Just Another Artifact: The Necessity of the Empirical Experience of AI

Joanna Bryson
Phil Kime
Intelligent Systems Laboratory/
Centre for Cognitive Science
Edinburgh University, UK
`joanna@ai.mit.edu, phil.kime@ed.ac.uk`

February 9, 1996

Abstract

Why are grown scientists, let alone the general public, afraid that Artificial Intelligences (or "AIs") are going to take over the world, obsoleting humans? Why are people worried about the ethics of unplugging their computers, or whether AI spells the end of social order through the death of free will? In this paper we argue that these are manifestations of over-identification with AI, as well as general confusion over ethical standards. Any advances in AI technology will require relatively little change in our ethical views towards ourselves and artifacts, and where change might be useful, it is already required. In these cases, empirical experience of the general public with AI technology will actually be a useful catalyst.

Area(s):	Artificial Intelligence, Philosophy of Science and Mind, Ethics
Presentation Format:	Talk Only

Note: Both authors are students eligible for the Marr prize.

Just Another Artifact: The Necessity of the Empirical Experience of AI

Joanna Bryson
Phil Kime
Intelligent Systems Laboratory/
Centre for Cognitive Science
Edinburgh University, UK

1 Introduction

In the early March of 1995 Daniel Dennett chaired the EPIIC Workshop on Darwinism and Artificial Intelligence at Tufts University. Participants besides Dennett included Murray Gell-Mann, David Haig, John Holland, Kevin Kelly, Patti Maes, Bruce Mazlish, Marvin Minsky, Hans Moravec, Seymour Papert, Oliver Selfridge, Sherry Turkle, Rodney Brooks and Karl Sims. On the second day of the event the group began discussing a future world in which artificially intelligent machines ran the world, made all the decisions, and would inevitably eventually decide humanity was not worth worrying about or taking care of any more. Brooks later said, "I got called conservative for saying that wouldn't necessarily happen. I didn't say it *couldn't* happen, I said it *might* not, and they called me a conservative!"

This was not an isolated event. On October 5, 1994, Vernor Vinge (a professor at San Diego State University) gave a lecture at MIT entitled "The Coming Technological Singularity: How to Survive in the Post-Human Era." His abstract was the following:

> Within thirty years, we will have the technological means to create superhuman intelligence. Shortly after, the human era will be ended. I will attempt to make this claim plausible by outlining several different approaches to superhuman intelligence. [...] 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?

What can lead some computer scientists to believe that our society is endangered by

AI? After all, computer programs, including those classified as "artificial intelligence," are purpose-built artifacts designed and controlled by human beings. While computers can accelerate and magnify our mistakes to do more damage than an unaided individual is capable of, the same could be said of levers and pulleys. We believe the exaggerated fear of AI is symptomatic of a larger problem, a general confusion about the nature of humanity and the role of ethics in society. Much of this can in turn be understood in terms of an artificial distancing from empirical empathy. In this paper we will start with a brief exploration of how rules of relations are incorporated our society, that is, of our ethical constructs and how they relate to our sense of self. We will then discuss and attempt to dismiss exaggerated fears and hopes for the future of AI that are rooted in self concept. Finally, we will discuss what the true ethical conundrums for AI are and will be. We hope to demonstrate that these same issues face us already, and that AI may in fact *help* us to absorb technology into our ethical system.

2 Foundations

2.1 Empiricism and Ethics

When we mentioned earlier that many ethical conflicts stem from lack of empirical empathy, what did we mean? There are many things, the very idea of which elicit strong positive or negative reactions in many people. Artificial Intelligence is one of them. Taken in abstraction, AI seems to gnaw away at our sense of identity and threaten our values. This same feeling is common with emigrants and immigrants upon arriving in new cultures and gradually subsides after empirical immersion in the new culture. Similarly, gaining experience in working with AI is the best way to defuse the many harmless ethical implications of AI. The same has been true of the industrial and computer revolutions but as with these, people continue to worry that *new* technology may be unnerving, even as they come to take current technology for granted. We will explore these fears in the following sections, but currently we undertake to explore the phenomenon of empathy and its role in social structure. Most arguments in this domain are based heavily on ethical intuition. Intuitions provide sources of unrest regarding AI research; it is therefore important to approach the problems at this level.

In the history of philosophy, many have held that the primary or even sole purpose of ethics is to maintain a functional degree of social homogeneity, or to protect the social organism (for example [Hobbes 47]). To this extent, the details of an ethical system might be seen as somewhat arbitrary. Certainly the ethics of a culture adapts in many ways according to the caprice of the environment. One consequence of this is that a self-regulatory nature is an important aspect of ethical behaviour. Socially speaking, external control is far too resource intensive. If we were prevented from crime merely by the threat of external punishment rather than internal codes of ethics, the police force would have to number a good percentage of the population.

Our primary ethical reactions are often empathetic. Thus ethics requires some degree of perceived analogy between ourselves and that with which we should empathise. The degree of empathy we feel is a roughly sliding scale, with ourselves and our families tending to be at the top, followed by our neighbours and other people with whom we acknowledge commonality. In non-human animals, it peaks with dogs, cats, horses and the like and tails off famously with rodents and insects. Empathy seems to be rooted in the extent to which we can identify with another, and this identification may be described in psycho-pathological terms. Children have well-documented life-long identity confusion with their parents, possibly founded in a failure of categorisation occuring between the discriminating self and other in infancy. (Parents, perhaps confusingly, often have the same goals as the child, such as keeping it warm and happy.) In turn, parents can understand children as perpetuations of their own bodies and knowledge.

If we come to understand our more complicated relationships with friends, state and religion in terms of transference from familial relationships (as might be suggested by [Lakoff & Johnson 80]) then we might quite literally feel that actions that benefit them are in our own best interest.

Thus transferance and identification are two manifestations of our underlying empathetic tendencies. Together, they may provide important elements of our social and ethical structures. Other potential elements are our tendency to be extremely over-optimistic in our estimates of our own levels of control and likelihoods for success, even in situations with which we have experience [Taylor & Brown 88]. These biases can contribute to our willingness to participate in situations like war and capitalism, which can have lower probable return for some individuals than for their society as a whole. If self-interest is the root of our ethics, then it is easy to see how ethical systems can be self-regulating.

2.2 Incorporating AI

Therefore the question, on this intuitive empathetic ground, is whether or not an AI application or robot can elicit perceptions of analogy necessary to engender the required degree of empathy for ethical feelings. A most important aspect of these empathetic reactions is that they must be *robust*. Something that we fail to empathise with over time and also through differing situations will have difficultly in holding an ethical status. For example, we may empathise with the plight of a Queen bee when we see her mating trials on a nature documentary but loose this when we see her eating her young. AI workers need to realise that if an ethical appreciation of their creations is at all possible, the empathetic reactions of humans must demonstrate constancy. To do so, empathy requires empirical reinforcement in many situations and over time. Learning ethical feelings is an empirical matter and not something you can convince people of by arguments about "virtual consciousness" and the like. Contrary to the intentions of many philosophers, change in belief is rarely achieved by anything other than hard, often irrational, empirical work.

An important area of controversy is that surrounding the attitude of religion to AI research. This is a particularly sensitive area as it is sometimes suggested that AI, in particular the evolutionary sections thereof, are a direct threat to fundamental religious doctrines and thereby our ethical feelings. However, we wish to suggest that as the advent of AI rode on the back of the communications revolution, there are reasons why AI can positively contribute towards a cohesion of ethical feeling. It is transparent to all that religions tend to be mutually exclusive, particularly those in widely differing cultures. This fact was much explored in the heyday of cultural relativism. What is often under-appreciated is that the communications revolution breaks down many barriers that insulate orthogonal cultures, thus bringing conflicting beliefs into sharper relief and closer contact. One only has to peruse any of the social or politically oriented Usenet groups to see this. Since ethics are often linked with culture and religion, cultural hostility or restructuring can lead to ethical breakdown. A unifying theme in this increasing confusion would be a benefit but cannot be expected to stem from those systems that are so heavily centred in their own culture, like organised religion. AI and other sciences are, in part, attempting to address the fundamental themes of human characteristics from a certain point of view. This can, we think, have a rather unifying effect on the perceptions leading to a beneficial empathy.

AI is an important contribution to systematising and helping us to see some kind of pattern in our empathy and thus our ethical feelings. We are presented with a robot that we empathise with in some way. This is a startling occurrence as it contradicts our feelings about inanimate objects, metal, plastic, automation and such. Thus we are helped to explore what we are doing when we empathise. Empirical interactions in such situations are extremely important as they provide a real event where our feelings are tested. In effect, AI can be an empirical proving ground for critical ethics. In examining "automatic" objects, we help ourselves become less ethically automatic. John Stuart Mill was a great advocate of critical thought in absolutely all areas of human endeavour. This was quite a radical position at the time. We are suggesting that AI is a tool enabling this noble pursuit to continue into an age where the conceptual complexities of an issue often obscure the fruits to be gathered from empirical experience of it. We rarely explore the ways in which we are encouraged to empathise: we empathise at random much of the time with inanimate objects, animals and sometimes even people.

As we mentioned, an important aspect of ethical behaviour is its self-regulatory nature. We believe that coherent self-regulation is really only possible through critical thought.

Critical thought obtained through conceptual analysis is a luxury (if it is a luxury) available to very few. Most people are prompted to thought by the world: by experiences in reality. Thoughts about empathy and the appropriate objects thereof are thus entertained as a result of contact with possible objects of empathy. Of extreme value here are objects of an ambiguous status with which it is not clear whether we should or should not empathise. Such objects are invaluable in modifying our self-regulating codes of practise. More importantly, we argue that one feels more empathy with that which displays some form of self-regulatory behaviour itself, thus raising the ethical status of AI creations rather than lowering our own status. To object that we might be "tricked" into ethical feelings for a robot is both misguided and a *petitio principii*. If to be "tricked" is to be guided by analogy and empathy then most of our ethical feelings towards humans are tricks also. We cannot be forced into empathising with an object because of the appropriateness of a predicates such as "perceives" or "is conscious." We can be convinced by interaction with that object, regardless of its qualities however, which is how we begin many of our empathetic relationships.

3 Spectres of AI

In the introduction, we stated that both our exaggerated fears and our exaggerated affection for AI are based on over-identification. This may seem at odds with the previous section, where we state that empirical experience with the systems will help us to incorporate them into our ethical framework, and indeed make our ethical framework more robust. The problem with our exaggerated concepts of AI is that they are not based on deep experience with or understanding of real systems, but rather on distorted images of both ourselves and AI. Because computers display behaviours which we have previously seen only in humans, such as language and mathematics, we can initially be quickly moved to unjustified empathy [Weizenbaum 66]. The authors, gravitating towards the behaviour based AI community, would argue strongly that our intelligence, being body based, parallel, and rooted in reflexes and imitative learning of culture, has much more to do with animal intelligence than with the mathematically based logics of computer science (see further [Brooks 91]). Classically humanity has distanced itself from the rest of nature using precisely such talents as mathematics and language. Hopefully having more diverse counter-examples of such paradigms of humanity will give us a better sense of self, and thus a better basis for a progressive ethical structure.

In this section, we address the major misapprehensions about AI in terms of misidentification with computers or machines. In the following section we will discuss the social and ethical place of real, and realistically perceived, AI.

3.1 Reductionism and the Loss of Free Will

One source of fear about AI is simply that, if it works, it proves that our intelligence (which we classically consider to be the same as "us") is the inevitable consequence of the interaction of simple, deterministic parts. It is the "inevitable" and "deterministic" parts of this description that trouble people; this would seem to prove that free will is an illusion, which would then (the argument goes) eliminate the concept of personal responsibility. Not only that but we could, in addition, create objects that were not responsible for anything they might do.

Responsibility, however, is a definition. It is a socially assigned role: we are responsible for what we do if society considers us to be. Similarly, free will is a powerful concept that helps us organise our own behaviour, regardless of arguments as to its reality. The fear that AI might reveal underlying determinism in our own behaviour is probably a vain hope given the state of AI today. However, theoretically, could this be a real problem? We think not. Any possible future account of the low-level deterministic elements constituting our behaviour would almost certainly be so complex that the decomposition in any particular case would be impossible. It is not something that could give rise to anything other than weak concern, akin to the concern people feel when they are told about everything being made of atoms: will their books fall through their desks? The atomicity of solids is never even vaguely apparent in a concrete case and thus this angst can never obtain a hold on us.

We are still so stunningly ignorant regarding the interrelationships of the various parts and states of our bodies that it is unlikely that we will obtain a picture that allows prediction and control: two things that opponents of AI fear the most. What few people realize is that no atom beyond helium has yet been successfully modelled from first principles of mathematics and particle physics, yet our knowledge of physics is staggeringly advanced next to our knowledge of biology. Just as chemists are still forced to think at the level of atoms and molecules, and drug manufacturers still use trial, error and intuition, finding the "first principles" of intelligence would almost certainly not give us prediction and control of human behaviour. Certainly from the individual perspective, free will will continue to be our qualitative experience.

Interestingly, Hume had a powerful argument to the effect that deterministic operation is *necessary* to our idea of responsibility [Hume 48]. An act without determination is an act without cause and this is a random event. Random events with no cause are no basis for attributions of responsibility. Whilst we would not like to here advocate the accompanying view of causality and metaphysics, this raises an important point. Determinism has some very useful and desirable consequents as regards responsibility. In recent years, we have become obsessed with responsibility; there has been an explosion in the number of external and internal business accountability agencies, particularly in the public sphere. We introduce strict determinism where there is none, in order to facilitate attribution of responsibility. Now, one central feature of human artifacts is that we generally know how they work in minute detail. The products of AI, being human creations, are open to the sort of analysis that pre-existing natural artifacts such as people and hopelessly complex emergent artifacts such as societies are not. Thus, attributions of responsibility cut a considerably less daunting figure in this case. To find out exactly why something happened in an AI system, we can debug. To find out the same from a human, we can only guess, ask, theorise and in most well known instances of this endeavour, conduct a police interview. In an important sense, responsibility can be better pinpointed when you better understand the artifact. Thus we feel that there are positive *benefits* to the maintenence of responsibility where AI is concerned.

3.2 Identification and Obligation

There are two possible consequences of over-identifying with a machine. The first, as just mentioned, is that it lowers your own opinion of yourself. The other is that it elevates the worth of the machine. This, we propose, is the basis for both exaggerated fears and affinity for the promise of artificial intelligence. For example, in identifying with our machines, we endow them with the rights and privileges of ethical protection. We are somewhat embarrassed by this sense of obligation, and so try to rationalise it, while at the same time distancing ourselves. For example, many people think it would be unethical to unplug a computer if it were conscious. But what is consciousness? If its apparent complexity turns out to be more a result of the complexity of the system in which it is embedded, then consciousness itself may be relatively straight forward. If we find a need for implementing a simple form of it, does that change our ethical obligation? What if consciousness is just a shell script on unix (or a batch command in DOS) that loops?

Of course, part of our attribution of humanity to artifacts we identify with is out of our own desire for power. We would like to be more intelligent, to live longer, to be stronger. If not ourselves, then our progeny might have these characteristics. So we hear scientists and science fiction writers speak of self-replicating space ships being 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!). Science fiction has thoroughly examined the idea of robots that are deserving of citizen status. But these works of literature are exploring what it means to be human, not what it means to be a computer.

Our fears of obsolescence are the reverse side of our hopes for immortality. Again, we are thinking more about what would happen if we were to have the attributes of a machine than of what kinds of machines we actually build. Although the Internet worm of the mid eighties demonstrated that significant damage can be done by unexpected computer access, computer security is a heavily researched and practiced area of science. Protecting ourselves from hacking motivated by human intelligence is essentially identical to protecting ourselves from something a machine "decided" to do. Being hacked is bad regardless of who or what does the hacking.

Essentially, machines are created to fill useful roles (or occasionally as art). 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. There is a threat here not that machines out of maliciousness will take over the world, but that every human endeavour will eventually be "better" accomplished by a machine. But 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 accomplishments. What, for example, happens when people choose to write a letter by hand instead of sending email, because writing "looses 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 for you. An obsession with the results of action rather the actions themselves is not something that AI needs to address. It is a fault of our culture and we should not fear to tackle it [Burns 69].

4 Ethics and AI

Our argument here is not that there are no ethical considerations to creating and using AI technology. Rather, we are arguing that the *nature* of our ethical obligations are 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," depending on how we wish to use them. But if we do choose to create systems that could meaningfully be described in such terms, then we should take the same precautions against their potential flaws we do against human errors and fallibilities already. Nondeterministic computing should be audited regularly, by humans, and possibly monitored by other parallel programs with substantially different architectures. These kinds of precautions are already standard even for conventional computer programs in critical applications, such as manned space flights.

AI systems are already being used by major industries. For example, computer manufacturers use expert systems (AI programs embodying some amount 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 now 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 would ever be arrested directly as a consequence of these programs: exceptions are simply flagged and turned over to human account officers, who phone the customer and verify whether they were making the unexpected purchases. 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 these systems have been integrated by usage and empirical experience rather than theoretical assurances of desirable attributes.

What about our ethical obligations to intelligent systems? To answer this we must refer back to the fact we have already mentioned — that ethical systems can be seen to be largely arbitrary, and that one of their attributes is the maintenance of social order. Since we are provided with a new ethical quandary, we are to some extent free to create a new ethical standard. This standard should be consistent with our overall code of ethics, and should presumably contribute towards a social order we find desirable. As we have argued, the change need not be strongly qualitative; there need be no harsh transition. Instead, it is an ongoing process in which we are already participating.

Ethics generally leads us to be altruistic towards things which are like us. If we decide that the axis of "likeness" 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 culture? After all, much of what we are — our language, our music, our socialising — is less inborn than learned. If an artifact becomes a vessel of our culture, should we treat it with the same respect as a person? This question is already with us. Books, buildings, art, songs and languages all embody the intellectual output of members of our species. If it is 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. While 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 the lifestyle of a region. But even a defensive war is still a sacrifice of life to cultural artifact. It is justified because the cultural artifact is seen to so enrich the lives of the population as to merit the risk of loss of life to some percentage of them.

Another possible criteria is *contributing* to culture, rather than simply retaining it. We suggest that AI is a tool of creativity, not a creator in itself. The creators are those who designed the AI system. This would place a "creative" AI system on the same standard of any work of art or scholarship, for example the Mona Lisa or "On the Origin of Species". Regardless of the model of contribution, the moral conundrums are the same here as above. The question is, if an artifact is retaining or generating more intellectual information than a person, should people be allowed to die to preserve it? This is a difficult problem, but as we have said, 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. In AI we actually have a luxury not afforded to da Vinci — we can to some extent resolve the difficulty by ensuring our systems are not unique. With computer science we have the ability to ensure that every aspect of our work is replicable, and "backed up". Even if a system learns from irreproducible experience, the resulting internal hardware states could normally be preserved. So, in a sense, AI reduces the possibility of such ethical problems. Replicability is replaceability and thus deciding between things is easier. Indeed, this important property of computer-based systems allows us to be positively biased in favour of people as they are unique and qualitatively irreplaceable.

If an AI machine 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. As we said, a purpose of ethics is to maintain social order, and as such it has always involved balancing obligations to various sources: to yourself, your family, your country, your dog. 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.

5 Conclusion

Our conclusion then is that in creating and using intelligent artifacts we do need to consider ethical and social dangers, but to no greater respect than we should show "conventional" technology. If they control our defence systems, our utility supplies or our political propaganda, we should take exactly the same care with them we should currently pay our computers, piping and media. To the extent that they serve, like us, as vessels of our culture, we owe them the same respect we owe libraries, architecture and works of art. We believe the extent to which people feel exaggerated fears or obligations with respect to AI and robotics is a consequence of their over-identification with the artifacts due to superficial, unfamiliar similarities such as a machine's use of language. It also indicates our society's uncertainty in ethics which is in turn symptomatic of the stresses on our self-regulatory strategies resulting from our suddenly heterogenous culture.

We have stressed the importance of empirical experience in forming ethical obligations in new situations, and suggested that experience and understanding of real AI programs may actually assist us in our ethical quandaries. They could help us form better models of what it is to be human, and thus provide us with a better basis for empathetic decisions. Empathy and therefore ethics is dependent on real experience rather than theoretical urgings. Those worried about the acceptability of AI technology in our culture should worry less about anthropomorphising and more about making AI visible and understood where it already exists, such as in our video cameras and our credit system.

The cognitive science community, being composed of researchers in artificial intelligence, artificial life and complex systems, not to mention those concerned with philosophy and human behaviour, is particularly well suited for analysing, understanding and even modelling the role of ethics in society. In this paper, we have undertaken to use some of our knowledge of human psychology, intelligent systems of agents, and philosophy of mind and ethics to describe what the problems and solution spaces look like in an area that has so far been either neglected or misunderstood by most of the AI community. Our understanding of AI strongly effects our understanding of ourselves and society, and these in turn affect our attitudes and funding of AI. Since the problems we have addressed are both scientifically interesting and important to our society, we hope to see much more discussion and even experimental psychological research in this area.

Acknowledgements

Though the ideas in this paper largely came from discussions between the authors, its motivation came from discourse and disagreements with Rodney Brooks and Anne Foerst. Thanks also to Daniel Dennett and Oded Maron, and to Brooks and Lynn Stein for bringing philosophers and theologians into the MIT AI lab.

References

[Brooks 91] Rodney A. Brooks. Intelligence without reason. In *Proceedings of the 1991 International Joint Conference on Artificial Intelligence*, pages 569–595, August 1991.

[Burns 69] Robert Burns. *Poems and Songs.* Oxford University Press, Oxford, James Kinsley edition, 1969.

[Hobbes 47] Thomas Hobbes. *Leviathon.* London, Michael Oaskeshott edition, 1947.

[Hume 48] David Hume. *Essay Concerning Human Understanding.* London, 1748.

[Lakoff & Johnson 80] George Lakoff and Mark Johnson. *Metaphors We Live By.* University of Chicago Press, Chicago, Illinois, 1980.

[Taylor & Brown 88] S.E. Taylor and J.D. Brown. Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin,* (103):193–210, 1988.

[Weizenbaum 66] Weizenbaum. Eliza - a computer program for the study of natural language communication between man and machine. *Communications of the Association for Computing Machinery,* 9(1):36–44, 1966.