Just an Artifact: Why Machines Are Perceived as Moral Agents*

Joanna J. Bryson
University of Bath
Bath BA2 7AY, UK
j.j.bryson@bath.ac.uk

Philip P. Kime
Zürich, CH
Philip@kime.org.uk

Abstract

How obliged can we be to AI, and how much danger does it pose us? A surprising proportion of our society holds exaggerated fears or hopes for AI, such as the fear of robot world conquest, or the hope that AI will indefinitely perpetuate our culture. These misapprehensions are symptomatic of a larger problem—a confusion about the nature and origins of ethics and its role in society. While AI technologies do pose promises and threats, these are not qualitatively different from those posed by other artifacts of our culture which are largely ignored: from factories to advertising, weapons to political systems. Ethical systems are based on notions of identity, and the exaggerated hopes and fears of AI derive from our cultures having not yet accommodated the fact that language and reasoning are no longer uniquely human. The experience of AI may improve our ethical intuitions and self-understanding, potentially helping our societies make better-informed decisions on serious ethical dilemmas.

1 Introduction

> "Within thirty years, we will have the technological means to create superhuman intelligence. Shortly after, the human era will be ended. Can the Singularity be avoided? If not to be avoided, can events be guided so that we may survive? What does survival even mean in a Post-Human Era?"—Vernor Vinge, *The Coming Technological Singularity: How to Survive in the Post-Human Era* (1995).

> "Technologists are providing almost religious visions, and their ideas are resonating in some ways with the same idea of the Rapture." — Eric Horvitz, *Scientists Worry Machines May Outsmart Man* [Markoff, 2009].

Not all computer scientists consider world conquest by machines probable, or even possible. However, such fears have been a persistent part of our culture, not only in fiction but also in scientific writings [de Garis, 1990]. What can lead even computer scientists to believe that AI endangers our society? Computer programs, including those classified as Artificial Intelligence (AI), are purpose-built artifacts designed, commissioned and operated by human beings. Computers can accelerate and magnify our mistakes to do more damage than an unaided individual, yet the same could be said of levers, pulleys and organised government.

We believe exaggerated fears of, and hopes for, AI are symptomatic of a larger problem—a general confusion about the nature of humanity and the role of ethics in society. To the category of *exaggerated fear* we assign the notions of ambitious or machine-loyal AIs that make selfish decisions about the relative importance of their own intelligence, growth or energy. The category of *exaggerated hopes* includes the expectation that machine intelligence will perpetuate either individual or planetary experience and culture past the normal life expectancy of a human individual or the human species. Our thesis is that these are false concerns, which can distract us from the real dangers of AI technologies. The real dangers of AI are no different from those of other artifacts in our culture: from factories to advertising, weapons to political systems. The danger of these systems is the potential for misuse, either through carelessness or malevolence, by the people who control them.

Social ethics is derived from each individuals' personal sense of obligation. The proximate (but not ultimate [West *et al.*, 2011]) mechanism of that obligation is an individual's identification with its object. This explains the misplaced hopes and fears for AI if they come from individuals' inappropriate identification with machine intelligence. Yet AI, properly understood, might be used to help us rationalise our ethical systems, leaving us time to address the real threats to our societies and cultures, including those stemming from the misuse of advanced technology such as AI.

2 Ethical Obligation

Understanding ethical issues requires the understanding of ethics itself, particularly its function. Naturally, the diversity of this subject prohibits a summary here of general tenants, but for our purposes we concentrate on the idea that a significant element of ethics, whether by its essential nature or more weakly, by its manifest effects, is to maintain

*An earlier version of this work was partially published in *Proceedings of the 15th International Congress on Cybernetics*.

a functional degree of social homogeneity. This is a classic theme in functionalist sociology [Parsons, 1991]. This is sometimes expressed as the function of protecting the social organism [Hobbes, 1651; Hölldobler and Wilson, 2008]. In other words, ethics has evolved into a contributor to human social cohesion. As with any evolved system, some details of a particular ethical system may consequently appear arbitrary. In any evolved order, there is a purely contingent historical element which determines the variation subsequently subjected to evolutionary pressure. Further, the tendency for "speciesisation" (in this case cultural and/or family identity) may lead to idiosyncrasies [Cronin, 1991]. However, some concepts are nearly universal, such as prohibition of necessarily socially destructive behaviours, such as murder, or support for locally unifying forces, such as family and religion.

One important aspect of any ethical system is that it should be as much as possible self-regulatory [Dennett, 2006]. Socially speaking, external control is far too resource-intensive. For example, if we were to be prevented from crime solely by the threat of external punishment rather than by internalised codes of ethics, the police force would have to number a good percentage of the population, and would itself be difficult to regulate. In fact, however, many forms of individual ethical or altruistic behaviour are generated without conscious weighing of negative consequence to the individual. Most people "naturally" conform to such social expectations as protecting and supporting their own family members, or taking care not to damage the property of others.

A significant basis for such ethical reactions is empathy or identification with another entity. We care for people or objects that we would *feel badly for* if they were hurt or damaged. Thus ethics requires some degree of perceived analogy between ourselves and that with which we empathise. Our empathy and sense of ethical obligation tend to be highly correlated, with our future and past selves. Our families tending to be at the top, followed by our neighbours and other people with whom we acknowledge commonality. A proximate explanation for this gradient is the ease of identification in such cases. Children have well-documented life-long identity confusion with their parents, possibly founded in a failure of categorisation occurring between the discrimination of self and other in infancy [Damon, 1983]. Parents, perhaps confusingly, often have the same goals as the child—they support the child's physical and psychological needs. This complicates discrimination of self on the basis of reaction to intention. In turn, parents can view their children as perpetuations of their own bodies and even lives. Ultimately, such identity confusion with close kin is obviously adaptive.

If identification can be seen to diffuse through a community, we can claim a likely strong correlation at least between identification and internalised sense of ethical obligation. If we come to understand our more complicated relationships with friends, state and religion through transference or metaphor with familial relationships, then we might quite literally feel that actions that benefit these select others are in our own best interest. If this sort of generalised 'self interest' is the root of our ethics, then our innate self-regulation may be relied on to govern our ethical behaviour.

3 Over-Identification with AI

We propose that misidentification with machine intelligence leads to false ethical evaluations of AI's potentials and threats. What could lead us to over-identify with machines? Quite simply, a misconception of human life that puts the capabilities of language, mathematics and 'reason' as its key characteristics. Historically, the things in our world that are most like humans are other animals, thus ironically we define our identity in terms of the ways we differ from them. Beliefs regarding the 'essential' human characteristics may be a consequence of a general drive for identity which creates a desire to separate ourselves from entities with which we are similar enough to threaten our own sense of uniqueness and utility.

Forming a human society again has historically required valuing the lives of the humans in the community over the lives of other animals [Aiello, 1997]. However, the most obvious metrics for identity with respect to animals lead to an undervaluing of the emotional, visceral and aesthetic in our society. Consequences include a neglect and denial of emotional experiences—a theme familiar not only from popular psychology but more recently from behavioural economics [Henrich *et al.*, 2001]. Further, we have difficulty understanding the behaviour of our companions and ourselves as we attempt to impose intentional models on what may well be instinctive, heuristic or emotion-driven responses. Such failures of understanding can lead to poor predictions and unnecessary conflicts. The intractability of formal AI results from trying to model everything with these easily-understood fully-justified systems which sadly never fit the real world [Dreyfus, 1992; Vernon, 2010].

If identification is central to our sense of ethical obligation, then over-identifying with a machine displaying some aspect of artificial intelligence holds two dangers. Firstly, we may believe the machine to be a participant in our society, which would confuse our understanding of machines and their potential dangers and capabilities. Second, we may over-value the machine when making our own ethical judgements and balancing our own obligations [Bryson, 2010].

The statement that we over-identify with machine intelligence is, of course, itself a judgement. This evaluation can be made on two different levels, one technical and the other ethical. The technical is easier to demonstrate: the general population ascribes much higher levels of intelligent capabilities to machines than machines generally possess. The ethical evaluation is more subjective. It is quite possible that some percentage of the people reading this paper would consider the aspects of culture potentially embodied in computer programs as equally or more valuable than some or all individual human beings. We do not attempt to address this issue directly here, but rather seek only to point out that this problem is not restricted to AI, but rather is a problem for many forms of artifact, including fine art and political systems. The technical argument is sufficient for this section. This section focuses on fears and hopes specific to AI which we attempt to demonstrate as unfounded.

4 Identification and Obligation

There are two possible consequences of over-identifying with a machine. The first is that it lowers one's own opinion of oneself. The fear of functionalism is an example of this [Dennett, 2003]. The other possible consequence is the inappropriate elevation of the worth of the machine. This can be the basis for both exaggerated fears of and affinity for AI. For example, in identifying with our machines, we endow them with the rights and privileges of ethical status. We are somewhat embarrassed by this sense of obligation, and so try to rationalise it, while at the same time distancing ourselves. Many people think it would be unethical to unplug a computer if it were conscious. Naturally, this hypothetical gets us nowhere since *consciousness* is an aggregated concept over multiple phenomena with no single agreed definition [Dennett, 2001]. However, we can see where obligation, consciousness and identification overlap without having to have any idea of what consciousness "is".

Take the idea of memory — a computing concept that most of the public have at least some sense of. Memory is inescapably suffused with many connotations having nothing to do with computers. Consciousness and memory are related in obvious ways to do with continuity of awareness, responsibility stemming from remembered actions etc. Computers have memory in some sense not utterly unrelated to the human sense of the word, even though for computers memory is something you can purchase and add. Computers remember things, otherwise they would contain no programs and perform no tasks. Their memory allows detection and recovery from errors and inconsistencies, storage of experience and either exact or generalised recall of what is stored at another time—whether on the same machine or another compatible one [Date, 2004]. Thus if explicit memory is relevant to consciousness [Dennett and Akins, 2008] then it is not utterly absurd to speak of machine consciousness.Further, the attributes of both humans and machines which make it possible to so speak are actually a decided benefit. If such benefits are related to a sense of obligation towards something—for example an obligation towards something which has a continuity of memory, then obligation is due to non-human entities. This is already apparent in the obligations societies feel towards libraries—after all, they 'remember' information for us, they have a continuity of this 'memory' and it is partly this which accounts for our sense of obligation. Of course, neither this definition of consciousness nor anything else so simple-minded could be considered entirely adequate. However, this example does show that certain of the concepts which we hold dear to in ourselves as against machines, if construed in pragmatically sensible ways, can exhibit characteristics which might seem to require a sense of obligation.

Failing to recognise fundamental differences between human and machine intelligence still leads to mistakes. For example, because we consider human intelligence to be the only intelligence, and because we infamously desire power, we tend to assume that any other intelligent system desires power. Deep Blue can defeat most humans at chess, but it has absolutely no representation of power or human society anywhere in its program, with the possible exception of the metric values of various chess pieces and positions [Hsu, 2002]. While a human chess player might be inspired and motivated to excel by the feeling of power, this neither necessarily nor in fact holds of computer chess programs.

Of course, it is easy to program a computer to print, or even say "I want to rule the world!!!" It is also possible to program a system to preferentially select behaviours that give it more power or resources. Ray's [1995] artificial-life system Tierra, at its essence, just develops programs that compete for computer resources such as disk and processing time. Internet worms often compete for processor time, some disabling computers world-wide by monopolising their processors. However, these actions can happen without intent either by the programmer or the program [Eichin and Rochlis, 1989]. Worms no more *intend* to out-compete other processes than a bomb intends to destroy a city.

The 1988 Internet worm demonstrates another aspect of identifying with artificial intelligence. Its creator desired to have a copy of his program running on as many different machines as possible. This ambition for a creation is again similar to the ambition parents may impose on their children. We would like to be more intelligent, to live longer, to be stronger; if not actually ourselves, then our progeny with whom we identify should have these characteristics. Identifying with biological children is evolutionarily adaptive, and it is possible that identifying with intellectual offspring is culturally so. A consequence of this is both scientists and science fiction writers sometimes speaking of self-replicating space ships becoming the ultimate receptacle of Earthly intelligence — as if surviving to the end of the universe were significantly better than surviving to the end of the solar system [Sagan and Newman, 1983; Helmreich, 1997]. Science fiction has also thoroughly examined the idea of robots that are deserving of citizen status [Bryson, 2010]. It is important to understand that these works of literature are exploring what it means to be human, not what it means to be a computer. Largely, they are an attempt to legitimise the particular mixture of psychological elements which humans, presumably purely contingently, have evolved to demonstrate. The problem is that such attempts are essentially and utterly human-centric — they attempt to necessarily link desire for power, love, community etc. with the development of intelligence, while in fact the human mixture of all of these is at least partly determined by contingencies of evolution.

Machines in contrast are created for a purpose. We develop artifacts to perform tasks for us, and while they may eliminate the need for various human labours, they do not eliminate the need or desire for us to live our lives. The threat is not that machines out of maliciousness will take over the world, but that every human endeavour will eventually be 'better' accomplished by a machine. What makes this concern mistaken is that what ultimately matters to us is not the actual accomplishments of our lives (for which there is no real, objective metric to measure value) but the performing of the actions that leads to accomplishments. What we value is what we actually sense; it always has an element of the aesthetic and emotional. What, for example, happens when people choose to write a letter by hand instead of sending

email, because writing "loses some of its essence" when it is too easy? Perhaps they are eccentric, but perhaps they have recognised something extremely valuable about their own experience, and that of their letter's recipient. It is ludicrous to think of a machine falling in love for you, of it enjoying victory or gossip on your behalf. An obsession with the results of action rather than the actions themselves is not the fault of AI, but a problem our culture needs to address. If AI puts this crucial issue into sharp relief, all the better. The experience that results from our actions is, we claim, what the highest value for an individual human.

5 AI and Ethics

As we stated earlier, our argument is not that there are no ethical considerations to creating and using AI technology. Rather, we are arguing that the *nature* of those ethical obligations is often misconstrued. We should neither fear the motives of a system nor trust its common sense. AI systems need not have either "motives" or "common sense", but where we choose to create systems that can meaningfully be described in such terms, we should take the same precautions against their potential flaws as we do against similar human errors and fallibilities, such as auditing. These kinds of precautions are already standard even for conventional computer programs in critical applications, such as manned space flights [Sklaroff, 1976].

AI has been commonly used in industry since the late twentieth century. For example, computer manufacturers use expert systems (AI programs designed around some notion of 'common sense') for checking circuit design. But this is just one step in the manufacturing process; real circuits are later built and tested. Similarly, credit card companies use machine learning to build profiles of their customer spending habits, in an attempt to recognise as soon as possible whether a card is being used by a thief. But no one is arrested directly as a consequence of these programs: exceptions are flagged and turned over to human account officers, who phone the customer and verify whether they were making the unexpected purchases. A card may be automatically disabled, but then the customer can telephone their provider and discuss the situation themselves. These are examples of AI systems reasonably and responsibly integrated into our culture. There is no reason for a disturbing qualitative jump in practise as usage and empirical experience rather than theoretical assurances of desirable attributes have integrated these systems.

What about our ethical obligations to intelligent systems? To answer this we refer back to the fact that ethical systems, while having sometimes arbitrary evolved features, are essentially involved in maintaining social order. Since we are provided with a new ethical quandary, we are to some extent free to create a new ethical standard. Of course, new standards must not be overly disruptive of the overall existing code of ethics, and should generally contribute towards a social order our society finds suitable. The best way to achieve these desiderata is if the change is not abrupt, and indeed as we have argued there does not need to be a harsh transition. Rather, ethical adjustment to encompass AI is and will be an ongoing process—one in which we are already participating.

Ethics generally leads us to be altruistic towards members of our society that we identify with, or even altruistically violent against those who threaten it [Nakamaru and Iwasa, 2006]. For identification to operate there must be some sense of 'likeness' that leads us to empathic understanding. If such a sense is genetic or biological, then we will always have absolutely no ethical obligation towards an artifact. But what if we choose our standard to be cultural? After all, much of what we consider to define our identity—our language, our music, our social networks—is less inborn than learned. If an artifact becomes a vessel for our culture, should we treat it with the same respect as a person? Again, this question is valid and difficult, but also already with us. Books, buildings, art, songs and languages all embody the intellectual output of members of our species. If it were unethical to destroy a building or to burn books, then it would be unethical to destroy a machine that retains and communicates the same kinds of information.

The difficulty with choosing retention of culture as a criteria for ethical consideration is that it raises the possibility that some machine *could* become considered more important than some human. Before dismissing this idea out of hand, consider that again, this problem is already with us. We routinely place human life as significantly less important than sustaining cultural artifacts such as political systems, economic systems, and religions. Whilst warfare may be considered unethical, people are far less inclined to deem it so when it is seen as being defensive, when an aggressor is threatening a way of life—a language group, a religion, an economy. But even a defensive war is still a sacrifice of life to cultural artifact. It is justified when the cultural artifact is agreed to so enrich the lives of the population as to merit the risk of loss of life to some percentage of them.

A less romantic but equally compelling example is automobiles [Williams, 1991]. Road accidents claim many lives every year all over the world. A complete solution to this problem would be to ban cars altogether. However, even ignoring the enormous economic and therefore political pressure to save automobile transport and industry, the fact is that these are privately owned artifacts, and everyone who owns them is making an independent decision about their net worth. The convenience or perhaps the power of owning this particular artifact is perceived as being so huge that the risk to human life is, in some sense, less important.

Besides receptacles of culture, another possible criterion of ethical status is that of *contributing* to culture. There is an entire discipline of computational creativity [Wiggins, 2006]. AI-generated art and music can pass the Turing test—that is, it can be mistaken for a human production [Cope, 2005]. Robots may even be used as nannies or personal trainers [Sharkey and Sharkey, 2010; Lowe, 2010]. We would like to suggest that it would be considerably less disruptive to our existing ethical system (and therefore society) if AI is considered a *tool* of creativity, not a creator in itself. The creators are those who design and/or operate the AI.

This sort of perspective would allow us to place a creative AI system on the same standard of any work of art or scholarship, for example a major astronomical observatory, the Mona Lisa or *On the Origin of Species* [Darwin, 1859]. The ques-

tion is, if an artifact is retaining or generating more intellectual information than a person, should people be allowed to die to preserve it? This difficult problem is again one that faces us with or without AI. Resources are spent to protect the Mona Lisa that could in theory be spent on medicine or food. But AI allows us a luxury not afforded to da Vinci, though afforded to Darwin. With books and AI we can ensure our contributions are not unique [Bryson, 2000]. Every aspect of our work is replicable, and can be "backed up". Even if a system learns from irreproducible experience, the resulting internal states could normally be preserved. So, in a sense, AI helps reduce the possibility of such ethical problems. Replicability is replaceability and thus deciding between the importance of things is easier. This property of computer-based systems allows us to be unabashedly biased in favour of people and other evolved systems, as they are unique and irreplaceable.

If AI machines could generate scientific theory, art, money, or some other cultural commodity significantly faster than humans, would we not be obligated to devote *all* of our resources to building such machines? Obviously not. Ethics has always involved balancing obligations to various sources: to yourself, your family, your city, your country. There can be no reason except nihilism that human civilisation should chose a set of ethics that values artificial intelligence to the exclusion of our own existence [Sagan and Newman, 1983].

In sum, we are not arguing there *is* no problem of assigning ethical status to AI. Rather, our point is that each of the problems the threat or promise of AI draws to our attention are actually ones we already have, and yet seldom consider. The challenge is to more clearly see the decisions we are making concerning the relative worth of human experience and our cultural artifacts, and to integrate intelligent artifacts into that balance accordingly. Treating AI as some new and dangerous source of ethical conundrums may in fact just play into the hands of those profiting from the current status quo. The illusion of a terrifyingly different world in which we are at the mercy of machines is simply a trick of perspective: we imagine the far future and compare it to our concept of the present without taking into account an intervening time of gradual change. In fact, we are currently and always in that "intervening" time, and investment in immensely powerful and sometimes destructive cultural artifacts is already a centuries-old industry. The process is underway. There is no Singularity.

6 Ethics and AI

AI could in fact be used to help rationalise human ethical decisions. Our technology gives us means to do extensive damage not previously possible, but it also allows us to understand our world, including ourselves, as never before. As we come to understand the evolution of our culture and society, we can also make decisions and take actions that have direct impact on that evolution. AI systems may actually play a key role in this process They can be used to more rapidly recognise and allow us to reinforce valuable emerging ethical norms [Anderson and Anderson, 2008].

As we have indicated earlier, one of the key problems for our society is the development of internalised ethical systems. In particular, several traditional means of communicating imposed senses of identity, such as religion and family, are severely challenged by our new densely populated, multi-cultural, informed and empowered citizenry. The issue of forming identity is now more than ever an issue for public education. It is important for students to undertake to understand themselves, including the understanding of their dependence on their society and environment. AI is in fact an ideal tool for working on this problem. Already, students learn through computer simulations about dynamic systems such as the global environment or local ecosystems [Hantsaridou *et al.*, 2005]. Equally important might be the introduction to the classroom of AI systems allowing the students to model aspects of their own behaviour.

If students were allowed to build AI 'portraits' of themselves and each other, this would offer both an opportunity to learn about their own behaviour (e.g. the necessity of having multiple, conflicting goals), and to have a more explicit model of how to resolve them. Direct experience of such simulations might also help reduce the confusion between mechanism and human, by demonstrating some of the more subtle and revealing ways in which computers fail the Turing Test.In the future, children should have no more difficulty disambiguating an artificial companion from a friend than they would a painted portrait.

Our conclusion then is that in creating and using intelligent artifacts we do need to consider ethical and social dangers, but in no greater sense than we should with more conventional technology. If AI is to control our defence systems, our utility supplies or our political propaganda, we should take exactly the same care with them we should currently be paying our computers, power utilities and media. If it accelerates our capacity for change, we must also use it to increase our ability to track and govern change. To the extent that AI comes to serve, like us, as vessels of our culture, we owe it the same respect we owe libraries, books, architecture and works of art.

We have argued that the extent to which people feel exaggerated fears of or obligations towards AI and robotics artifacts is a consequence of their over-identification with the artifacts due to superficial, unfamiliar similarities such as a machine's use of language or reason. It also indicates our society's uncertain understanding of ethics, which, having so far largely evolved unguided within cultures and societies, is currently hard pressed to keep up with the rate of cultural change. We propose that the empirical experience of AI might be one way to help us to understand intuitively not only that AI is not us, but also the nature of the various relations which *are* ethically significant.

We have emphasised the importance of identity born of such experience in forming ethical obligations.AI simulations and responsibly-advertised domestic robotics could help us form better models not only of what it is to be a machine, but also by comparison of what it is to be human, and thus provide us with a better basis for empathetic and ethical decisions. Empathy and therefore real every-day human ethics is dependent on real individual experience rather than theoretical arguments. Those worried about the acceptability of AI technology in our culture should worry less about anthropomorphic fallacies and more about making AI visible and understood where it already exists.

References

[Aiello, 1997] Leslie C. Aiello. Brains and guts in human evolution: The Expensive Tissue Hypothesis. *Brazilian Journal of Genetics*, 20(1), 1997.

[Anderson and Anderson, 2008] M. Anderson and S. Anderson. Developing a general, interactive approach to codifying ethical principles. In Ted Metzler, editor, *Proceedings of AAAI Workshop Human Implications of Human-Robot Interaction*, Chicago, July 2008.

[Bryson, 2000] Joanna J. Bryson. A proposal for the Humanoid Agent-builders League (HAL). In John Barnden, editor, *AISB'00 Symposium on Artificial Intelligence, Ethics and (Quasi-)Human Rights*, pages 1–6, 2000.

[Bryson, 2010] Joanna J. Bryson. Robots should be slaves. In Yorick Wilks, editor, *Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues*, pages 63–74. John Benjamins, Amsterdam, March 2010.

[Cope, 2005] David Cope. *Computer models of musical creativity.* MIT Press, Cambridge, MA, 2005.

[Cronin, 1991] Helena Cronin. *The Ant and the Peacock*. Cambridge University Press, 1991.

[Damon, 1983] William Damon. *Social and Personality Development: Infancy Through Adolescence*. W. W. Norton & Co., New York, 1983.

[Darwin, 1859] Charles Darwin. *On the Origin of Species by Means of Natural Selection*. John Murray, London, 1859.

[Date, 2004] C. J. Date. *An Introduction to Database Systems*. Addison Wesley, Reading, MA, 2004.

[de Garis, 1990] Hugo de Garis. The 21^{st} century artilect: Moral dilemmas concerning the ultra intelligent machine. In *Revue Interationale de Philosophie*, 1990.

[Dennett and Akins, 2008] Daniel C. Dennett and Kathleen Akins. Multiple drafts model. *Scholarpedia*, 3(4):4321, 2008.

[Dennett, 2001] Daniel C. Dennett. Are we explaining consciousness yet? *Cognition*, 79:221–237, 2001.

[Dennett, 2003] Daniel C. Dennett. *Freedom Evolves*. Viking, 2003.

[Dennett, 2006] Daniel C. Dennett. *Breaking the spell: Religion as a natural phenomenon*. Viking, 2006.

[Dreyfus, 1992] Hubert L. Dreyfus. *What Computers Still Can't Do*. MIT Press, Cambridge, MA, 1992.

[Eichin and Rochlis, 1989] Mark W. Eichin and Jon A. Rochlis. With microscope and tweezers: An analysis of the Internet virus of November 1988. In *The Proceedings of the IEEE Symposium on Security and Privacy*, pages 326–343, May 1989.

[Hantsaridou *et al.*, 2005] A. P. Hantsaridou, A. Th. Theodorakakos, and H. M. Polatoglou. A didactic module for undertaking climate simulation experiments. *European Journal of Physics*, 26(5):727–736, 2005.

[Helmreich, 1997] S. Helmreich. The spiritual in artificial life: Recombining science and religion in a computational culture medium. *Science as Culture*, 6(3):363–395, 1997.

[Henrich *et al.*, 2001] J. Henrich, R. Boyd, S. Bowles, C. Camerer, E. Fehr, H. Gintis, and R. McElreath. In search of homo economicus: behavioral experiments in 15 small-scale societies. *American Economic Review*, 91(2):73–78, 2001.

[Hobbes, 1651] Thomas Hobbes. *Leviathan*. London, Michael Oaskeshott edition, 1651.

[Hölldobler and Wilson, 2008] B. Hölldobler and E. O. Wilson. *The Superorganism*. Norton, London, 2008.

[Hsu, 2002] Feng-hsiung Hsu. *Behind Deep Blue: Building the Computer that Defeated the World Chess Champion*. Princeton University Press, 2002.

[Lowe, 2010] Will Lowe. Identifying your accompanist. In Yorick Wilks, editor, *Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues*, pages 95–100. John Benjamins, Amsterdam, March 2010.

[Markoff, 2009] John Markoff. Scientists worry machines may outsmart man. *The New York Times*, 26 July 2009.

[Nakamaru and Iwasa, 2006] Mayuko Nakamaru and Yoh Iwasa. The coevolution of altruism and punishment: Role of the selfish punisher. *Journal of Theoretical Biology*, 240(3):475–488, 2006.

[Parsons, 1991] Talcott Parsons. *The Social System*. Routledge, 1991.

[Ray, 1995] T. S. Ray. An evolutionary approach to synthetic biology: Zen and the art of creating life. In Christopher G. Langton, editor, *Artificial Life: An Overview*, pages 179–210. MIT Press, Cambridge, MA, 1995.

[Sagan and Newman, 1983] Carl Sagan and William I. Newman. The solipsist approach to extraterrestrial intelligence. *Quarterly Journal of the Royal Astronomical Society*, 24:113–121, 1983.

[Sharkey and Sharkey, 2010] Noel Sharkey and Amanda Sharkey. The crying shame of robot nannies: an ethical appraisal. *Interaction Studies*, 11(2):161–313, June 2010.

[Sklaroff, 1976] J. R. Sklaroff. Redundancy management technique for space shuttle computers. *IBM Journal of Research and Development*, 20(1):20–28, 1976.

[Vernon, 2010] David Vernon. Enaction as a conceptual framework for developmental cognitive robotics. *Journal of Behavioral Robotics*, 1(2):89–98, 2010.

[West *et al.*, 2011] Stuart A. West, Claire El Mouden, and Andy Gardner. Sixteen common misconceptions about the evolution of cooperation in humans. *Evolution and Human Behavior*, 2011. *in press*.

[Wiggins, 2006] Geraint A. Wiggins. Searching for computational creativity. *New Generation Computing*, 24(3):209–222, 2006.

[Williams, 1991] Heathcote Williams. *Autogeddon*. Jonathan Cape Ltd., London, 1991.