

# An Outline for Determining the Ethics of Artificial Intelligence

Richard Lucas

Centre for Applied Philosophy and Public Ethics

Charles Sturt University

The central question of this paper is: Can there be an ethical artificial intelligence (AI)?

What do I mean when I say: Can an AI be ethical? On the face of it this seems like a question with an obvious answer; no. So at first glance either that is the answer or it is the case that the question just seems odd to ask in the first place: Of course AIs can't be ethical only people can. There are other possibilities of course. Some say that other categories of beings such as animals (Clark, 1977) and groups (McMahon, 2001) are able to be moral.

In spite of the seeming common sense (How can my web browser possibly be ethical??) of the above there is a body of literature that does take the question seriously. In the literature four answers are contemplated; YES, NO, WHO KNOWS?, and WHO CARES?.

Due to the limited size of this paper I will address this last answer, WHO CARES, in detail and merely sketch an outline of the other answers.

## *A What makes the question worth asking - WHO CARES?*

Upon examination, that the question: Can AIs be ethical?, is worth asking is, straightforward. There are four reasons why this is true: relinquishing control, uncritical acceptance, technologizing society, and the other important questions that might be addressed once its answers are spelt out in detail. Examples showing the importance of each of these will be included in the paragraphs following.

RELINQUISHING CONTROL. If AIs were to have no effective intrusion into our lives, that is, if they were simply figments of the imaginings of science fiction, then, whether they could be ethical would be of interest to only those who engage in thought experiments. But AIs do intrude. Of great interest, but peripheral to the question of this paper is the whether they ought to intrude. [There are of course different sorts of intrusion. For example, I feel compelled to answer emails but that kind of intrusion is, at least in part, of my choice; I do not have to answer them - it just seems prudent to do so. Others may feel differently about their emails. The kinds of intrusion I have in mind here are of two sorts: the sorts that we have seemingly no choice but to interact with and those that are intimately bound up with cultural change. The first kind are those that, in some way, force themselves upon us in the normal course of our daily lives. For example, I cannot choose to not be affected by the AI-controlled automatic pilot in the aircraft that I fly in. Having chosen to get cash late at night there are few choices other than automatic teller machines (ATMs) that I can interact with. The second kind are those that cause and are the result of significant changes to culture generally and human, that is to say personal, interaction in particular. Both the automatic pilot and ATM examples above also fit into this category of intrusion. Specifically in the ATM example they have caused and have been caused by social change. It seems that there is little we can do to avoid interacting with such technology without significant inconvenience.]

Some argue that it is right for them to intrude in areas where they would perform better than we. Moor (1985) has argued just this view but noted an important exception: in the determining our values. This cannot be the whole story though. There are many areas of human achievement that even if AI did perform better, we still ought to do them. Amongst examples are those tasks we do for the simple pleasures of doing them: exercise. For our own physical and, probably psychological well-being we ought to engage in some form of physical exercise. There are other tasks such as those that Lenman (2001) argues for; those that fulfill our need to participate in life as human beings. Here he finds whole classes of tasks (composing music, doing science, talk with other humans, etc.) that we need to perform simply to show that we are *in* life and not merely consuming life.

This need for us to do things aside, given that it is clear that AIs do intrude into our lives the question of interest for this paper is how much control does that intrusion involve?

AIs are ever encroaching on domains of activity previously thought available to people only. Even more, this encroachment is with our compliance: we relinquish more and more control of our lives to AI-controlled technology. (Arguably we create and use technology to extend our domain of control but the result is that, frequently, things do get out of our control. The reasons why this seems to be the case are not explored in this paper.) An example of this is the social restructuring caused by the ATMs referred to above. This restructuring of cash dispersal in turn leads to changed expectations about the kinds and amount of cash that we will chose to have. ATM technology determines which denominations of currency that are available to be dispensed. This in turn forces us into taking a different amount than we might desire if that desired amount does not fit in with the denominations held by the ATM. The ATMs that I frequent dispense only two denominations; twenty and fifty dollar units. If I want to take out an amount that is not an integral multiple of these two numbers then I have to select an amount which is larger than what I want just to get my desired amount. If I want \$30 then I have to take out at least \$40 which is more than I want; in a worse case the ATM might be out of twenties in which case I must take out \$50 – 1 more than planned. Now this sort of inconvenience contributes, at least in part, to another social change; the use of plastic (credit, and less usually, debit) cards; relinquishing our previously held ideas about currency, finance, and, in the end, independence.

A consequence of this relinquishing is that we are taking less and less active part in decisions which have moral import. Evidence the Russian airline disaster in which a plane load of children going on holiday and an American cargo plane crashed into each other over Switzerland. In this case “The voice recorders show that a Swiss air traffic controller's order for a Russian pilot to descend contradicted the cockpit warning system's command for the Tu-154 to climb, the investigators said. The automatic cockpit warning systems issued simultaneous instructions for the Russian passenger jet to climb and a cargo jet to descend about 45 seconds before they ultimately collided over southern Germany, killing all 71 people on board. But one second after the on-board system warnings, the Zurich tower, which was in charge of directing the planes even though they were flying over Germany, told the Russian plane to descend, German investigators said, citing voice recorders from both planes recovered at the crash site.” (CBS, 2002) In this case clearly this relinquishing has had disastrous consequences.

But what has this relinquishing of control to AIs got to do with them being ethical? The answer to this question lies in the assumptions that we have when we do the same relinquishing with human beings. That is, what assumptions do we operate with when we give over control of parts of our lives to other humans? It is the case that whenever we give over control of parts of our lives to other humans we do two things: the first is that we establish that these other humans can and will carry out our wishes for that part of our lives and, secondly, we hold them responsible for the actions that they carry out as part of that control (See Wolgast (1992) for how responsibility plays out in these sorts of circumstances). Does this hold when we do the same thing with AIs? It seems that we might be justified in asking AIs to have the same capacities and carry the same responsibility. However, the 'carrying out of our wishes' necessarily includes some sort of assumption about, or determination of, the moral consideration that AIs might have (See Johnson and Powers, *Computers as Surrogate Agents* (1994) for a thoughtful and clear account of moral consideration of this sort). That is, moral consideration usually implies that the agent carrying out our wishes take into account our moral values. Instead of demanding that they take only our values into account in their acting we might accord them some sort of independence in moral values and trust that these values are similar enough to ours to get the desired outcome. Whatever stance we take however, it is the case that in all circumstances where moral delegation occurs we assume that some sort of (imposed or inherent) moral capability is present in the being that has been delegated to. The issue I am exploring in this paper is, how far beyond being merely the extension of a programmer's morality can an AI's morality be. That is say, to what extent would/do AIs have autonomy of the sort that is needed for moral agency, how extensive would the features or characteristics AIs have, have to be? This morality, when considered fully, would as well as moral consideration, include notions of moral worth, moral constraint, moral personhood, and being morally praiseworthy.

Another reason to care: UNCRITICAL ACCEPTANCE.

Take Turing's now famous quote:

"The original question, 'Can computers think?' I believe to be too meaningless to deserve discussion. Nevertheless I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted. I believe further that no useful purpose is served by concealing these beliefs." (Turing, 1950, p.442)

The common interpretation of this statement is that machines *will* think. What is more likely, and important for this paper, is that people will *believe* that machines can think. This belief is, to some, all that matters. The thinking goes something like: If it is believed that AIs can think then why not believe that they can be ethical? This line of thought might be explored in detail in the WHO KNOWS answer and involves the Intentional Stance Theory proposed by such philosophers as Dennett (1987). Even if the Intentional Stance is all that is needed, and in spite of the current state of affairs of the nature of AIs, (To date the action taking capabilities of AIs has been limited to giving advice to people. There are exceptions of course; think of automatic pilots and automatic teller machines, but these are, so far, fairly well restricted in their effect on people. That is, they have limited power to act in-the-world.) people do seem to have a tendency to adopt an attitude of acceptance. If this attitude of acceptance is extended to the (moral) decision making and action taking of AIs then their autonomy will grow and they will become more independent of us. The possible impacts of this independence ought to give us cause for pause and reflection.

TECHNOLOGIZING SOCIETY. The third reason to care if AIs can be ethical is the affect that they might have in changing society if they were able to be ethical. One affect might be that the incorporation of machine agents into human practices will accelerate and deepen as artefacts simulate basic social capacities: dependence upon them will grow. Human relations will be technologised to the extent that such artefacts are able to participate as agents in social interaction rather than merely mediate it. The encounter with these artefacts will occur earlier and earlier in human development. They will thereby take part in the sociocultural learning by which skilled practices, and the values they express, are transmitted. The attribution of human like agency to artefacts will change the image of both machines and of human beings. As Mumford (1963) and McLuhan (1966) both realised, technology shapes the cultural conditions within which people develop the shared skills and values that allow them to live together. These conditions now include agent-like artefacts with which human beings will need to co-exist. Of course, there are some, Pickering (2000) for example, that are not at all concerned by this.

Given the destructiveness of contemporary society, an examination of the additional influence that an ethical AI would have in the technologising of human social relations is timely.

The final reason: WHAT ELSE? It seems that we ought to investigate the question of AIs being ethical even if only to be able to answer a host of additional questions that depend upon the answer to the paper question. Knowing whether AIs can be ethical leads to questions such as: Ought we to consider more carefully the degree to which we give AIs effective control over morally charged parts of our lives? What moral controls ought to be built into AIs?, and What does this mean for *our* notions of moral responsibility? What does this mean for our *taking* of moral responsibility? While these questions are also worthy of examination they are beyond the scope of this paper. What is not beyond the scope of this paper is the following imperative: Given that the pervasive incursion of AIs into our lives is unlikely to wane, it is incumbent upon us to examine the possibility and especially, desirability and warrant, of constricting, controlling, and containing their ethical impact. For that we need to know the nature and extent of their ethicality.

It is to these ends that the paper's question is posed.

## ***B      A precise examination of the structure and content of the question***

To provide a thoughtful and considered examination of the question of AIs being ethical requires care and attention to the precise structure of the question and the words chosen. It is in this spirit that I now explicate what I mean by the words in the question: CAN there BE an *ethical AI*?

Now, for practical considerations (ie containing the size and complexity of this paper) the scope of the terms 'AI' and 'ethical' are, here, circumscribed. This is done in the next section where I state my grounding assumptions. This then leaves the terms CAN and BE; how to clarify them?

For many the need to clarify the terms such as CAN and BE might seem odd. Surely everyone knows what is meant by them. Unfortunately both in philosophy and AI technical language

is frequently used to make specific and detailed points about matters of subtly and fine distinction. It is this technical use of ordinary words that might lead some uninitiated readers to assume the ordinary usage when a specific understanding is required. [When an author deliberately trades on this misreading they engage in what Saul (1993) calls the hijacking of language; something he considers to be an immoral act.] This misunderstanding of the use of terms can lead the reader to then misunderstand the points that the writer is trying to make. For example, many AI researchers use the term CAN in a very loose way. They can mean anything from, simply having the possible, potential, capacity to do something, to the more specific, narrow, sense of merely being able to physically act in a particular way. For my purposes here and for most moral theories this sort of usage is simply not adequate. For CAN what is needed is more than mere physical causality, more than physical behaviour.

I use the term CAN in a unproblematic sense, meaning, possibly, or able to, without actually implying that an agent might actually carry out the action associated with the verb. I also mean to use CAN in the strong sense which encompasses, understanding, analyzing/deciding, and acting. I mean for it to be used as such in the idea expressed by Kant when he claimed that to will the act is to will the means. By this Kant meant that if we make a rule then we must, necessarily, have the means to carry out that rule. It seems straightforward that, to make a rule or law (especially a moral law) and not know that we have the means to follow it or carry it out is at least shortsighted and probably nonsensical (This is notwithstanding the Little Engine That Could when it expressed faith in saying 'I think I can, I think I can'). Kant meant that if we make a rule then, we must both, have a way (the means) to carry out that rule and, *know* that we have a way of carrying out the rule, before postulating the rule that ought to be followed.

I use the term BE in a similar way. Here, BE is an active verb and not meant to be restricted to merely existing. It includes notions of reasoning, intending, deciding, and acting. Is it the intention (which implies motivation) of the AI that is the heart of the matter? Is it the reasoning process that is ethical, that is, is the ethical stance taken by a AI taken because the AI has made a decision which can be considered to be arrived at through ethical deliberation? Is it the theory/guideline because the AI has followed some ethical theory, code, rules, or the like? Is it action, that is, is the action taken by an AI to be considered the ethical part? All of these possible meanings are taken into account in providing an answer to the paper question Can AIs be ethical?

There is another sense in which BE is used and that is in the way it is used in – Can AIs be fat? I do not use this sense of BE rather than the sense meant in my paper question because in this, 'fat', sense the question is passive, asking a quantitative question, asking a question about what I (and Kant) will call the sensible world. Now this sort of BE can be answered in some empirical way. My, Can AIs be ethical? question rather is psychological, as well as anticipating possible affirmative answers to the questions at the beginning of this paragraph. It is the non-sensible question that is of most interest to me in this paper.

Also I put a lot of emphasis on this usage of BE rather than emphasising ETHICAL which would seem to be the more important of these two terms. The reason for this emphasis on BE, or rather the not emphasising ETHICAL, is twofold; firstly to emphasise the psychological and secondly to not get bogged down in the disputes over what counts as ETHICAL.

## **C      *Assumptions and Limitations***

To make the topic of this paper, and the whole enterprise generally, of a manageable size I make the following assumptions.

- I      I take the terms ethics and morality to be synonyms. When I use these terms I intend that the reader assume I am addressing the following: the classical questions of value, the way we determine questions of right and wrong, and answers to the question ‘How should I live?’**
  
- II     In this paper I limit my examination of ethics to Kant’s moral theory. There are, of course, many different (sometimes wildly different) interpretations of Kant’s moral theory. Anyone engaged in this program ought, as much as possible, refer to Kant’s writing directly and call upon Kant researchers for explanatory material.**

I accept this uncritically, not because it is flawless but rather because I am investigating its applicability as a moral theory for AIs. What this acceptance requires will be begun to be spelt out in detail in Kant for all Rational Beings.

- III    I need not take a stand on the mind-body problem nor on functionalism. I take it that these are not crucial to my discussion. The mind-body problem is not crucial because I am open to the many interpretations of this problem that are able to encompass being ethical. Functionalism is not crucial because either, holding, or not holding, a functionalist attitude may be consistent with being ethical.**
  
- IV    I concede that both humans and AIs are both some kind of machine; admittedly different kinds of machines but machines none the less. Much of the peripheral literature that might be of use to the program proposed in this paper makes much of the argument: humans are machines, AIs are machines, therefore humans are AIs. Logically the conclusion does not follow from the premises. If it were to then the term machine loses most, if not all, useful meaning. Under the common definitions of machine there are many things which count as machines but are not the same. While both are machines no one would, for example, say that spacecraft and thermometers are the same.**

I assume without argument that the fact of both being machines is insufficient to disqualify them from being ethical. If one is a machine and also ethical also assume that merely being different kinds of machines is insufficient to disqualify the other one of them from being ethical.

- V     I assume that humans and AIs are substantially and relevantly different. These differences will be spelt out in the paper. The only question is whether these differences prevent AIs being ethical.**

**VI** I do not take a position on the classical question, Can AIs think? I leave it as an open question whether the characteristics necessary for thinking bear any relationship with the characteristics necessary for AIs being moral. Now many would take issue with this stance with something like: Surely thinking must be absolutely essential to being moral: How can an unthinking thing make moral decisions without engaging in something that could be called thinking? There is some work (See Brooks' 1991 *Intelligence without reason*. I also refer the reader to Waller's 1996 paper *Moral Commitment without Objectivity or Illusion: Comments on Ruse and Woolcock*.) which proposes related questions and claim, for example, that intelligence is possible without such common notions as reason and commitment. Considering such questions and examining in sufficient detail what other researchers have written is beyond the capacity of this paper. It is not that such questions are unimportant, rather that they are too large to deal with adequately here.

**VII** I also do not take a position on such related issues as AI intelligence and AI consciousness. Naturally this might lead someone to object with the question: How can something be (morally) responsible if it is not aware of what it is doing? Surely consciousness is essential to being moral and to exclude it is to remove much of what it means to be moral. Note that not taking a position allows me to, later, consider various definitions of, for example, consciousness that are at odds with classical uses of the term. See, for example, Birnbacher's 1995 work on artificial consciousness or Dennett's 1995 paper entitled *Cog: Steps Towards Consciousness in Robots* as examples of work that might be examined for applicability. As these would take the discussion outside the bounds of this paper I take no position on either their value or their suitability.

I do not want to be closed to any particular instantiation of these characteristics. I leave it as an open question whether the characteristics necessary for being intelligent or conscious bear any relationship with the characteristics necessary for AIs being moral.

**VIII** I limit my examination of AIs to those that are instantiations of Turing Machines (these are idealized abstract machines that have the following characteristics: an indefinite memory, an instruction set, automatic sequential operation, programmes, and symbol manipulation capacity.) and hence, von Neumann architecture AIs. My notion of AI includes notions such as autonomous software agents but does not include quantum AIs

**IX** I do not consider human-AI hybrids such as cyborgs and transhumans. Rather than be confused about which characteristics of these hybrids originate with their humanity and which originate with their AI-ness I take the simpler and pure case of AI-type entities which do not have any characteristics which can be attributed, uniquely, to humans. I want it to be as clear as possible that the sorts of entities that I am concerned with are not human beings and the case for hybrids is unclear.

## A *Which Ethics?*

I have already referred to Kant's moral theory as a possible starting point for which moral theories that might be chosen to examine to answer the question: Can AIs be ethical? Of all the moral theories to choose from, why Kant's? I could have chosen either of the other two main moral theories, consequentialism or virtue theory, or any one of a number of lesser known (to the common man at least, if not the philosopher) moral theories such as Rawls' contract theory. It is even true that some of the authors who have written on the central question in my paper have chosen explanations not from moral theories to guide and control AI behaviour. They have used instead things like Asimov's Three Laws of Robotics. That these cannot work has been shown by myself (Lucas, 2003) and others (Clarke, 1993 for example) previously and those arguments will not be repeated here. While each of these other moral theories has much to recommend it, Kant's was chosen for, principally, two reasons; common sense and 'all rational beings'. These two reasons tie it to the concern raised by Turing some fifty years ago. Turing's concern was that people will, without questioning, accept AIs as thinking things (see Turing's quote earlier in this paper). It is this worry that gives us cause to pay attention to two things: what the *common understanding* of what AIs are capable of is, and, what *kinds of things* are capable of being moral. Kant, unlike the others, has something to say about each of these two concerns: common sense and 'all rational beings'.

### 1 Common Sense

In a rather lengthy passage Kant states the reasons for abiding with common or ordinary human understanding. He says that "ordinary human reason, ... knows well how to distinguish, what is good, what bad, and what is consistent inconsistent with duty." and that all we need do is "draw its attention to its own principle (in the manner of Socrates), thus showing that neither science nor philosophy is needed in order to know what one has do in order to be honest and good, and even wise and virtuous." (Kant, 1785, 404).

Ordinary or common understanding is well (better) equipped to deal with moral problems than the trained mind (ie philosophers) and ought to be used in determining what matters.

#### All Rational Beings

Clearly, Kant was interested in rational beings and not simply human beings:

"It may be added that unless we wish to deny to the concept of morality all truth and relation to a possible object, we cannot dispute that its law is of such widespread significance as to hold not merely for men but for *all rational beings* as such – not merely subject to contingent conditions and exceptions, but with *absolute necessity*"(Kant, 1785, 408).

Finally Kant's conception of rational being encompassed a Supreme rational being as well as supernatural ones.

That Kant's moral theory explicitly makes the two notions, common understanding and rational beings, important factors also makes it an ideal starting point, morally speaking, with which to explore the, possible, ethics of AIs.

## ***B Why von Neumann architecture AIs?***

The common conception of AIs (Science fiction, novels and movies, apart) is that of the kind of things that do our word processing, calculate our spreadsheets, control our automatic tellers and autopilots, keep our cars running, and, provide search engine results. (I am, of course, assuming that such AIs are acting in the physical world and as such are not purely logical machines that have no causal impact on the physical world. In short I view AIs to be both physical and logical and, not purely logical entities such as ideal Turing machines described earlier.) All of these are accomplished on what are known as digital AIs and it to these that we will look to when examining AIs for the capacity to be moral.

Modern digital AIs are von Neumann architecture machines and have the following characteristics: an arithmetic/logic unit, memory, a central processing unit, and input/output mechanisms. It is these sorts of machines that are used to do most of the modelling of AIs, as well as, human, brains, thought processes, and such like. It is these that are referred to when researchers discuss the relationship between AIs and human beings.

Additionally, from the work of Turing and von Neumann we know that all computation can be accomplished on a Turing Machine and, hence, instantiated in a von Neumann architecture AI. This paper does not consider machines that cannot compute in the sense of computing as understood as being capable of instantiated algorithms..

However two concepts that are implicit in such AIs need to be spelt out. These are: technological sufficiency and computation.

TECHNOLOGICAL SUFFICIENCY. Arguments have been put that AIs are not technologically advanced enough to be thought the equals of humans. Two common reasons are given for this; complexity and fallibility. The first, complexity is usually stated as the emergence theory of intelligence (or any other human characteristic, say emotion, you may care to posit). This paper does not examine this theory as there is no rejoinder to “but it is not complex enough” if it fails to appear (as so far, it has failed to do so); there is no proof except for such characteristics to simply appear, big-bang-like in an AI; the so-called emergence theory of – whatever characteristic that is lacking. (Something like the emergence theory of intelligence *is* usually called on in arguments supporting this line of thought *but* there are other difficulties. For example, it could be argued that the earth, containing as, it does, people (who have some property,  $x$ ), is sufficiently complex to have a property, say,  $x$ . This does not seem to be the case: Even Gaia hypothesisists do not go this far.) The second, fallibility, says that the technology is simply not good enough (yet) to contain the relevant feature, say being ethical. To overcome the objection of technological deficiency I make use of the notion of perfect technology.

### *Perfect Technology*

Perfect technology, first proposed by McMenamin and Palmer (1984), is the notion that all the components that make up an AI are as good (technologically) and as they might possibly be. It is introduced to overcome two obstacles: the first is that of the previous paragraph, fallibility; the second is that the current state of technology is incapable of being ethical. In answer to both of these, to make any progress on this issue, AI technology must be better than it currently is. The question becomes - how much better? Any technologically specific answer – should it fail to deliver the desired results – would seem to be prey to the same

problem that complexity faces; not yet. Perfect technology overcomes these obstacles. For, if being perfect fails to deliver the desired results then technology will never achieve what I am asking of it – be ethical.

What then, exactly, is perfect technology? As stated in the previous section, von Neumann architecture AIs are made up of an arithmetic/logic unit, memory, a central processing unit, and input/output mechanisms. Following McMenamin and Palmer (1984) I put together the arithmetic/logic unit, central processing unit, and input/output mechanisms into *processors* and put the memory into *containers*. That is to say, processors carry out the activities of the AI. The containers move the data between processors and memory as well as store the data for use by processors. Note that the containers do more than merely contain, they transport the data between where it is stored and where it is manipulated.

Now we can describe perfect technology. To quote McMenamin and Palmer (1984):

“If its technology were perfect, a system would have a perfect processor and a perfect container. A perfect processor would be able to do anything and everything instantly; that is, it would have infinite capabilities and infinite workload capacity. It would cost nothing, consume no energy, take no space, generate no heat, never make a mistake, and never break down.

A perfect container would have many of the same virtues. It wouldn't cost anything, and it would be able to hold an infinite amount of data. Any processor would be able to access conveniently the data it carried.”  
(McManamin and Palmer, 1984, p.16)

As containers also transport the data to and from processors, being perfect means that there are no concerns about whether the data that ended up at one end of this transaction was exactly the same data that started out.

‘Doing anything’ and ‘access conveniently’ means doing these thing in *zero* time.

COMPUTATION. I take computation to be the instantiation of algorithms in von Neumann architecture machines. I take an algorithm to be:

“a finite procedure, written in a fixed symbolic vocabulary , governed by precise instructions, moving in discrete steps, 1, 2, 3, ..., whose execution requires no insight, cleverness, intuition, or perspicuity, and that sooner or later comes to an end.” (Berlinski, 2000, p. xviv)

Berlinski means to say that an *effective* algorithm has these properties, especially the final part concerning the coming to an end. From this then we can say that to be an *effective* computation is to be the instantiation of an effective algorithm. This paper does not examine the class of problems that are known to be non-computable. The only question of interest of computation for this paper is whether the idea of ethics is computable. Additionally, the only condition for computation of interest here is for it to be realizable on some actual AI. Of necessity, for a computation to be realized it must be expressed in some language capable of being programmed on some, particular, AI.

### **C      *An outline of a possible program***

The remainder of paper is devoted to giving some detail to a possible program that could be explored to, in part, address some of the issues raised so far. This program ought to focus on: an examination of the arguments put by others; introduce a schema for Artificial Ethics ( $\mathcal{A}$ ); and draw some conclusions about the progress so far.

## 1 Part A – Examination

This examination has two elements that need to be attended to: a detailed examination of moral theories, Kant is used as an example here, and a series of critiques of the substantive written material for the three major answers to the question of whether AIs can be ethical. That is, WHO KNOWS, NO, and YES. This ordering is to deal with the two more straightforward cases, WHO KNOWS and NO, first before proceeding with the more difficult answer, YES.

**KANT FOR ALL RATIONAL BEINGS.** This ought to isolate those parts of Kant's moral theory that are addressed to all rational beings from those parts that are aimed at mere human beings. That is this division ought to provide the extra-human core to his moral theory. This of necessity involves deciding what the term *rational beings* includes and excludes.

**WHO KNOWS.** This ought to deal with the possibility that perhaps it is not possible to know if AIs can be ethical. Perhaps it does not matter if we know that AIs can be ethical, all that really matters is that they *seem* to us to be ethical. There are two very different notions behind these kinds of replies; answerability and, appearance and believability.

The following is some preliminary work on this answer. The first kind of WHO KNOWS reply addresses the notion of being able, at all, to answer the question of AIs being ethical. It might be unanswerable because it is; unknowable, undecidable, uncomputable, incomplete, or subject to scepticism. If it is unknowable then we cannot proceed further: there is nothing, ever, for us to say. If the question, is undecidable then we don't know if an answer can be found; maybe, maybe not. If it cannot be computed at all (because it is not the sort of thing that can be computed) or is computationally unfeasible (See Flannery (2001), p.154 for a layman's introduction to this topic), then it falls into the NO answer category. If it is incomplete then it is an example of the Turing Halting Problem and is also beyond the scope of this paper and will not be investigated here. If it is vulnerable to the sceptical objection then this might be from the general problem of scepticism about other minds. That is to say, how do we know that an AI has a (sufficient) mind to be able to be ethical? I acknowledge that there are those, such as Brooks (1991a), who take the position that we can have such things as intelligence without reason as well as those who argue for its essentialness. However, to contain the largess of this paper and not get bogged down in what are, essentially, not the major focus of this paper I accept, without argument, that they do have a mind sufficient to be ethical.

The second kind of WHO KNOWS reply, those that address appearance and believability, are described in the literature review following contains what is known as the as-if or Intentional Stance, which says that so long as they (for any particular they, worms, thermometers for example, but for my purposes, AIs) believably behave as if they do then we ought to accept that they are without actually knowing that they are. All that matters is that AIs act as if they *were* ethical: All that matters is that they are believable.

NO. Any program must analyze the literature that says AIs cannot be ethical. Any no answer to the question, Can AIs be ethical? would need to show either, what it is about human beings that make them moral agents that AIs lack, or that there is something about moral agency that makes it impossible for non-humans to be ethical no matter what their makeup might be. This would need to be stated taking into account the assumptions and limitations stated above.

YES. Of course the most interesting part of this program would be the YES answer. In a preliminary search I found nearly 100 sources which, in some (usually indirect) way, answer the program question in the affirmative. While, most of these do not examine the possibility of there being a *Kantian* moral AI as such (using the example given in this paper), most make the general YES claim with the thrust of their arguments centring around the idea that there are some reasons to think so and no compelling reasons to believe that they, at least in principle, cannot be so. Most of the affirmative authors say so without being committed to any of the positions taken up by those who have been placed in the WHO KNOWS camp.

There are a small number of authors (Coleman, Floridi, Driver, Stuart, and 6 are prominent among these) who have said that AIs can be (Kantian) moral persons.

## 2 Part B – Proposals

This second part of the program ought find the fit between the particular moral theory chosen (Kant's moral theory is the example chosen in this paper) and the conception of AIs outlined earlier. This ought to answer for example, questions such as: How well do AIs address the question of whether AIs, according to <insert moral theory here> can possibly be, such beings.

This second part also ought to propose a schema (what I shall label  $\mathcal{A}$ ) whereby any AI might be assessed against a criteria to determine its candidature for being an ethical person generally (and not merely a Kantian moral person). It might be the case that the isolated parts of Kant's moral theory identified earlier in the program could be used to form the basis of the development of the  $\mathcal{A}$  schema. After the schema is introduced the characteristics of existing AIs are compared to the schema and an evaluation made. Finally recommendations are made about the necessary criteria for any future AIs to be ethical persons

Of course any decent program ought to conclude with a summary concerning the state of current AIs being ethical and the possibilities for future AIs being so.

Carrying out this program would be an enormous task but not one to be abandoned because of that nor ought it be simplified to make the task manageable.

## D Bibliography

6, Perri. *Morals for Robots and Cyborgs, Ethics, society and public policy in the age of autonomous intelligent machines*. Middlesex: Bull Information Systems Ltd; 1999.

Berlinski, David. *The Advent of the Algorithm*. New York: Harcourt Inc.; 1999

- Birnbacher, D. Artificial consciousness. in Metzinger, T. *Conscious Experience*. Thorverton England: Imprint Academic; 1995; pp. 489-503.
- Brooks, Rodney Allen. Intelligence without reason. *IJCAI-91*. 1991:569-595.
- CBS. *Russian Airliner Got Mixed Signals* [Web Page]. 2002 Jul 8; Accessed 2002 Nov 26. Available at: <http://www.cbsnews.com/stories/2002/07/01/world/main513959.shtml>
- Clark, Stephen R. L. *The Moral Status of Animals*. Oxford: Oxford University Press ; 1977.
- Clarke, Roger. *Robot Rules, OK? - An Examination of Asimov's 'Laws of Robotics' Fiction*. IEEE Computer. December 1993, January 1994.
- Coleman, Kari Gwen. *Could Computers Be Kantian Persons?* [Web Page]. 1996 Sep 2; Accessed 2002 Nov 15. Available at: <http://www.andrew.cmu.edu/course/80-136/kari.html>.
- Coleman, Kari Gwen. *On the Moral Standing of Computers*: Rensselaer Polytechnic Institute; 1996 May.
- Dennett, Daniel C. Cog: Steps Towards Consciousness in Robots. in Metzinger, T. *Conscious Experience*. Thorverton England: Imprint Academic; 1995; pp. 471-487.
- Dennett, Daniel C. *The Intentional Stance*. The MIT Press; 1987.
- Flannery, Sarah and Flannery, David. *In Code: A Mathematical Journey*. New York: Workman Publishing; 2001.
- Floridi, Luciano and Sanders, J. W. *On the morality of artificial agents* [Web Page]. Accessed 2001 Dec 24. Available at: <http://www.wolfson.ox.ac.uk/~floridi/pdf/maa.pdf>
- Johnson, Deborah G. and Powers, Thomas M. *Computers as Surrogate Agents*. in: Bynum, Terrell Ward; Pouloudi, Nancy; Rogerson, Simon, and Spyrou, Thomas, Editor. *Ethcomp 2004 - Challenges for the Citizen of the Information Society*; Syros, Greece. University of the Aegean; 2004: 422-435.
- Johnson, Steven. *Emergence: The Connected Lives of Ants, Brains, Cities, and Software* [Web Page]. Accessed 2004 Jul 21.
- Kant, Immanuel. *Groundwork of the Metaphysic of Morals*. New York: Harper Torchbooks; 1992 [1785]; ISBN: 0-06-131159-6.
- Kavanaugh, John F. *Who Count as Persons?: Human Identity and the Ethics of Killing*. Washington D.C.: Georgetown University Press; 2001; ISBN: 0-87840-836-3.
- Kyberg, Henry E. How to Settle an Argument. in: Ford, Kenneth M.; Glymour, Clark, and Hayes , Patrick J., Editor. *Android Epistemology*. MIT Press; 1995; pp. 267-278.

- Lenman, Jimmy. On Becoming Redundant or What Computers Shouldn't Do. *Journal of Applied Philosophy*. 2001; 18:1-11.
- Lucas, Richard. Moral Theories for Autonomous Software Agents. in. *Papers and Abstracts, 2003 Computers and Philosophy Conference*; University House, Australian National University, Canberra, Australia.
- McLuhan, Marshall. Cybernation and Culture. in: Dechert, Charles R., Editor. *The Social Impact of Cybernetics*. Notre Dame: University of Notre Dame Press; 1966; pp. 95-108.
- McMahon, Christopher. *Collective Rationality and Collective Reasoning*. Cambridge: Cambridge University Press; 2001; ISBN: 0521011787.
- Moor, James H. Are There Decisions Computers Should Never Make? in: Johnson, Deborah G. and Snapper, John W., Editors. *Ethical Issues in the Use of Computers*. Belmont California: Wadsworth Publishing Company; 1985; pp. 120-130.
- Mumford, Lewis. *Technics and Civilization*. NY: Harcourt Brace and World; 1963.
- Pickering, John. *Agents and Ethics* [Web Page]. 2000 Apr; Accessed 2001 Jun 18. Available at: <http://www.cs.bham.ac.uk/~jab/AISB-00/Rights/Abstracts/pickering.html>.
- Saul, John Ralston. *Voltaire's Bastards – The Dictatorship of Reason in the West*. Toronto: Penguin Books; 1993.
- Sharlow, Mark F. *Can Machines Have First-Person Properties?* [Web Page]. 2001; Accessed 2002 May 7. Available at: <http://maxkpages.com/files/markphilosophy/firstpers.htm>.
- Steinhart, Eric. Emergent values for automatons: Ethical problems of life in the generalized internet. *Ethics & Information Technology*. 1999; 1155-160.
- Stuart, Susan. *Should Artificial Systems Have Rights?* [Web Page]. 2001; Accessed 2001 Jun 17. Available at: [http://www.justlikethat.com/client/philosophy\\_user/ai\\_paper/artrights.html](http://www.justlikethat.com/client/philosophy_user/ai_paper/artrights.html).
- Turing, Alan Mathison. Computing Machinery and Intelligence. *Mind*. 1950 Oct; LIX(236):433-460.
- Waller, Bruce N. Moral Commitment without Objectivity or Illusion: Comments on Ruse and Woolcock. *Biology and Philosophy*. 1996; 11:245-254.
- Wolgast, Elizabeth. *Ethics of an Artificial Person*. Stanford: Stanford University Press; 1992.