Why is AI so scary?

Michael L. Anderson
Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20742

October 16, 2005

Abstract

This article is a review of the following three books, and explores the question of why artificial intelligence is viewed with fear by many people. Thomas M. Georges *Digital Soul: Intelligent Machines and Human Values* (Boulder, CO: Westview Press), 2004. Stefan Helmreich *Silicon Second Nature: Culturing Artificial Life in a Digital World* (Berkeley, CA: University of California Press), 2000. Anne Foerst *God in the Machine: What Robots Teach Us About Humanity and God* (New York: Dutton), 2004.

By far the most common response I get when I tell people about my job is: "AI? Wow. Scary." I have to fight the impulse to roll my eyes, and don't always succeed. Generally I opt to soothe their fears with the appraisal that true AI is a long way off, but if I am feeling a bit pedantic I might also point out that there are very many currently existing technologies—nuclear power and various weapon systems come to mind, not to mention an overburdened energy grid that leaves us rather vulnerable—that are far more scary than anything AI has to offer. Strangely, this doesn't make them feel any better, and they are typically dubious, besides. But why is this so? If it is true, as I think it is, that other technologies are much more dangerous than AI, then why does AI nevertheless continue to seem uniquely threatening?

It is doubtless the case that the many "AI takes over the world" stories have something to do with it, but I'd like instead to consider two possibilities less often explored. Remember that there are actually three things in play here: the technology, its designers, and the public. If there is nothing uniquely dangerous about the technology itself, then maybe there is something about AI scientists, or about the public, that can make it seem, or perhaps actually render it so. In *Digital Soul: Intelligent Machines and Human Values*, Thomas Georges explores some of these possibilities.

Digital Soul is an extended, engagingly written plea for us to consider the moral and social implications of AI before it's too late. What does it mean to be "intelligent"? How different will machine intelligence be from human intelligence? Will these differences matter? What is the nature of freedom? Can physical beings ever be truly free? Which things deserve to be treated with moral consideration, and what kind of consideration do they deserve? Will we be morally permitted to make intelligent, autonomous machines work for us? How will this change our economy, or social fabric? How will we value hard work—or what will we value instead—in an android-driven future?

That the project of AI will eventually succeed Georges takes for granted, and *Digital Soul*, like most books of its kind, makes the immense leap from current technology to the possibility of super-intelligent machines, without much idea of how that could be done, nor really what kind of intelligence will be the likely result. This is a weakness even in the context of the moral and social discussions that provide the core content of the book. For instance, Georges assumes that the human loss of control over intelligent machines is inevitable, because such machines will not be useful unless they are autonomous, and this means that we won't control them. Moreover, he argues, even if we *could* control such machines, it would be immoral to do so, in the same way that it is immoral for us to control one another. It is not clear to me that either of these points follows quite so easily from the premise. For perhaps certain kinds or degrees of autonomy are compatible with control, utility, *and* morality, and without knowing the exact nature of the machines we are discussing, it is not possible to make this determination. There are many places like this where closer attention to actual technical details could have enhanced and guided the moral imagination. Still, whether the particular *conclusions* Georges reaches are sustainable is ultimately less important than the various issues themselves, and he correctly identifies a host of things that we will indeed eventually have to grapple with as we move ever closer to true machine intelligence.

One of the more unique and interesting things about *Digital Soul* is that, despite its detailing of some of the radical consequences that could result from the creation of intelligent machines, up to and including a complete revaluation of our own self-image, its tone is less often warning and dire than it is prophetic and hopeful. Indeed, Georges positively embraces the changes that AI will force in our culture and moral code, which he sees as largely dysfunctional, outdated, mystical and stupid. Georges turns out to be a kind of hyper-rationalist—he's a physicist by training—and moral relativist, just the sort of person one would expect to think both that AI is possible, and that its existence would help improve (or sweep aside) our irrational, mystical moral practices. He hopes we will come to see ourselves as purely physical beings, without immaterial souls or transcendental wills, whose intelligence is in fact a matter of complex, iterated interactions with our environment and one another. Having seen that the notion of pure individual autonomy is illusory, he hopes we will come to accept our deep interdependence, and will start shaping our environment and social structures so as to enhance peaceful cooperation rather than individual freedom.

This returns us to our theme, for however positive and appealing such a vision of the future will be to some, there are very many people quite attached to the set of inherited beliefs and practices—cultural, religious, and moral—that Georges dismisses as mystical and irrational, and hopes that AI will be instrumental in replacing. Insofar as he is right in predicting this effect of AI, then, AI represents a threat not as a technology, but as a social movement in which a rational, scientific world-view prevails over older cultural and religious beliefs. Note the difference between this and the notion that AI scientists will design machines that will turn against us and wreak physical havoc. The machines of Georges' imagination are already, by their very nature, against us (the majority of us, anyway), in that they represent a challenge to some of our oldest and most deeply-held beliefs. Such machines *may* destroy us, not by killing or enslaving anyone, but by being the vehicles through which the world-view of their builders triumphs, utterly changing our notions of who and what we are. If this is right then humans, as we now think of them, really will cease to exist. Maybe AI *is* scary!

George's vision will certainly seem dire and threatening to many, but perhaps it is just the idiosyncratic view of one scientist, and not even one directly involved in building intelligent machines. Is it really the case that AI scientists think differently from most everyone else, that they have such a different world-view that, were this view to become culturally predominant, the changes would be concrete, vast, and drastic? The answer coming from *Silicon Second Nature: Culturing Artificial Life in a Digital World*, by Stefan Helmreich, is largely yes (but in some ways, no).

Helmreich is an anthropologist who did his fieldwork studying the science and scientists of Artificial Life (ALife), primarily at the Santa Fe Institute in New Mexico. To use the oversimplified, somewhat counterproductive, but nevertheless resonant imagery of recent U.S. presidential campaigns, he did indeed find that ALife scientists, wherever they actually live, are blue-staters—more liberal and less religious than their red-state counterparts. In fact, every scientist he interviewed was an atheist without belief in souls, the afterlife, or anything resembling transcendental freedom. Instead, they were mostly reductive physicalists, who shared a belief that it will be possible to create life in various forms and media, once they find the right design. Moreover, he argues, with the support of many telling examples, that the machines scientists build will in fact embody the (nearly complete) set of beliefs that they bring to the drafting table. Not just the technical ideas about how computer vision or automated reasoning works, but also the more theoretical, culturally-informed ideas what vision, or thinking, or autonomy (or gender, or race, or sex) *is*, what it is for, what its value is, and what it means. In fact, he would deny this distinction between the "technical" and the "theoretical" or "cultural", for ideas always contain all of these, and it is not generally possible to distill each from the others. Intelligent machines, then, will be like the AI scientists who build them, which is perhaps good news for secular rationalists, but worse news for feminists, as Helmreich found little evidence that the members of this highly male-dominated field have anything but fairly traditional ideas about gender, gender roles, and the social structures they typically support. And, of course, it can be seen as terrible news for the religiously inclined.

However, Helmreich's book focuses less on the *future* of ALife, on what intelligent artificial life *will* be like, and more on what the creations of ALife *are* like, and how this reflects the particular social, cultural, physical, and economic circumstances of their designers. For this reason, *Silicon Second Nature* is an extremely interesting book, and well worth a look from AI scientists. Much of it will be hard going for those with no background in, or little tolerance for, the peculiar vocabulary of cultural studies, with its tendency to turn nouns into gerunds and make plurals from singulars (culturing, gendering, knowledges, subjectivities), and its fascination with the flock of "-isms" (heterosexism, masculinism) that define the "unconscious hegemony" of western technoscientific thought. Even so, Helmreich's analyses are nuanced and highly sympathetic to scientific practice, and this book does not read as yet another relativist critique of scientific autonomy and objectivity.

It is much more interesting than that, and, indeed, could well provide to AI the kind of externally-mediated self-understanding that leads to greater insight and creativity. For when unconscious, unrecognized assumptions and limitations are held up for scrutiny, it becomes possible to discard them, entering unexplored territory.

This being said, there is nevertheless no doubt that the book will ruffle feathers. Anyone who does not see such things as gender roles and the distribution of wealth (including money, time and autonomy) as highly contingent arrangements maintained only with great (if often unrecognized) effort is likely to be a bit unsettled by Helmreich's observations. And those who suppose that the ways of thinking both supported by, and supporting, such contingent constructs do not (cannot) find their way into the practice of science will likewise be upset by this report. But while it is true (in my judgment) that some of his analyses are stretched past the breaking point—especially those that claim to find a tinge of racial or even racist content in the ALife understanding of the primitive—the accumulation of valid and surprising examples of scientific claims clearly influenced less by empirical considerations and more by ideas of "naturalness" or "normality"—notions that are highly culturally infused—strongly argues for careful consideration of his overall thesis. Simple examples include pair-wise cross-breeding in genetic algorithms (why not different or more complex forms genetic mixing?), and the treatment of the evolution of "immortal" creatures by one simulation as a bug to be fixed, rather than as an interesting result (a more religious person might have had quite a different reaction).

To put the whole argument a different way, think what might happen if the boys of AI (and it is mostly boys) were asked to design the first girl robot. It seems entirely likely that the result would reflect not just the technical expertise and scientific knowledge of its (her) designers, but also their various ideas—many of them prereflective cultural inheritances—about what constitutes the feminine. Well, you can imagine what might result, and we should hardly be surprised to be confronted with some cross between Commander Data and Lara Croft. The lingering question left by Helmreich's observations is how many of the core concepts of AI—mind, representation, autonomy, emotion, self, desire, reward, reproduction, etc.—are more like the typical computer jock's ideas about girls than we'd prefer to admit.

Besides the discovery that our machines reflect who we are and what we believe, which seems to support Georges' prediction that the existence of such machines will in part represent a triumph of a secular scientific world-view, there is one other recurring theme from the book that is worth remarking: many of Helmreich's subjects report moments, often transformative ones, in which they found themselves treating their creations *as* alive, as real entities in their own right. These experiences range from simply *seeing* the screen patterns of John Conway's "Game of Life" as entities with beliefs and intentions, to the near sub-conscious feeling, reported by one scientist, that while his simulation was running it was as if there were another presence sharing his office. What's interesting about these reports is that these scientists did not *decide* to treat their creations as alive; rather the behavior of the creations somehow *caused* the scientists to think in this way. Georges notes something similar when he remarks that, as people imagine themselves interacting with intelligent machines, that they cannot help *but* think of them as living things, with many of the ethical obligations this implies.

This element of our relationship, or potential relationship, with intelligent machines is also highlighted in Anne Foerst's recent book, *God in the Machine: What Robots Teach Us About Humanity and God*. We are, it seems, programmed (if you will excuse the word) to respond with a certain respect and deference to interactive things, and the more interactive, the more deeply felt is the claim to respect that these interlocutors make on us. Foerst reports on some experiments done by Stanford sociologist Clifford Nass, wherein he asked subjects to test some interactive tutorial software. The software was designed to perform poorly, and after the testing was complete, the same computer asked the subjects to evaluate its performance. Surprisingly, given the objectively poor performance of the software, subjects' responses were generally positive. Immediately after, the subjects were moved to a different room, where, working at a different computer, they were again asked to evaluate the performance of the tutorial software. Here the responses were somewhat less favorable. Finally, the subjects were interviewed by a human, and only in this situation were they very negative (and apparently more truthful) about the performance of the software. These results suggest that the subjects felt the need to be polite and positive when evaluating the performance of the first computer while *using* that very computer, just as they would when asked to evaluate the performance of a person to his or her face—and this effect was strong enough that it carried over to a "relative" of the first computer, just as politeness might carry over to an evaluee's family or friends. Although, when questioned, the subjects denied they had any obligation to be polite to computers, their behavior tells quite a different story.

I think, somewhat paradoxically, that this is another source of the fear of, or at least the resistance to, intelligent machines: not that they will be better than us, or turn on us, but just that they will interact with us in such a way that we will feel connected to them, and this connection will obligate us in specific ways that, thinking about it now, make us uncomfortable. To get some idea of what difference this sense of connection can make, imagine having a personal assistant that can be commanded at will; and now consider how much more complicated and difficult that relationship would be if you felt the assistant deserved to be treated with a bit of respect. Just as some bosses are loath to let their human assistants cross this line into personhood, so are very many more people wary of allowing their computers to do so. And yet this may be the attitude to which we are naturally disposed; it is hard *not* to feel this obligation, and therefore hard *not* to make the interaction more complex and personal. This sense of connectedness has other effects as well, and may even enhance the basic fear that the machines will take over. For, imagine being threatened by a stranger; and now imagine feeling the same threat from someone you had taken to be a friend. The objective danger might be exactly equal, yet being threatened by the friend—someone trusted and close—would be vastly more painful and bewildering. Thus, even though we may acknowledge the objectively greater threats posed by other technologies like nuclear power, the threat coming from AI may *feel* somehow worse. Nevertheless, it does seem that the primary ground for the fear of AI—that is, the thing that makes AI uniquely threatening—is a fear of the growing dominance of the world-view embedded in the project of AI itself. It is this issue that is the main subject of Foerst's book.

God in the Machine is written with the breathless style of a technical correspondent for *E!* magazine, and words like "fascinate", "excite" and their energetic relatives fill the pages, as the reader is taken to meet fascinating people enthralled by AI, and its fabulously interesting and exciting machines. Such is the fate of serious subjects in mass-marketed books like this one. And this *is* a serious book, or at least an earnest book on a serious topic: the clash of world-views between secular-scientific AI (fairly represented by Minsky's comment that human beings are "meat machines") versus the foundations of religious belief and practice, and the reasons to think that this apparently deep divide can be narrowed, and perhaps bridged or even erased.

The main strategy of the book is to point out the various shared concerns between religion and AI—a concern for the nature of human beings, and most especially, given Foerst's focus on robotics (she was a post-doc in Rodney

Brooks' lab at MIT), for the significance of, and the challenges resulting from, our embodiment—and to describe these concerns in language general, abstract or metaphorical enough that it can indeed begin to seem that AI and Theology are treating the same subject matter, albeit equipped with slightly different disciplinary tools (and funding sources). Thus, for instance, she treats the Jewish golem stories—about the creation of a human figure from clay, which is then animated by inscribing it with the name of God—as describing a kind of theologically grounded proto-AI. This interpretation allows her to make the reverse claim that modern AI is in fact a quasi-religious practice, a reverent imitation of God's creative abilities. This is an interesting idea, but of necessity glosses over some of the differences in detail between the aims and motivations of sixteenth century Jewish rabbis and contemporary scientists. As a result of this overall, highly conciliatory approach, Foerst is led to an account of religion in which God is understood as something broad and general like the supreme animating force of the world, so that different religious beliefs can be seen seen as ultimately compatible, culturally relative expressions of the same basic idea. Similarly, she softens scientific truth claims by reminding us of the various ways in which these claims are culturally encoded and relative to theoretical schemas. It does seem true that, if religion involves little more than recognizing a supreme animating force in the universe, and science makes no claim to establishing the truth about the nature and origins of that force (or of the universe as a whole), then there is a good possibility that these world-views are ultimately compatible. However, in my judgment such interpretations artificially lower the bar, making reconciliation between AI and religion—human practices that in fact sometimes stake incompatible claims to the truth—seem a good deal easier than it in fact is.

In the course of her exploration (and re-interpretation) of the competing world-views of religion and AI, Foerst ranges over many, many different subjects, each treated only briefly and sketchily, with the result that anyone who knows anything about any of the fields discussed—AI and its history, Cartesian philosophy, machine learning, the nature and limits of human knowledge, the basis of religious belief, narrative accounts of human personhood, embodied cognition, etc.—is likely to come away frustrated. Worse, indeed so much so that it approaches the level of an internal contradiction, is that, unlike Helmreich, she is not at all deft, probing, or apparently self-conscious about her own allegiance to the subjectivist cultural relativism within which the reconciliation of religion and AI is seen as possible. Although she frequently reminds us of the “obvious” fact that the key terms of AI and theology—“intelligence”, “life”, “intuition”, or even “God”—are cultural constructions with sometimes shared but no objective meaning, she nevertheless is quite happy to tell us what we *know* about the brain, or about human behavior, from neuroscience, psychology and cognitive science. Thus, she relies on the authority of science to motivate her accounts of human nature, even as she is lead to undermine that authority to achieve her objective of cultural reconciliation.

These are significant flaws (and there are others, including basic scientific mistakes like confusing aphasia, a difficulty in language use, with bodily neglect, a disruption of the body image), but perhaps the main weakness of the book—and the main reason its proposals for reconciliation are likely to fail—is Foerst’s overwhelmingly liberal and open approach to theology and to the interpretation of religious meaning. She writes, for instance, that “[b]uilding Cog and Kismet”, two robots at MIT modeled on human beings, “made us modest in our admiration for God’s creation.” Although she admits that “[p]robably no one except [her] in the team would formulate this feeling with the same religious words,” she is nevertheless confident that “the sentiment was exactly [the same].” I am not so sure. Indeed, it seems to me that such religious words only express the same sentiment as that felt by a typical MIT scientist if one’s understanding of God is so highly abstract that treating something as a divine creation means nothing more than recognizing it as a highly complex and impressive bit of machinery. A more demanding God—of the sort we routinely encounter in Hebrew and Christian scripture—might expect a great deal more from us in the face of His creation. Modesty could indeed require respect for the *mystery* of creation, entirely ruling out the scientific study, or at least the unnatural reproduction, of His works (an injunction that could affect not just AI, but stem cell research, cloning, and many other things besides). Such possibilities do not enter into Foerst’s liberal theological picture. For her, exploring and understanding our bodies is of necessity an act that brings us closer to God, who after all became flesh in the person of Jesus. Well, that’s one possible interpretation, to be sure, but it seems destined to remain a minority view.

The case is likewise with her take on evolution, the acceptance of which is central to the scientific enterprise, and increasingly a part of research in artificial intelligence. There would seem to be a major disagreement on this point between the religious and the scientific establishments, but Foerst is dismissive, writing: “only a few very extremist Christians actually believe that evolution and creation are in conflict." Would that it were so. In fact, according to a recent survey by the Pew Forum (August 30, 2005), fully 42% of the U.S.—more than 125 million people—believe that evolution and creationism are in conflict, while only 18% accept their compatibility. The numbers may be more to Foerst's liking in Europe, but probably not in Africa and South America, in both of which places Evangelical Christianity is growing at a rapid clip. We are not, in any event, dealing with the beliefs of "only a few".

I bring this up not because it is clear to me that liberal theology offers an incorrect understanding of God, or of the obligations of religious believers, but because it should be clear to everyone that the understanding it offers is not universally shared by the faithful. Likewise, although many scientists are cautious in staking their claims to know the truth (for we must be constantly aware that any of our theories could turn out to be mistaken), it is nevertheless among the foundational principles of science that the truth is humanly accessible, and the way to discover it is through the patient, repeated, and careful application of scientific methods. In the end, Foerst has created two straw figures, not to knock them down, but to marry them, and while these scarecrows may be compatible, it is far from clear that their real-world models are.

This is unfortunate, because there is indeed an important and on-going clash of world-views, creating fear, anger and frustration on all sides—and AI scientists are in the middle of it, whether we want to be or not. The longer we go without a searching and intelligent exploration of these differences—an exploration that takes both sides seriously, neither diluting nor dismissing their self-understandings—the more dire, and unpredictable, will be the eventual outcome. Where there is common ground, we should find it and build on it, and where there is genuine difference, we should see to what degree each can advance its own causes without damage to the other. And where, finally, there are differences irreconcilable and mutually antagonistic, there we will have some hard choices to make; but we should make them in the fullest possible awareness of their nature, meaning and likely consequences. The muddled skirmishes currently being fought in public policy forums from school boards to drug advisory panels offer no such illumination.