions.

Second, mutualities are not necessarily symmetries. My own analysis suggests that persons and artifacts do not appear to constitute each other in the same way. In particular, I would argue that we need a rearticulation of asymmetry that somehow retains the recognition of hybrids, cyborgs, and quasi-objects made visible through technoscience studies, while simultaneously recovering certain subject/object positionings – even orderings – among persons and artifacts and their consequences. In his essay "On actor-network theory: A few clarifications," Latour reiterates that:

> An ‘actor’ in ANT is a semiotic definition – an actant – that is, something that acts or to which activity is granted by others. It implies no special motivation of human individual actors, or of humans in general. An actant can literally be anything provided it is granted to be the source of an action (1997, p. 4, original emphasis).

I want to focus our attention on this last phrase; that is "provided it is granted to be the source of an action." Who or what within the situated practices of constituting agency does this work of "granting" of actor status? Does that form of agency distribute evenly across the human-nonhuman network?

Analyses that describe the active role of artifacts in the configuration of networks generally seem to imply other actors standing just offshore for whom technologies act as delegates, translator, mediators; that is, human engineers, designers, users, etc. I want to suggest that the persistent presence of designers/users in technoscientific discourse is more than a recalcitrant residue of modernism: that it reflects a durable asymmetry among human and nonhuman actors. On that basis I am after an ontology that can tie humans and nonhumans together without erasing the culturally and historically constituted differences among them. Those differences include the fact that persons just are those actants that conceive and initiate technological projects, and configure material-semiotic networks, however much we may be simultaneously interpellated into and through them. And as well, I want to keep in view the ways in which it matters when things travel across the human/artifact boundary, when objects are subjectified (e.g. machines made not actants but actors) and subjects objectified (e.g. practices made methods, knowledges made commodities).

Taking ultrasound technology as a case in point, Barad (1998) is working this problem by maintaining a distinction between phenomena and the knowledge practices that render those phenomena as particular social/material objects. On this view phenomena are in continuous production, reification and re-fashionation through their "intra-actions" with stable and shifting configurations of apparatus. Applied to the question of agency, I want to argue that we are witnessing a reiteration of traditional modernist notions of agency, at the same time – even through – the intra-actions of that notion with new computational media.

**Humans, machines and Others**

Recent writings on the human-nonhuman boundary from within the humanities and social sciences have been concerned to recover for those enterprises aspects of the world – specifically "nature" and "technology" – so far excluded from consideration as proper sociological subjects (see for example Latour 1992). My own project is clearly indebted to these efforts, which provide the reconceptualizations needed to move outside the frame of purification and opposition on which modernist theorizing and disciplinary boundaries depend. My engagement with these concerns, however, came first in the context not of the humanities but of technology and engineering, where the situation is in important respects reversed. Far from being excluded, nature and technology in the regimes of research and development are centered, while "the social" is separated out and relegated to the margins. It is the centered machine in this context that creates its marginalized human Others.11

The way forward, clearly, is not simply to recenter the social as the proper subject matter of the natural sciences and engineering, as yet one more step in a kind of pendulum swing across the disciplines. Rather, the problem lies in the very constitution of the sciences as
either human or natural, social or technological. More relevant to the concerns taken up here, the problem lies with any expansionist enterprise that seeks to project one particular frame as inevitable: in this case, the frame of a world in which to be overwhelmed is natural, and in which machines must acquire human agency if we are to maintain effective sociality.

**Difference without separation**

My concern, in sum, is that far from recognizing the deep interrelations of humans and artifacts, contemporary discourses of machine agency simply shift the site of agency from people to their machine progeny. This does not mean that human power is lost. Rather, as in claims for the independent voice of nature in the practice of positivist science, the assertion of autonomy for technology depends upon the obscuring of human authorship. And just as claims for nature's autonomy ensure the power of the scientist, it is precisely in the obscuring of authorship that the power of the technologist is renewed. The history of the materialization of machines is intimately and intricately tied to the history of "regulatory (power/knowledge) practices and their participatory role in the production of bodies" (Barad 1998, p. 23). I want to keep our eye on the ways in which autonomous machine agency, however subserviently constructed, might be consistent with regulatory practices aimed at producing certain kinds of humans (employers, workers, consumers) and excluding others.

As a way of recovering difference without separation, recent feminist theorists have proposed a form of what Barad names "agental realism" (1995, 1998), a position that takes reality to be "sedimented out of the process of making the world intelligible through certain practices and not others" (1998, p. 20). If we are responsible for what exists, what is the reality that current material/discursive practices regarding computational artifacts make intelligible, and what is excluded? To answer this question Barad argues (following Haraway) that we need a simultaneous account of the intra-actions of humans and nonhumans and of their asymmetries and differences. This requires remembering that boundaries between humans and machines are not naturally given but constructed, in particular historical ways and with particular historical consequences. As Barad puts it:

> Boundaries are not given by their enemies, they are necessary for making meanings, but this does not make them innocent. Boundaries have real material consequences – cuts are agentially positioned and accountability is mandatory ... Our goal should not be to find less false boundaries for all spacetime, but reliable, accountable, located temporary boundaries, which we should anticipate will quickly close in against us (p. 187).

It is on this understanding that I am arguing for a certain restoration of the boundaries between humans and machines, not as the restoration of some natural order but as a historically and culturally specific act of refiguring. The point of the refiguring is twofold. First, to recover certain qualities – interactivity, intelligence – from trivialization. And second, to restore authorship and thereby accountability to our relations with artifacts. As Barad argues:

> Which shifts actually occur matter ... We are responsible for the world in which we live not because it is an arbitrary construction of our choosing, but because it is sedimented out of particular practices that we have a role in shaping (1998, p. 16).

It is in this spirit that my aim here is not to prohibit uses of the figure of interaction to describe our relations with computational artifacts, but rather to heighten our degree of critical awareness. Machines (along with the various other materials that participate in making realities) can be understood not only as passive objects but as "active actants" (Lee and Brown, 1994). The point, however, is that the price of recognizing the agency of artifacts need not be the denial of our own. Agency – and associated accountabilities – reside neither in us or nor in our artifacts, but in our intra-actions. The question, following Barad, is how to figure our differences in such a way that we can intra-act responsibly and productively with and through them.

**Endnotes**

1 The word "imaginary" is a term of art in recent cultural studies (see Marcus 1995, p. 4, Verran 1998). It shares the nice quality with the more colloquial term "imagination" of evoking both vision and fantasy. In addition, however, it is used to refer to the ways in which how we see and what we imagine the world to be is shaped not only by our individual experiences but by the specific cultural and historical resources that the world makes available to us, based on our particular location within it. The general location of the imaginary I am citing here are the circulating information and communication networks of what we might call the hyper-developed countries of Europe and North America.

2 For closely considered arguments regarding the problems with this premise see e.g. Dreyfus 1992; Collins 1990; Button et al 1995.

3 For a partisan review of recent developments see Clark 1997.


5 I am thinking in particular of the rhetorics of ownership, management, free agency and service in relation to end of the century global capitalism.


7 The phrase "material-semiotic" is meant to indicate the ways in which the "natural" and the "cultural," or the material and the meaningful, are inextricably intertwined.

8 I am grateful to Naoki Ueno for referring me to this passage, and for other helpful comments.

9 For a related argument regarding the modernist, humanistic discourse of AI see Sack 1997.

10 My thanks to Yasuko Kamatoko for bringing this question to my attention.

11 Others are defined by Lee and Brown (1994) as "all those entities and areas of inquiry that are rendered problematic by expansionist projects, be they formally political or theoretical" (p. 773).

**References**


This page is maintained by Pennie Drinkall, Cath Gorton and John Law

Last change 16 June, 2001