Heteronomous Humans and Autonomous Agents: 
Toward Artificial Relational Intelligence

Abstract
The notion of “autonomy” is a central concept and a generative metaphor in many AI approaches and systems. It also embodies a tension that is inherent to a persistent and sustained trend in AI that can be called “autonomist AI,” whose objective is to build systems that are, on the one hand, complex and intelligent enough to initiate actions on their own, and, on the other, simple enough to be understandable and controllable by human beings. Tracing the origins of autonomist AI in some of the basic tenets of modernity, I show how the above tension is manifested in theories of affect, morality, and knowledge. I argue that these tensions arise largely because of adherence to a substantivist view, and propose a reversal to what I call Artificial Relational Intelligence.

Introduction
On February 16, 2011, law enforcement officers in California arrested Chris Butler, also known as P.I. Mom, along with Norman Wielsch, a Narcotic Enforcement Team Commander, on charges of embezzlement, burglary, and conspiracy, as well as drug-related crimes. Butler, the founder of a company that provided service to women suspicious of cheating husbands, had signed a contract with Lifetime Television for a reality TV show on the same topic, where his alleged operations on behalf of betrayed wives were featured as true stories. His TV shows, as such, crossed the boundary between fiction and reality in a delicate manner, involving some of his real-world clients as actors. In parallel with this, in collusion with Wielsch, he was involved in illegal operations involving the resale of narcotics confiscated by law enforcement. In an interesting turn of events, which led to his arrest, TV and FBI cameras were capturing the same scene that was player out at once as real and as fantasy (Bearman 2012).

In an interview with reporters after his arrest, Butler mused, “The problem with people is that they want to believe you. You give them a little, and they take it from there.” What he calls a “problem,” however, is in fact a well-understood human trait with its own advantages and downsides. On the advantageous side, linguists, psychologists, and cultural anthropologists have shown that the capability of human beings to attribute beliefs and intentions to others is at the root of human sociality (e.g., Levinson). On the flip side, however, this trait can lead to attributions that go beyond the realm of the real, the sensible, or even the probable. This latter aspect is well known to AI practitioners who, as far back as 1976, came to notice the ubiquity of the Eliza Effect in human-machine interactions (Weizenbaum 1976). Initially considered as “the susceptibility of people to read far more understanding than is warranted into strings of symbols – especially words – strung together by computers” (Hofstadter 1995), the phenomenon seems to be much more general and prevalent, extending beyond strings of symbols to include actions, appearances, and affects. Ekbja (2008) proposes Generalized Eliza Effect (GEE) to refer to this broader phenomenon, and to show the penchant among AI practitioners to inadvertently use GEE to deal with tensions that often arise between their scientific and engineering aspirations, and to sometimes lure
people into unrealistic claims about AI systems, committing in the process what he calls the Attribution Fallacy. However, as mentioned above, attribution has a light and positive side that also needs to be considered in AI, and this is what I would like to pursue here. The concept of “autonomy” provides a useful angle for this purpose.

This is how the article proceeds from here. We start with a brief examination of “autonomous systems” in AI and the way they are designed and represented in AI literature — in particular, the tendency in AI to erase the “supplements” that surround and enable the functioning of AI systems. Then, we trace the origins of this tendency in a substantivist philosophy, discuss its articulation in AI theories of affect, morality, and knowledge, and show the paradoxes that it faces in theory and practice. Finally, situating these paradoxes in the broader socio—historical developments of modernity, we explore the possibility of a reversal that could allow us to deal with them in a productive way.

**Autonomous Systems in AI**

The idea of “autonomy” as a generative metaphor has shaped a great deal of thinking and research in AI. A persistent trend, which can be called “Autonomist AI,” has survived paradigmatic changes in approach and technique. This trend is perhaps most explicitly represented in the research that is conducted on “autonomous agents.” The standard definitions of these agents describe them as follows:

> Autonomous agents are computational systems that inhabit some complex dynamic environment, sense and act autonomously in this environment, and by doing so realize a set of goals or tasks for which they are designed. (Maes 1995: 108)

> An autonomous agent is a system situated within and a part of an environment that senses that environment and acts on it, over time, in pursuit of its own agenda and so as to effect what it senses in the future. (Franklin and Graesser 1996: 25)

Although somewhat different in details — e.g., in how they talk about the agent’s “agenda,” as something that is “its own” or for which it is “designed” — the common premise in both of these (and other definitions) is that they conceptualize autonomous agents as inhabiting an environment. However, as Elisabeth Wilson (2010) points out, “AI researchers have perhaps been more attentive, philosophically, to the autonomy that emerges for this agent (“its own agenda”) than in how this autonomy has been and continues to be constituted through relations to a milieu (“within and part of an environment”).

In addition, to the extent that “environment” is featured in accounts of autonomous agents, it acquires a very narrow and often underdetermined character. That is, environment is understood as a set of features and properties that the agent senses and acts upon. What is often lost in this conceptualization is the fact that the environment also supports the agent in carrying out its actions (Vygotsky 1978). This is a basic but important insight.
the flouting of which has serious implications for how we think about the world, about the nature of intelligence (natural or artificial), and about how we design systems (Agre 1997).

Jacque Derrida makes this point using the playful rubric of “dangerous supplements,” which, according to him, have two different significations: “A surplus, a plentitude, enriching another plentitude, the fullest measure of presence,” or alternatively, “an adjunct, a subaltern instance which takes-(the)-place [tient-lieu]” (1976: 144-45). In other words, that which supplements (e.g., environmental scaffoldings) can be simply a surplus that can be ignored but also as something that substitutes. Here Derrida is attracting our attention to the important role of supplements, the things that sometimes “take the place” of something.

Common practice in AI tends to focus on the first signification of supplements at the expense of the second one. This is best manifest in public portrayals of AI systems. A cursory comparison of Figures 1(a) and 1(b) shows, for instance, how the portrayal of the humanoid robot in the media as a solitary and “autonomous agent” erases the complex surrounding support structure provided by human beings, devices, and infrastructures, in the same fashion that media portrayals of chess contests between humans and machines take out of the picture the critical role of humans as surrogate players, spectators, analysts, and so forth (see Figures 2a, b, c; cf. Suchman 2007). The point, however, is more than the simple issue of media representation, and is symptomatic of a deep-rooted philosophy that can is referred to as “substantivism.”

According to this view, an attribute or property P (e.g., intelligence, affect, morality, expertise, etc.) of the members of a social group G (humans, animals, robots, etc.) is the real and substantive possession of the members of that group (Collins & Evans 2007). Accordingly, the behaviors of the members of G can be largely determined by their membership in the group. An opposing view, which is sometimes called “relational,” would posit that P is a set of attributions by the social group G — P is in the eye of the beholder, in other words. You are an expert on a topic such as medicine, for instance, if medical experts judge you as one. By the same token, you are “intelligent” if the members of the group of intelligent entities (humans?) judge you to be intelligent. This radical version of the relational view might be difficult to defend because, in its attempt to avoid the essentialism of the substantivist view, it puts an undue emphasis on outside relationships. A different version of the relational view, however, is conceivable, which would go like this: P is crucially dependent on the performative capabilities of other entities in G that are “outside” the individual. From this perspective, P is in the act of performance and participation; it is in the capability to interact and relate meaningfully with relevant others; or, to put it most succinctly, P is a mediation (Ekbia 2009).

Both of these views — the substantivist and the relational (in its various versions) — can be traced in the history of AI, starting with its early origins in Alan Turing who speculated that “instead of trying to produce a programme to simulate the adult mind, why not rather try to produce one which simulates the
child’s?’ (1950). Basically, Turing, advocating a relational view, wonders if by putting childlike AI systems in the right situation we might be able to turn them into recognizably intelligent beings among adult humans. In fact, in this reading, the Turing Test itself is a clear demonstration of the relational view, as it is organized around the notion of “performance.”

More recent incarnations of the relational view can be found in the interactionist thread in AI, which despite the best efforts of its advocates has not been able to win a noticeable space in AI research and practice, giving way to the dominant substantivist view (Agre 1997b). Many different parameters account for this course of development in AI: the philosophical tradition and the dominance of Cartesian dualism, the development of computing and the prevalence of the von Neumann architecture, the individualist and sensationalist cultural milieu of the time, the largely militaristic funding structure in the US, close academic affiliation with cognitive psychology as opposed to social sciences, and so forth (Ekbia 2008). We do not have space to examine all of these parameters here. However, to understand the origins of that particular strand of the substantivist view that we called autonomist AI we need to step back and examine the historical context of AI developments as a modernist project.

**Human Heteronomy and the Paradox of Modernism**

Autonomy, as a kind of freedom, was introduced into modern philosophy in contrast to “negative” freedom — the right to do as one pleases, unimpeded by others (Rousseau 1997). It had to do with the willingness of citizens to surrender some of their rights as individuals and “to think of their social membership as essential, not merely accidental, to who they are” (Neuhouser 2011: 479-80). As such, at its origin, autonomy was not conceived individualistically. In particular, it was not related to the ideal of autonomous judgment in the sense of deciding for oneself. Furthermore, the kind of moral freedom brought about by autonomy was considered by Rousseau to be superior to other natural or civic freedoms: “for the impulsion of mere appetite is slavery, and obedience to the law one has prescribed for oneself is freedom” (1997: I.8, iii). Given the insatiable character of human desires — something that separates us from all other animals — he finds this positive moral freedom to be the only way for humans to become “truly the master of himself [sic]” by submitting to a law that they themselves have prescribed.

The notion of self-prescribed law brings us to the etymology of the term — *auto* (“self”) and *nomous* (“law”) — and its inherent relationship to democracy as a form of governance. However, it also brings out an inherent tension and paradox in Rousseau’s formulation, which commentators, philosophers, and modern societies in general have dealt with ever since. This paradox has to do with another fundamental feature of the human condition — namely, our dependence on others:

> Plants are fashioned by cultivation, and men by education. If man were born big and strong, his size and strength would be useless to him until he had learned how to use them; they would create prejudice against him; and, left to himself, he would die miserably before knowing his needs. We complain of
The state of infancy; we do not see that, if man had not begun
by being a child, the human race would have perished.”
(Emile: p. 67) ii

The emphasis that Rousseau puts on education and, more broadly, on the
ongoing neediness and dependence of human beings on other fellow humans
reveals a central dilemma of his thinking, but also a central paradox of modern
times — namely, the paradox of autonomy versus neediness, of voluntarism
versus coercion:

Rousseau’s system is somewhat paradoxical… The standard to
which will must conform — is itself non-voluntaristic;
therefore, contradictory. The standard that gives will its object
is the very negation of voluntarism. (Riley, 1982: 121)

This neediness and dependence of human beings on others, coupled with
inevitable asymmetries and social inequalities among individuals, creates the
ground for social domination and the undermining of autonomy — hence, the
notion of “dangerous supplement” that Derrida has playfully written about
(see previous section). The danger calls for a political solution, of which
Rousseau’s Social Contract is a foundational articulation. Despite the danger,
however, Rousseau does not counsel self-sufficiency because the independent
and unattached beings of this imaginary scenario would be not only devoid of
affect and love but also language, reason, and virtue — even selfhood itself
(Neuhauser 2011: 487). This is at the root of the paradox of Rousseau’s system
that attracts our attention to the opposite term heteronomous: subject to
different laws, according to the Oxford Dictionary. iii

Rousseau’s concept has influenced political and moral philosophy, and also
political thought, in the last three centuries, particularly the political ideology
of various versions of liberalism that have come to play a dominant role in
contemporary Western societies. Michel Foucault discusses the paradox of
liberalism as having to do with the tension between freedom and coercion in
liberal ideology. Freedom is the purported historical gift of liberalism to
humanity. However, liberalism does not guarantee, provide, or even respect
freedom. What liberalism purports to do is to produce what we need to be free —
the conditions, organizations, instruments, etc. that create the possibility for
the production of socio-economic, legal, and political freedoms. And the way it
does this is by “the establishment of limitations, controls, forms of coercion,
and obligations relying on threat, etcetera” (Foucault 2008: 64). Consequently,
freedom in liberal societies has to be constantly produced and accomplished; it
is not a given. This duality between freedom and the coercive instruments and
techniques for producing freedom led to what Foucault called the “paradox of
liberalism.”

In summary, the paradox of modernism finds various shapes and forms in
modern societies, which seek to establish a social order that, on the one hand,
provides individual freedom and autonomy, and, on the other, subject those
same individuals to a collective will that is variously conceived and
implemented depending on the dominant ideology and political order. In other
words, modernity creates individuals who are autonomous in certain ways and
heteronomous in other ways; it takes back with one hand something that it
gives to the individual with the other.

**Autonomy and the Paradox of AI**

Given this state of affairs, perhaps we can consider the general interest in
“autonomy,” especially in the context of AI research, as a projection of the
modern humanity’s desire for independence and freedom, on the one hand, and
their despair in understanding their neediness and dependence. Whether or
not we take it as a projection, one can discern a tension in AI, where
practitioners seek to build systems that are intelligent, powerful, and
autonomous, on the one hand, but that can also be intelligible, flexible, and
predictable (if not controllable). At one level, therefore, the paradox can be
seen in the tension between “intelligence” and “intelligibility” (Suchman 2007).
At another level, it can also arguably be seen in how AI research has dealt with
specific topics such as affects, ethics, knowledge, and so forth. In dealing with
the paradox, AI research has largely tilted toward a monadic understanding of
these topics, underplaying their dyadic and social aspects. In the following
discussion, I would like to illustrate how a substantivist philosophy underlies
this bias in AI research.

**Affect: Monadic or Dyadic?**

Affects and emotions have been of interest to AI from early on, starting with
Turing and continuing since then. Herbert Simon, for instance, emphasized the
relationship between affect and cognition as early as 1960’s (Simon 1967).
Generally, both substantive and relational views on affects can be identified in
AI, and in computing in general. While the former considers affects as discrete
states, internal to the individual, and transmitted in a loss-free manner from
people to others or to computational systems, the latter understands them as
dynamic, culturally mediated, and socially constructed and experienced
(Boehner et al. 2005). The substantive view is perhaps best represented in
Donald Norman (xxx)’s model, which closely mimics Card et al (1995) ’s
model of human mind (see Figure 3). The relational view, on the other hand, is
best captured in the idea of an affective loop that posits that affect flows in
dyadic interactions between the individual and another person or system.

The relational view of affect has a long and established history in
psychoanalysis, as well as social and developmental psychology, where the
“self” is believed to emerge in interactions with others, and not as built-in.
Feelings, according to Rosaldo (1983), “are not substances to be discovered in
our blood but social practices organized by stories that we both enact and tell.”
In a similar fashion, psychotherapy also seeks to build autonomy through
relatedness. In fact, the practice of psychotherapy itself can be understood as a
process through which autonomy is accomplished through the inter-subjective
regulation of affect. This represents a paradox of psychotherapy (Wilson 2010:
85):

Ideally, psychotherapy builds not sovereign subjects but
individuals who can both recognize their own self-states and
modulate those states in relation to others. The inter-subjective
regulation of affect is one of the means by which such
autonomy emerges. In this sense, psychotherapy is an instance
of a more general dynamic: all modes of autonomy are acquired affectively and relationally.”

Child psychologists have similarly argued that infant development is a dyadic process. The famous proclamation by Donald Winnicott (1966) speaks to this point: “There is no such thing as a baby” — i.e., a baby cannot exist alone, but is always and essentially part of a relationship. Along the same lines, Fonagy et al. (2002: 4) argue as follows:

The baby’s experience of himself as an organism with a mind or psychological self is not a genetic given. It is a structure that evolves from infancy through childhood, and its development critically depends upon interaction with more mature minds, who are benign and reflective in their turn.

These observations lead to a dyadic notion of affect that operates through what can be called an “affective loop,” which is at work even in those cerebral situations that seem to be far removed from emotional attachment — e.g., in championship chess. For instance, Kasparov’s defeat by the Deep Blue can be equally attributed to the power of the machines as to his failure to establish an affective loop with the opponent (Figure 2d). As Wilson (2010: 19) points out:

Kasparov’s customary tactics of intimidation aren’t simply a projection onto the opponent — a kind of one-sided attack. Rather, Kasparov, when he is most effective, recruits his opponents into an affective intimacy, albeit intimacy rooted in fear. The pertinent issue is not the emotion in Kasparov (Is he angry? Is he afraid?), as if he operates as an affective monad (an isolated talent); rather it is the emotional relationality between Kasparov and his opponent that governs, in part, whose intelligence will prevail.

One can, indeed, read Turing in a similar manner. Rather than emphasizing particular affects such as fear, joy, or anger, and their instantiation in machines, he is interested in how affectivity cultivates relationships between agents (specifically between humans and machines):

Without much in the way of available theory about the development of mind, Turing nonetheless seems to intuit that interiorities (human and artificial) are built mutually, intersubjectively… At important junctures [e.g., the biographical anecdote reported by Hodges (1983) in the conversation between Turing and his partner Arnold Murray about dreams], Turing imagines thinking and feeling to be chiasmatically related rather than opposed and disjunctive. (Wilson 2010: 21)

Against these proposals and precursors, the approach of Autonmist AI to the question of affects and emotions has been largely monadic, seeking to inscribe and attach affects to individual artifacts, which leads to a neglect of the dyadic character of affect and the intersubjective aspects of emotion. Generally, in AI
research, emotions reflect drive states but do not have much motivational force by themselves. This is, for instance, how the affective mechanisms of the “baby” robot Kismet are described by its designers (Breazeal & Scassellati 2000: 55):

When in the homeostatic regime, a drive spreads activation to those [emotional] processes characterized by positive valence and balanced arousal. This corresponds to a “contented” affective state. When in the under-stimulated regime, a drive spreads activation to those processes characterized by negative valence and low arousal. This corresponds to a “bored” affective state that can eventually build to “sorrow.” When in the overwhelmed regime, a drive …

In brief, Kismet operates in three regimes, where different levels of arousal give rise to appropriate emotional states. While quite novel, the approach in the design of Kismet is based on a theory of drives and affects, which is in contrast with the views of psychologists such as Sylvan Tomkins, who also happened to have had a long-term interest in social and affective robots, long before AI projects such as Kismet came into being. According to Tomkins, our behaviors are largely regulated by affects, which are sustained and general in character, as opposed to drives, which are spatially and temporally specific and hence weak in motivating behavior. Affects, as such, take priority over drives. The hunger drive, foundational to behaviorism and also to Freud’s theory of sexuality, for instance, is not powerful by itself. It becomes urgent (and so able to compel behavior) when it is amplified by, say, distress or enjoyment. It can similarly be attenuated or blocked by disgust or fear. Sexual drive is similarly diminished by shame, fear, apathy, or surprise. Humans act “not only by a responsiveness to drive signals but by a responsiveness to whatever circumstances activate positive and negative affect” (1962: 22). Therefore, “the creation of a humanomaton would require an affect system,” according to Tomkins (1963: 18), not a drive system as we saw in Kismet.

Kismet, however, suffers from a more serious shortcoming in terms of its affective behavior. Its affective states, expressed in nine different facial expressions (happy, sad, angry, etc.), lack an important intersubjective affect that operates in the space between aspiration and dependency — namely, “shame.” A negative affect triggered when positive affects of interest and enjoyment are obstructed (I want, but …), shame brings to halt facial communication with the eyes down, head down, and blushing. Many psychologists and cultural anthropologists have argued that shame is the quintessential intersubjective and dyadic affect, which regulates social behavior — an aspect that should have been of natural interest to the designers of Kismet.

The dominance of drives instead of affects and the absence of shame have led Wilson to conclude that:

In much of the Kismet project, internal states are deduced rather than felt, intellectually discerned rather than
sympathetically known. Mutuality is executed rather than sensed... The Kismet project is drawn to emotion, but then loses its nerve. (p. 55)

... one thing we might be able to declare on examining an autonomous robot like Kismet is that there is no such thing as an autonomous robot” (Wilson: 74).

**Morality:**
Similar issues in the area of morality and machine ethics, which is concerned with ensuring that the behavior of machines toward human users, and perhaps other machines as well, is ethically acceptable. Examples of research where this might be an issue are autonomous robots (social and home-based robots), autonomous driving, and autonomous drones and armed vehicles. In dealing with these topics, we also see the substantive and relational views in opposition. A substantivist sees morality as the outcome of individual reasoning, embodied in implicit or explicit ethical agents that are programmed to behave ethically, respectively, without an explicit representation of ethical principles or based on the calculation of the best action in ethical dilemmas using ethical principles (Moor 2006). A relational view, on the other hand, considers morality as the mutual accomplishment of individuals embedded in worlds of social values. From this perspective, guilt and shame are seen as moral affects necessary to constrain the individuated self from dangerous and asocial acts of impulse, lust, and violence (Rosaldo 1983).

The substantivist view has led some researchers to contend that “machines with a level of autonomy requiring ethical deliberation are already with us, and that their number and level of autonomy are likely to increase. Therefore, “the liability already exists, and machine ethics is necessary as a means to mitigate it” (Anderson & Anderson 2007). This leads these authors to the rather utopian idea that the ultimate goal of machine ethics is to create a machine that is an explicit ethical agent — a “Humans 2.0,” which would be a better version of human beings:

Machines, though, might have an advantage over human beings in terms of behaving ethically... human beings, as biological entities in competition with others, may have evolved into beings with a genetic predisposition toward unethical behavior as a survival mechanism. Now, though, we have the chance to create entities that lack this predisposition, entities that might even inspire us to behave more ethically.

Interestingly, the substantivist view behind the AI approach allows these authors to make the more general essentialist claim about human beings as genetically disposed toward unethical behavior — a claim that would not sit well with those who expect a more balanced and nuanced image of human morality. Ironically, the same substantivist view gives rise to dystopian fears of a “postbiological” future, where, “The human race has been swept away ... usurped by its own artificial progeny” (Moravec 1988). It is, indeed, hard not to read the resonances of an ancient mythology in these images. Commenting
on the related case of HAL in Arthur Clark’s celebrated fiction, Bloomfield & Vurdubakis (2008) wonder:

For what is HAL’s crime but the Original Sin? Moderns, having created thinking machines in their own image, immediately expect that these machines will — just like they themselves did — attempt to usurp the powers of their creators… It is perhaps paradoxical but not unexpected that AI, the enterprise that is said to epitomize the workings of reason, is at the same time so heavily mythologized.

With these underpinnings, it is indeed hard to find work in machine ethics that could be considered a project in relational ethics. By way of contrast, however, it might be useful to point out Tomlin’s vision on morality (1963: 216):

Just as contempt strengthens the boundaries and barriers between individuals and groups, and is the instrument par excellence for the preservation of hierarchical, caste and class relationships, so is shared shame a prime instruments for strengthening the sense of mutuality and community whether it be between parent and child, friend and friend, or citizen and citizen. When one is ashamed of the other, that other is not only forced into shame but he is also reminded that the other is sufficiently concerned positively as well as negatively to feel ashamed of and for the other.

**Knowledge and Expertise: Tacit and Explicit**

As a final illustration of the contrast between substantivist and relational views, one can mention AI models of knowledge-intensive or expert systems. A substantivist view starts with the premise that knowledge can be explicitly captured in propositional statements connected by logic-based rules, whereas from a relational perspective tacit knowledge is the linchpin of human cognition. The question of tacit knowledge is probably most widely discussed by social scientists such as sociologists of science. Influenced by Wittgenstein’s philosophy of language and Michael Polanyi’s work on “tacit knowledge,” many of these commentators emphasize that a significant part of human knowledge is not directly accessible to conscious thought. More recent work categorizes tacit knowledge into different types, showing the intricacies of acquiring, sharing, and even talking about knowledge. Collins and Evans (2009), for instance, introduce what they call the “Periodic Table of Expertise.” Although the term itself is not perhaps suitably chosen (there is nothing “periodic” about types of expertise), there are interesting insights in the work, some of which have direct bearing for AI and for our discussion. One key insight is in the idea of “interactional expertise,” which has to do with the capability of talking in the *language* of a specialism in the absence of expertise in its *practice*. This kind of expertise, according to Collins and Evans, is quite ubiquitous, and can be found in many different situations and places — e.g., journalists or ethnographers studying a specialty; managers evaluating a specialist; peer-reviewers commenting on the merits of a manuscript or grant application; and so on. In successful scenarios of these situations, there is usually a progression from ‘interview’ to ‘discussion’ to ‘conversation’ between
the contributing expert and the interactional expert, throughout which the latter incrementally learns how to speak the language of the former.

This account of interactional expertise brings it one step closer to the possibility of AI and expert systems with some kind of access to tacit knowledge, compared to alternative views based on a more strict understanding of embodiment (e.g., Dreyfus). However, it still leaves a huge gap for systems such as Cyc that have been built on the premise of rule-based encoding of explicit knowledge. Cyc is one of the longest running AI projects, which started in the early 1980’s with the idea of encoding all encyclopedic knowledge, then shifted to what is called commonsense knowledge, and more recently it seeks to use semantic web techniques to pool the web as a source of knowledge. While this last move makes sense, given what we know about the tenets and principles of Cyc (Ekbia 2008), it is unlikely that the project can come even close to its alleged goal of becoming the indispensible knowledge platform of computing or the semantic web. One of these tenets has to do with what can be called an autonomist epistemology, which considers knowledge to be the possession of an individual agent or system.

By way of contrast, Watson developed by IBM works on a hypothesis-driven model, and given what we know about its architecture it seems to come closer to the mark in terms of relationality, although it is too early to make any judgments in terms of its direction (). In particular, the disembodied character of the system and its reliance on explicit representation poses serious challenges — e.g., in breaking the meaning barrier (Mitchell 2013).

**Looking Ahead:**
To summarize what we have discussed so far, Autonomist AI — the prevailing perspective based on a substantivist view of intelligence, knowledge, affect, morality, etc. — faces serious challenges in its goal of creating autonomous systems. In particular, it deals with an inherent paradox that manifests itself differently in different systems: developing affective agents regulated by drives, moral agents without shame, expert systems with no tacit knowledge, and so forth. We traced the historical origins of this paradox in the modernist project that seeks to create autonomous human beings endowed with freedom but constrained by their inherent neediness and their social obligations. AI research doubly augments this paradox through its techno-cultural imaginary. In its attempt to carry out “[Modernity’s] demiurgic ambition to exorcise the natural substance of a thing to substitute a synthetic one” (Baudrillard 1983), AI puts the modernist paradox in high relief.

Paradoxes, of course, do not have solutions. A productive strategy for dealing with paradox, as such, is “reversal” — that is, a strategy that flips the center and the margin, and thereby expands our understanding (Derrida 1982). In the case at hand, Autonomist AI, based on a substantivist view, puts individual properties and agendas of agents at the center, pushing “supplements” and dyadic relationships with others and the environment to the margins. A strategy of reversal would, therefore, put these at the center, and de-emphasize individual attributes.
I propose such a reversal with the aim of pursuing what can be called Artificial Relational Intelligence (ARI). ARI is a way of thinking about AI so as to make it more realistic and more humble, but no less interesting and challenging in its aims than current alternatives. This is a topic that I hope to pursue in later writing. To reiterate the power of participation in human behavior, though, I would like to end with a comment on the opening story of this article (Bearman 2112):

Everyone in that room broke the law, but they were celebrating it like a civic duty. That's how Chris allowed them to see it. Because if Chris had a genius for fantasy, it was that he understood that everyone had their own particular fantasy, and he could spot it and harness it, and weave it together with the rest of the people in his web…. The Moms wanted to be on TV. Norm wanted to feel powerful again. The media wanted a good story. The Candyman got a little fantasy date. Even Carl told me that before he first blew the whistle on Chris, he hesitated. Not just because he was scared, but because he, too was taken by Chris's grand vision.
References


Journal of American Society for Information Science and Technology, 60 (12), 2554-2566.


Notes

1 Butler’s story is also narrated in detail on National Public Radio: http://www.thisamericanlife.org/radio-archives/episode/447/transcript, and it has its own dedicated Wikipedia page: http://en.wikipedia.org/wiki/Chris_Butler_private_investigator

2 The close parallel between Rousseau’s view of the human condition and Turing’s idea of a childlike AI system is interesting.

3 Immanuel Kant, in his moral philosophy, has a different interpretation of “heteronomous”: acting in accordance with one’s desires rather than reason or moral duty.

4 I need to add that Collins and Evans (2010) present their account of expertise as a “substantive” theory, and in opposition to relational views. In my view, this is why their account runs into problems and dilemmas that they acknowledge (p. 76). More relevant to our purposes, however, what they consider “relational” is pure attribution, which is not the stance that I advocate here.
Figure 1. The robot Cog with and without supplements

Figure 2. (a) and (b) The scene of a chess championship match; (c) and (d) its portrayal in AI media
Figure 3. (a) Card et al. (1995)'s model of cognition; (b) Norman's model of human affects (From Boehner et al. 2005)