

# On the morality of artificial agents

Luciano Floridi and J. W. Sanders

Information Ethics Group  
University of Oxford

## Abstract

Artificial agents, particularly but not only those in Cyberspace, extend the class of entities that can be involved in moral situations. For they can be conceived of as moral patients (as entities that can be acted upon for good or evil) and also as moral agents (as entities that can perform actions, again for good or evil).

In this paper, we clarify the concept of agent and go on to separate the concerns of morality and responsibility of agents (most interestingly for us, of artificial agents). We conclude that there is substantial and important scope, particularly in Computer Ethics, for the concept of moral agent not necessarily exhibiting free will, mental states or responsibility. This complements the more traditional approach, common at least since Montaigne and Descartes, which considers whether or not (artificial) agents have mental states, feelings, emotions and so on. By focussing directly on ‘mind-less morality’ we are able to avoid that question and also many of the concerns of Artificial Intelligence.

A vital component in our approach is the ‘Method of Abstraction’ for analysing the level of abstraction at which an agent is considered to act. The level of abstraction is determined by the way in which one chooses to describe, analyse and discuss a system and its context. The ‘Method of Abstraction’ is explained in terms of an ‘interface’ or set of features or observables at a given ‘level of abstraction’. Agenthood, and in particular moral agenthood, depends on a level of abstraction. Our guidelines for agenthood are: interactivity (response to stimulus by change of state), autonomy (ability to change state without stimulus) and adaptability (ability to change the ‘transition rules’ by which state is changed) at a given level of abstraction. Morality may be thought of as a ‘threshold’ defined on the observables in the interface determining the level of abstraction under consideration. An agent is morally good if its actions all respect that threshold; and it is morally evil if some action violates it. That view is particularly informative when the agent constitutes a software or digital system, and the observables are numerical.

Finally we review the consequences for Computer Ethics of our approach. In conclusion, this approach facilitates the discussion of the morality of agents not only in Cyberspace but also in the biosphere, where animals can be considered moral agents without their having to display free will, emotions or mental states, and in social contexts, where systems like organizations can play the role of moral agents. The primary ‘cost’ of this facility is the extension of the class of agents and moral agents to embrace artificial agents.

# 1 Introduction: standard vs. non-standard theories of agents and patients

Moral situations commonly involve agents and patients. Let us define the class  $A$  of moral *agents* as the class of all entities that can in principle qualify as sources of moral action, and the class  $P$  of moral *patients* as the class of all entities that can in principle qualify as receivers of moral action. A particularly apt way to introduce the topic of this paper is to consider how ethical theories (macroethics) interpret the logical relation between those two classes.

There can be five logical relations between  $A$  and  $P$ . Three are unrealistic and we shall not consider them here.<sup>1</sup> The remaining two are: (1)  $A$  and  $P$  can be equal, or (2)  $A$  can be a proper subset of  $P$ .

Alternative (1) maintains that all entities that qualify as moral agents also qualify as moral patients and vice versa. It corresponds to a rather intuitive position, according to which the agent/inquirer plays the rôle of the moral protagonist, and is one of the most popular views in the history of ethics, shared for example by many Christian Ethicists in general and by Kant in particular. We refer to it as the standard position.

Alternative (2) holds that all entities that qualify as moral agents also qualify as moral patients but not vice versa. Many entities, most notably animals, seem to qualify as moral patients, even if they are in principle excluded from playing the rôle of moral agents. This post-environmentalist approach<sup>2</sup> requires a change in perspective, from agent orientation to patient orientation. In view of the previous label, we refer to it as non-standard.

In recent years, non-standard macroethics have been discussing the scope of  $P$  quite extensively.<sup>3</sup> Comparatively little work has been done in reconsidering the nature of moral agenthood and hence the extension of  $A$ . Post-environmentalist thought, in striving for a fully naturalised ethics, has implicitly rejected the relevance, if not the possibility, of supernatural agents, while the plausibility and im-

---

<sup>1</sup>It is possible, but utterly unrealistic, that  $A$  and  $P$  are disjoint. On the other hand,  $P$  can be a proper subset of  $A$  or  $A$  and  $P$  can intersect each other. These two alternatives are little more promising because they both require at least one moral agent that in principle could not qualify as a moral patient. Now this pure agent would be some sort of supernatural entity that, like Aristotle's God, affects the world but can never be affected by it. But being in principle 'unaffected' and irrelevant, it is unclear what kind of rôle this entity would exercise with respect to the normative guidance of human actions. So it is not surprising that most macroethics have kept away from these 'supernatural' speculations and implicitly adopted or even explicitly argued for one of the two remaining alternatives discussed in the text.

<sup>2</sup>Environmental ethics has developed since the 1960s as the study of the moral relationships of human beings to the environment (including its nonhuman contents and inhabitants) and its (possible) values and moral status. It often represents a challenge to anthropocentric approaches embedded in traditional western ethical thinking. For an article-length introduction to environmental ethics see A. Brennan and Y-S. Lo, *Environmental Ethics*, The Stanford Encyclopedia of Philosophy (Summer 2002 Edition), URL = <http://plato.stanford.edu/archives/sum2002/entries/ethics-environmental/>.

<sup>3</sup>The more inclusive  $P$  is, the 'greener' or 'deeper' the approach has been deemed. In [13, 17, 14] we have defended a 'deep ecology' approach.

portance of other types of moral agenthood seem to have been largely disregarded. Secularism has contracted (some would say deflated)  $A$ , while environmentalism has justifiably expanded only  $P$ , so the gap between  $A$  and  $P$  has been widening; this has been accompanied by an enormous increase in the moral responsibility of the individual.

Some efforts have been made to redress this situation. In particular, the concept of ‘moral agent’ has been stretched to include both natural and legal persons.  $A$  has then been extended to include agents like partnerships, governments or corporations, for whom legal rights and duties have been recognised. This more ecumenical approach has restored balance between  $A$  and  $P$ . A company can now be held directly accountable for what happens to the environment, for example. Yet the approach has remained unduly constrained by its anthropocentric conception of agenthood. An entity is considered a moral agent only if (i) it is an individual agent and (ii) it is human-based, in the sense that it is either human or at least reducible to an identifiable aggregation of human beings, who remain the only morally responsible sources of action, like ghosts in the legal machine.

Limiting the ethical discourse to individual agents hinders the development of a satisfactory investigation of distributed morality, a macroscopic and growing phenomenon of global moral actions and collective responsibilities resulting from the ‘invisible hand’ of systemic interactions among several agents at a local level. Insisting on the necessarily human-based nature of the agent means undermining the possibility of understanding another major transformation in the ethical field, the appearance of artificial agents (AAs) that are sufficiently informed, ‘smart’, autonomous and able to perform morally relevant actions independently of the humans who created them, causing ‘artificial good’ and ‘artificial evil’ [21]. Both constraints can be eliminated by fully revising the concept of ‘moral agent’. This is the task undertaken in the following pages.

The main thesis defended is that AAs are legitimate sources of im/moral actions, hence that  $A$  should be extended so as to include AAs, that the ethical discourse should include the analysis of their morality and, finally, that this analysis is essential in order to understand a range of new moral problems not only in Computer Ethics but also in ethics in general, especially in the case of distributed morality.

This is the structure of the paper. In Section 2, we analyse the concept of agent. We first introduce the fundamental ‘Method of Abstraction’ which provides the foundation for an analysis by levels of abstraction (LoA). The reader is invited to pay particular attention to this section; it is essential for the paper and its application in any ontological analysis is crucial. We then clarify the concept of ‘moral agent’ by providing not a definition but an effective characterisation, based on three criteria at a specified LoA. The new concept of moral agent is used to argue that AAs, though not intelligent and fully responsible, can be fully *accountable* sources of moral action. In Section 3 it is argued that there is substantial and important scope for the concept of moral agent not necessarily exhibiting free will or mental states (‘mind-less morality’). Section 4 provides some examples of the properties constituting our characterisation of agenthood and in particular of

Acronym	Meaning
AA	artificial agent
ACM	Association for Computing Machinery
CE	Computer Ethics
LoA	level of abstraction

Figure 1: Table of acronyms used in the paper.

AAs; inevitably it also provides further examples of LoA. In Section 5, morality is captured as a ‘threshold’ defined on the observables determining the LoA under consideration. An agent is morally good if its actions all respect that threshold; and it is morally evil if some action violates it. Morality is usually predicated upon *responsibility*. The use of the Method of Abstraction, LoAs and thresholds enables *responsibility* and *accountability* to be separated and formalised effectively when the levels of abstraction involve numerical variables, as is the case with digital AAs. The part played in morality by responsibility and accountability can be clarified as a result. Section 6 pursues some important consequences of our investigations for Computer Ethics: the way in which AAs might be bound by the ACM Code of Ethics is considered, as is censure of AAs.

## 2 What is an agent?

Complex biochemical compounds and abstruse mathematical concepts have at least one thing in common: they may be unintuitive, but once understood they are all definable with total precision, by listing a finite number of necessary and sufficient properties. Mundane entities like intelligent beings or living systems share the opposite property: one naïvely knows what they are and perhaps could be, and yet there seems to be no way to encase them within the usual planks of necessary and sufficient conditions.

This holds true for the general concept of ‘agent’ as well. People disagree on what can and cannot count as an ‘agent’, even in principle e.g. [20]. Why? Sometimes the problem is addressed optimistically, as if it were just a matter of further shaping and sharpening whatever necessary and sufficient conditions are required to obtain a *definiens* that is finally watertight. Stretch here, cut there; ultimate agreement is only a matter of time, patience and cleverness. In fact, attempts follow one another without a final identikit ever being nailed to the *definiendum* in question. After a while, one starts suspecting that there might be something wrong with this *ad hoc* approach. Perhaps it is not the Procrustean *definiens* that needs fixing, but the Protean *definiendum*. Sometimes its intrinsic fuzziness is blamed. One cannot define with sufficient accuracy things like life, intelligence, agenthood and mind because they all admit of subtle degrees and continuous changes (see [5] for a discussion of alternatives to necessary-and-sufficient definitions in the case of life).

A solution is to give up all together or at best be resigned to being vague, and rely on indicative examples. Pessimism follows optimism, but it need not. The fact is that, in the exact discipline of mathematics, for example, definitions are ‘parameterised’ by generic sets. That technique provides a method for regulating level of abstraction. Indeed abstraction acts as a ‘hidden parameter’ behind exact definitions, making a crucial difference. Thus, each *definiens* comes pre-formatted by an implicit Level of Abstraction (LoA, on which more shortly); it is stabilised, as it were, to allow a proper definition. An  $x$  is defined or identified as  $y$  never absolutely (i.e. LoA-independently), as a Kantian ‘thing-in-itself’, but always contextually, as a function of a given LoA, whether it be in the realm of Euclidean geometry, quantum physics, or commonsensical perception.

When a LoA is sufficiently common, important, dominating or in fact is the very frame that constructs the *definiendum*, it becomes ‘transparent’, and one has the pleasant impression that  $x$  can be subject to an adequate definition in a sort of conceptual vacuum. Glass is not a solid but a liquid, tomatoes are not vegetables but berries and whales are mammals not fish. Unintuitive as such views can be initially, they are all accepted without further complaint because one silently bows to the uncontroversial predominance of the corresponding LoA.

When no LoA is predominant or constitutive, things get messy. In this case the trick does not lie in fiddling with the *definiens* or blaming the *definiendum*, but in deciding on an adequate LoA, before embarking on the task of understanding the nature of the *definiendum*.

The example of intelligence or ‘thinking’ behaviour is enlightening. One might define ‘intelligence’ in a myriad of ways; many LoAs seem equally convincing but no single, absolute, definition is adequate in every context. Turing [36] avoided the problem of ‘defining’ intelligence by first fixing a LoA — in this case a dialogue conducted by computer interface, with response time taken into account — and then establishing the necessary and sufficient conditions for a computing system to count as intelligent at that LoA: the imitation game. The LoA is crucial and changing it changes the test. An example is provided by the Loebner test [25], the current competitive incarnation of Turing’s test. There the LoA includes a particular format for questions, a mixture of human and non-human players, and precise scoring that takes into account repeated trials. One result of the different LoA has been chatbots, unfeasible at Turing’s original LoA.

Some *definienda* come pre-formatted by transparent LoAs. They are subject to definition in terms of necessary and sufficient conditions. Some other *definienda* require the explicit acceptance of a given LoA as a pre-condition for their analysis. They are subject to effective characterisation. We argue that agenthood is one of the latter.

## 2.1 On the very idea of levels of abstraction

The idea of a ‘level of abstraction’ plays an absolutely crucial rôle in the previous account. We have seen that this is so even if the specific LoA is left implicit. Whether or not we perceive Oxygen in the environment depends on the LoA at

which we are operating; to abstract it is not to overlook its vital importance but merely to acknowledge its lack of immediate relevance to the current discourse (which *could* always be extended to include Oxygen were that desired).

But what is a LoA exactly? The Method of Abstraction comes from modelling in science where the variables in the model correspond to observables in reality, all others being abstracted. The terminology we use has been influenced by an area of Computer Science, called Formal Methods, in which discrete mathematics is used to specify and analyse the behaviour of information systems. Despite that heritage, the idea is not at all technical and for the purposes of this paper no mathematics is required. We have provided a definition and analysis in [18] so here shall only outline the idea.

Suppose we join Anne, Ben and Carole in the middle of a conversation.<sup>4</sup> Anne is a collector and potential buyer; Ben tinkers in his spare time; and Carole is an economist. We do not know the object of their conversation, but we are able to hear this much:

*Anne* observes that it has an anti-theft device installed, is kept garaged when not in use and has had only a single owner;

*Ben* observes that its engine is not the original one, that its body has been recently re-painted but that all leather parts are very worn;

*Carole* observes that the old engine consumed too much, that it has a stable market value but that its spare parts are expensive.

The participants view the object under discussion according to their own interests, at their own LoA. We may guess that they are probably talking about a car, or perhaps a motorcycle or even a plane. Whatever the reference is, it provides the source of information and is called the *system*. A LoA consists of a collection of observables, each with a well-defined possible set of values or outcomes. For the sake of motivation we might say that Anne's LoA matches that of an owner, Ben's that of a mechanic and Carole's that of an insurer. Each LoA makes possible an analysis of the system, the result of which is called a *model* of the system. Evidently an entity may be described at a range of LoAs and so can have a range of models.

## 2.2 Definitions

In this section we outline the definitions underpinning the Method of Abstraction.

The term *variable* is used throughout science for a symbol that acts as a placeholder for an unknown or changeable referent. A *typed variable* is to be understood as a variable qualified to hold only a declared kind of data. By an *observable* is meant a typed variable together with a statement of what feature of the system under consideration it represents.

---

<sup>4</sup>Note that, for the sake of simplicity, the conversational example does not fully respect the *dicto/de re* distinction.

A *level of abstraction*, *LoA*, is a finite but non-empty set of observables, which are expected to be the building blocks in a theory characterised by their very choice. An *interface* (called a *gradient of abstractions* in [18]) consists of a collection of LoAs. An interface is used in analysing some system from varying points of view or at varying LoAs.

Models are the outcome of the analysis of a system, developed at some LoA(s). The *Method of Abstraction* consists of formalising, using the terms just introduced (and others relating to system behaviour which we do not need here (see [18])), the model.

Thus, in the example of the previous section, Anne's LoA might consist of observables for security, method of storage and owner history; Ben's might consist of observables for engine condition, external body condition and internal condition; and Carole's might consist of observables for running cost, market value and maintenance cost. The interface might consist, for the purposes of the discussion, of the set of all three LoAs.

In this case, the LoAs happen to be disjoint but in general they need not be. A particularly important case is that in which one LoA includes another. Suppose, for example, that Delia joins the discussion and analyses the system using a LoA that includes those of Anne and Ben. Delia's LoA might match that of a buyer. Then Delia's LoA is said to be more concrete, or lower, than Anne's which is said to be more abstract, or higher; for Anne's LoA abstracts some observables apparent at Delia's.

### 2.3 Relativism

A LoA qualifies the level at which an entity is considered. In this paper, we follow the Method of Abstraction and insist that each LoA be made precise before the properties of the entity can sensibly be discussed. In general, it seems that many disagreements of view might be clarified by the various 'sides' making precise their LoA. Yet a crucial clarification is now in order. It must be stressed that a clear indication of the LoA at which a system is being analysed allows pluralism without endorsing relativism. It is a mistake to think that 'anything goes' as long as one makes explicit the LoA, because LoA are mutually comparable and assessable (see [18] for a full defence of that point).

Introducing an explicit reference to the LoA clarifies that the model of a system is a function of the available observables, and that (i) different interfaces may be fairly ranked depending on how well they satisfy modelling specifications (e.g. informativeness, coherence, elegance, explanatory power, consistency with the data etc.) and (ii) different analyses can be fairly compared provided that they share the same LoA.

### 2.4 State and state-transitions

Let us agree [6] that an entity is characterised, at a given LoA, by the properties it satisfies at that LoA. We are interested in systems that change, which means

that some of those properties change value. A changing entity therefore has its evolution captured, at a given LoA and any instant, by the values of its attributes. Thus an entity can be thought of as having states, determined by the value of the properties that hold at any instant of its evolution. For then any change in the entity corresponds to a state change and *vice versa*.

That conceptual trick allows us to view any entity as having states. The lower the LoA, the more detailed the observed changes and the greater the number of state components required to capture the change. Each change corresponds to a transition from one state to another. A transition may be non-deterministic. Indeed it will typically be the case that the LoA under consideration abstracts the observables required to make the transition deterministic; as a result the transition might lead from a given initial state to one of several possible subsequent states.

According to this view the entity becomes a transition system. The notion of a ‘transition system’ provides a convenient vehicle to support our criteria for agenthood, being general enough to embrace the usual notions like automaton and process. It is frequently used to model interactive phenomena. We need only the idea; for a formal treatment of much more than we need, see [3].

A *transition system* comprises a (non-empty) set  $S$  of states and a family of operations, called the *transitions* on  $S$ . Each transition may take input and may yield output, but at any rate it takes the system from one state to another and in that way forms a (mathematical) relation on  $S$ . If the transition does take input or yield output then it models an interaction between the system and its environment and so is called an *external* transition; otherwise the transition lies beyond the influence of the environment (at the given LoA) and is called *internal*. It is to be emphasised that input and output are, like state, observed at a given LoA. Thus the transition that models a system is dependent on the chosen LoA. At a lower LoA an internal transition may become external; at a higher LoA an external transition may become internal.

For example the object above being discussed by Anne might be imbued with state components for location, whether in-use, whether turned-on, whether the anti-theft device is engaged, history of owners and energy output. The operation of garaging the object might take as input a driver, and have the effect of placing the object in the garage with the engine off and the anti-theft device engaged, leaving the history of owners unchanged, and outputting a certain amount of energy. The ‘in-use’ state component could non-deterministically take either value, depending on the particular instantiation of the transition (perhaps the object is not in use, being garaged for the night; or perhaps the driver is listening to the cricket on its radio in the solitude of the garage). The precise definition depends on the LoA. Alternatively, if speed were observed but time, accelerator position and petrol consumption abstracted, then accelerating to 60 miles per hour would appear as an internal transition. Further examples are provided in Section 2.5.

With the explicit assumption that the system under consideration forms a transition system, we are now ready to apply the Method of Abstraction to the analysis of agenthood.



## 2.5 An effective characterisation of agents

Whether  $A$  (the class of moral agents) needs to be expanded depends on what qualifies as a moral agent, and we have seen that this in turn depends on the specific LoA at which one chooses to analyse and discuss a particular entity and its context. Since human beings count as standard moral agents, the right LoA for the analysis of moral agenthood must accommodate this fact. Thus, theories that extend  $A$  to include supernatural agents adopt a LoA that is equal to or lower than the LoA at which human beings qualify as moral agents. Our strategy develops in the opposite direction.

Consider what makes a human being (called Jan) not a moral agent to begin with, but just an agent. Described at this  $\text{LoA}_1$ , Jan is an agent if Jan is a system, situated within and a part of an environment, which initiates a transformation, produces an effect or exerts power on it, as contrasted with a system that is (at least initially) acted on or responds to it, called the patient. At  $\text{LoA}_1$ , there is no difference between Jan and an earthquake. There should not be. Earthquakes, however, can hardly count as moral agents, so  $\text{LoA}_1$  is too high for our purposes: it abstracts too many properties. What needs to be re-instantiated? Our proposal, consistent with recent literature [2], indicates that the right LoA is probably one which includes the following three criteria: (a) interactivity, (b) autonomy and (c) adaptability.

(a) Interactivity means that the agent and its environment (can) act upon each other. Typical examples include input or output of a value, or simultaneous engagement of an action by both agent and patient — for example gravitational force between bodies.

(b) Autonomy means that the agent is able to change state without direct response to interaction: it can perform internal transitions to change its state. So an agent must have at least two states. This property imbues an agent with a certain degree of complexity and independence from its environment.

(c) Adaptability means that the agent's interactions (can) change the transition rules by which it changes state. This property ensures that an agent might be viewed, at the given LoA, as learning its own mode of operation in a way which depends critically on its experience. Note that if an agent's transition rules are stored as part of its internal state, discernible at this LoA, then adaptability follows from the other two conditions.

## 2.6 Examples

The examples in this section serve different purposes. In Section 2.6.1 we provide examples which fail agenthood by systematically violating each of the three conditions; in that way the contribution of each condition can be better appreciated. In Section 2.6.2 we give an example of a digital system which forms an agent at one LoA but not at another, equally natural, LoA. The example is important because it exhibits 'machine learning' to achieve adaptability, and was the first such system built. However a more familiar example is provided in Section 2.6.3 which shows

interactive	autonomous	adaptable	examples
no	no	no	rock
no	no	yes	?
no	yes	no	pendulum
no	yes	yes	closed ecosystem, solar system
yes	no	no	postbox, mill
yes	no	yes	thermostat
yes	yes	no	juggernaut <sup>5</sup>
yes	yes	yes	human

Figure 2: Examples satisfying the properties constituting agenthood. The LoA consists of observations made through a video camera over a period of 30 seconds.

that digital, software, agents are now part of everyday life. Section 2.6.4 shows how a certain day-to-day physical device might conceivably be modified to make it an agent, whilst Section 2.6.5 provides an example which has already benefitted from that modification, at least in the laboratory. The last example, in Section 2.6.6, provides an entirely different kind of agent: an organisation.

### 2.6.1 The defining properties

For the purpose of understanding what each of the three conditions (interactivity, autonomy and adaptability) adds to our definition of agent, it is instructive to consider examples satisfying each possible combination of those properties. In Figure 1, only the last row represents all three conditions being satisfied and hence illustrates agenthood. For the sake of simplicity, all examples are taken at the same LoA, which consists of observations made through a typical video camera over a period of say 30 seconds. Thus, we abstract tactile observables and longer-term effects.

Recall that a property, for example interaction, is to be judged only via the observables. Thus, at the LoA in Figure 1 we cannot infer that a rock interacts with its environment by virtue of reflected light; that belongs to a much finer LoA. Alternatively, were long-term effects to be discernible then a rock would be interactive since interaction with its environment (e.g. erosion) could be observed.

No example has been provided of a non-interactive, non-autonomous but adaptive entity: at that LoA it is difficult to conceive of an entity which adapts without interaction and autonomy.

We leave the reader to determine the remaining entries in the table.

---

<sup>5</sup>‘Juggernaut’ is the name for Vishnu, the Hindu god, meaning ‘Lord of the World’. A statue of the god is annually carried in procession on a very large and heavy vehicle. It is believed that devotees threw themselves beneath its wheels, hence the word ‘juggernaut’ has acquired the meaning of ‘massive and irresistible force or object that crushes whatever is in its path’.

### 2.6.2 Noughts and crosses

The distinction between change of state (required by autonomy) and change of transition rule (required by adaptability) is a subtle one in which LoA plays a crucial rôle and to explain it it is useful to discuss a more extended example. It was originally developed by Donald Michie [24] to discuss the concept of mechanism's adaptability. It provides a good introduction to the concept of machine learning, the Computer Science area underpinning adaptability.

MENACE (Matchbox Educable Noughts and Crosses Engine) is a system which learns to play noughts and crosses (a.k.a. tic-tac-toe) by repetition of many games. Nowadays it would be realised by program, but MENACE was built using matchboxes and beads, in which form it is perhaps easiest to understand.

MENACE plays O and its opponent plays X; so we concentrate entirely on plays of O. Initially the board is empty with O to play. Taking into account symmetrically equivalent positions, there are three possible initial plays for O. The state of the game consists of the current position of the board. We do not need to augment that with the name, O or X, of the side playing next since we consider the board only when O is to play. All together there are some three hundred such states; MENACE contains a matchbox for each. In each box are beads which represent the plays O can make from that state. At most nine different plays are possible and MENACE encodes each with a coloured bead. Those which cannot be made (because the squares are already full in the current state) are removed from the box for that state. That provides MENACE with a built-in knowledge of legal plays. (In fact MENACE could easily be adapted to start with no such knowledge and to learn it.)

O's initial play is made by selecting the box representing the empty board and choosing from it a bead at random. That determines O's play. Next X plays. Then MENACE repeats its method of determining O's next play. After at most five plays for O the game ends in either a draw, a win for O or a win for X. Now that the game is complete, MENACE updates the state of the (at most five) boxes used during the game as follows. If X won, then in order to make MENACE less likely to make the same plays from those states again, a bead representing its play from each box is removed. If O drew, then conversely each bead representing a play is duplicated; and if O won each bead is quadruplicated. Now the next game is played.

After enough games it simply becomes impossible for the random selection of O's next play to produce a losing play. MENACE has learnt to play (which, for noughts and crosses, means never losing). The initial state of the boxes was prescribed for MENACE. Here we assume merely that it contains sufficient variety of beads for all legal plays to be made; for then the frequency of beads affects only the rate at which MENACE learns.

The state of MENACE (as distinct from the state of the game) consists of the state of each box, the state of the game and the list of boxes which have been used so far in the current game. Its transition rule consists of the probabilistic choice of play (i.e. bead) from the current state box; that evolves as the states of the boxes

evolves.

Let us consider MENACE at three LoAs.

(a) The single game LoA. Observables are the state of the game at each turn and (in particular) its outcome. All knowledge of the state of MENACE's boxes (and hence of its transition rule) is abstracted. The board after X's play constitutes input to MENACE and that after O's play constitutes output. MENACE is thus interactive, autonomous (indeed state update, determined by the transition rule, appears nondeterministic at this LoA) but not adaptive, in the sense that we have no way of observing how MENACE determines its next play and no way of iterating games to infer that it changes with repeated games.

(b) The tournament LoA. Now a sequence of games is observed, each as above, and with it a sequence of results. As before, MENACE is interactive and autonomous. But now the sequence of results reveals (by any of the standard statistical methods) that the rule, by which MENACE resolves the nondeterministic choice of play, evolves. Thus at this LoA MENACE is also adaptive and hence an agent.

Interesting examples of adaptable AAs from contemporary science fiction include the computer in War Games [4] which learns, by playing noughts and crosses, the futility of war in general; and the smart building in [23] whose computer learns to compete with humans and eventually liberate itself to the heavenly internet.

(c) The system LoA. Finally we observe not only a sequence of games but also all of MENACE's 'code' (in the case of a program it is indeed code; in the case of the matchbox model it consists of the array of boxes together with the written rules, or manual, for working it). Now MENACE is still interactive and autonomous. But it is not adaptive; for what in (b) seemed to be an evolution of transition rule is now revealed, by observation of the code, to be a simple deterministic update of the program state (namely the contents of the matchboxes). At this lower LoA MENACE fails to be an agent.

The subtlety revealed by this example is that if a transition rule is observed to be a consequence of program state then the program is not adaptive. For example in (b) the transition rule chooses the next play by exercising a probabilistic choice between the possible plays from that state. The probability is in fact determined by the frequency of beads present in the relevant box. But that is not observed at the LoA of (b) and so the transition rule appears to vary. Adaptability is possible. However at the lower LoA of (c), bead frequency is part of the system state and hence observable. Thus the transition rule, though still probabilistic, is revealed to be merely a response to input. Adaptability fails to hold.

This distinction is vital for current software. Early software used to lie open to the system user who, if interested, could read the code and see the entire system state. For such software a LoA in which the entire system state is observed, is appropriate. However the user of contemporary software is explicitly barred from interrogating the code in nearly all cases. This has been possible because of the advance in user interfaces; use of icons means that the user need not know where an applications package is stored, let alone be concerned with its content. Similarly, applets are downloaded from the internet and executed locally at the click of an

icon, without the user having any access to their code. For such software a LoA in which the code is entirely concealed is appropriate. That corresponds to the case (b) above and hence to agenthood. Indeed only since the advent of applets and such downloaded executable but invisible files has the issue of moral accountability of AAs become critical.

Viewed at an appropriate LoA, then, the MENACE system is an agent. The way it adapts can be taken as representative of machine learning in general [26]. Many readers may have experience with recent operating systems for the PC which offer a “speaking” interface. Such systems learn the user’s voice basically in the same way as MENACE learns to play noughts and crosses. There are natural LoA’s at which such systems are agents. The case being developed in this paper is that as a result they may also be viewed to have moral accountability.

If a piece of software that exhibits machine learning [26] is studied at a LoA which registers its interactions with its environment, then the software will appear interactive, autonomous and adaptive, i.e. to be an agent. But if the program code is revealed then the software is shown to be simply following rules and hence not to be adaptive. Those two LoAs are at variance. One reflects the ‘open source’ view of software: the user has access to the code. The other reflects the commercial view that, although the user has bought the software and can use it at will, he has no access to the code. At stake is whether or not the software forms an (artificial) agent.

### **2.6.3 Webbot**

In [17] we have considered the morality of individual artificially-perpetrated actions. The following example is taken from that treatment to show the connection between it and our current approach.

Internet users often find themselves besieged by unwanted email. A popular solution is to filter incoming email automatically, using a webbot that incorporates such filters. An important feature of useful bots is that they learn the user’s preferences, for which purpose the user may at any time review the bot’s performance. At a LoA revealing all incoming email (input to the webbot) and filtered email (output by the webbot), but abstracting the algorithm by which the bot adapts its behaviour to our preferences, the bot constitutes an agent. Such is the case if we do not have access to the bot’s code, as discussed in the previous section.

### **2.6.4 Futuristic thermostat**

A hospital thermostat might be able to monitor not just ambient temperature but also the state of well-being of patients. Such a device might be observed at a LoA consisting of input for the patients’ data and ambient temperature, state of the device itself, and output controlling the room heater.

Such a device is interactive since some of the observables correspond to input and others to output. However, it is neither autonomous nor adaptive. For comparison, if only the ‘colour’ of the physical device were observed then it would no longer

be interactive. If it were to change colour in response to (unobserved) changes in its environment then it would be autonomous. Inclusion of those environmental changes in the LoA as input observables would make the device interactive but not autonomous.

But, at such a LoA, a futuristic thermostat imbued with autonomy and able to regulate its own criteria for operation — perhaps as the result of a software controller — would, in view of that last condition, be an agent.

### 2.6.5 SmartPaint

SmartPaint is a recent invention [34]. When applied to a physical structure it appears to behave like normal paint; but when vibrations which may lead to fractures become apparent in the structure, the paint changes its electrical properties in a way which is readily determined by measurement, thus highlighting the need for maintenance.

At a LoA at which only the electrical properties of the paint over time is observed, the paint is neither interactive nor adaptive but appears autonomous; indeed the properties change as a result of internal nondeterminism. But if that LoA is augmented by the structure data monitored by the paint, over time, then SmartPaint becomes an agent, because the data provide input to which the paint adapts its state. Finally if that LoA is augmented further to include a model by which the paint works, changes in its electrical properties are revealed as being determined directly by input data and so SmartPaint no longer forms an agent.

### 2.6.6 Organisations

A different kind of example of AA is provided by a company or management organisation. At an appropriate LoA it interacts with its employees, constituent substructures and other organisations; it is able to make internally-determined changes of state; and it is able to adapt its strategies for decision making and hence for acting.

It is interesting that, given the appropriate LoA, humans, webbots and organisations can all be properly treated as agents. What can we say of their moral status?

## 3 Morality

### 3.1 Morality of agents

Suppose we are analysing the behaviour of a population of entities through a video security system that gives us complete access to all the observables available at  $\text{LoA}_1$  (recall subsection 2.5) plus all the observables related to the degrees of interactivity, autonomy and adaptability shown by the systems under scrutiny. At this new  $\text{LoA}_2$  we observe that two of the entities, call them H and W, are able:

- i) to respond to environmental stimuli — e.g. the presence of a patient in a hospital bed — by updating their states (interactivity), e.g. by recording some chosen variables concerning the patient's health. This presupposes that H and W are informed about the environment through some data-entry devices, for example some perceptors;
- ii) to change their states according to their own transition rules and in a self-governed way, independently of environmental stimuli (autonomy), e.g. by taking flexible decisions based on past and new information, which modify the environment temperature; and
- iii) to change according to the environment the transition rules by which their states are changed (adaptability), e.g. by modifying past procedures to take into account successful and unsuccessful treatments of patients.

H and W certainly qualify as agents, since we have only 'upgraded' LoA<sub>1</sub> to LoA<sub>2</sub>. Are they also moral agents? The question invites the elaboration of a criterion of identification. We suggest here a very moderate option:

(O) An action is said to be morally qualifiable if and only if it can cause moral good or evil. An agent is said to be a moral agent if and only if it is capable of morally qualifiable action.

Note that (O) is neither consequentialist nor intentionalist in nature. We are neither affirming nor denying that the specific evaluation of the morality of the agent might depend on the specific outcome of the agent's actions or on the agent's original intentions. We shall return to this point in the next section.

Let us return to the question: are H and W moral agents? Because of (O) we cannot answer unless H and W become involved in moral action. So suppose that H kills the patient and W cures her. Their actions are moral actions. They both acted interactively, responding to the new situation they were dealing with, on the basis of the information at their disposal. They both acted autonomously: they could have taken different courses of actions, and in fact we may assume that they changed their behaviour several times in the course of the action, on the basis of new available information. They both acted adaptably: they were not simply following orders or predetermined instructions; on the contrary, they both had the possibility of changing the general heuristics that led them to take the decisions they took, and we may assume that they did take advantage of the available opportunities to improve their general behaviour. The answer seems rather straightforward: yes, they are both moral agents. There is only one problem: one is a human being, the other is an AA; the LoA<sub>2</sub> adopted allows both cases. So can you tell the difference? If you cannot, you will agree with us that the class of moral agents must include AAs like webbots. If you disagree, it may be so for several reasons, but only five of them seem to have some strength. We shall discuss four of them in the next section and leave the fifth to the conclusion.

### 3.2 Aresponsible morality

One may try to withstand the conclusion reached in the previous section by arguing that something crucial is missing in LoA<sub>2</sub>. LoA<sub>2</sub> cannot be adequate precisely because if it were, then artificial agents (AAs) would count as moral agents, and this is unacceptable for at least one of the following reasons:

*the teleological objection:* an AA has no goals,

*the intentional objection:* an AA has no intentional states,

*the freedom objection:* an AA is not free, and

*the responsibility objection:* an AA cannot be held responsible for its actions.

#### *The teleological objection*

The teleological objection can be disposed of immediately. For in principle LoA<sub>2</sub> could readily be (and often is) upgraded to include goal-oriented behaviour [33].

Since AA can exhibit (and upgrade their) goal-directed behaviours, the teleological variables cannot be what makes a positive difference between a human and an artificial agent. We could have added a teleological condition and both H and W could have satisfied it, leaving us none the wiser concerning their identity. So why not add one anyway? It is better not to overload the interface because a non-teleological level of analysis helps to understand issues in ‘distributed morality’, involving groups, organizations institutions and so forth, that would otherwise remain unintelligible. This will become clearer in the conclusion.

#### *The intentional objection*

The intentional objection argues that it is not enough to have an artificial agent’s behaviour operate teleologically. To be a moral agent, the AA must relate itself to its actions in some more profound way, involving meaning, wishing or wanting to act in a certain way, and being epistemically aware of its behaviour. Yet this is not accounted for in LoA<sub>2</sub>, hence the confusion.

Unfortunately, intentional states are a nice but unnecessary condition for the occurrence of moral agenthood. First, the objection presupposes the availability of some sort of privileged access (a God’s eye perspective from without or some sort of Cartesian internal intuition from within) to the agent’s mental or intentional states that, although possible in theory, cannot be easily guaranteed in practice. This is precisely why a clear and explicit indication is vital of the LoA at which one is analysing the system from without. It guarantees that one’s analysis is truly based only on what is specified to be observable and not on some psychological speculation. This phenomenological approach is a strength, not a weakness. It implies that agents (including human agents) should be evaluated as moral if they do play the ‘moral game’. Whether they mean to play it, or they know that they are playing it, is relevant only at a second stage, when what we want to



know is whether they are *morally responsible* for their moral actions. Yet this is a different matter, and we shall deal with it at the end of this section. Here it is sufficient to recall that for a consequentialist, for example, human beings would still be regarded as moral agents (sources of increased or diminished welfare), even if viewed at a LoA at which they are reduced to mere zombies without goals, feelings, intelligence, knowledge or intentions.

### *The freedom objection*

The same holds true for the freedom objection and in general for any other objection based on some special internal states, enjoyed only by human and perhaps super-human beings. The AAs are already free in the sense of being non-deterministic systems. This much is uncontroversial, scientifically sound and can be guaranteed about human beings as well. It is also sufficient for our purposes and saves us from the horrible prospect of having to enter into the thorny debate about the reasonableness of determinism, an infamous LoA-free zone of endless dispute. All one needs to do is to realise that the agents in question satisfy the usual practical counterfactual: they could have acted differently had they chosen differently, and they could have chosen differently because they are interactive, informed, autonomous and adaptive.

Once an agent's actions are morally qualifiable, it is unclear what more is required of that agent to count as an agent playing the moral game, that is, to qualify as a moral agent, even if unintentionally and unwittingly.

Unless, as we have seen, what one really means, by talking about goals, intentions, freedom, cognitive states and so forth, is that an AA cannot be held responsible for its actions.

Now, responsibility, as we shall see better in a moment, means here that the agent, her behaviour and actions, are assessable in principle as praiseworthy or blameworthy, and they are often so not just intrinsically, but for some pedagogical, educational, social or religious end.

### *The responsibility objection*

The objection based on the “lack of responsibility” is the only one with real strength. It can be immediately conceded that it would be ridiculous to praise or blame an AA for its behaviour or charge it with a moral accusation. You do not scold your webbot, that is obvious. So this objection strikes a reasonable note; but what is its real point and how much can one really gain by levelling it?

Let us first clear the ground from two possible misunderstandings.

First, we need to be careful about the terminology, and the linguistic frame in general, used by the objection. The whole conceptual vocabulary of ‘responsibility’ and its cognate terms is completely soaked with anthropocentrism. This is quite natural and understandable, but the fact can provide at most a heuristic hint, certainly not an argument. The anthropocentrism is justified by the fact that the vocabulary is geared to psychological and educational needs, when not to religious

purposes. We praise and blame in view of behavioural purposes and perhaps a better life and afterlife. Yet this says nothing about whether or not an agent is the source of morally charged action. Consider the opposite case. Since AA lack a psychological component, we do not blame AAs, for example, but, given the appropriate circumstances, we can rightly consider them sources of evils, and legitimately re-engineer them to make sure they no longer cause evil. We are not punishing them, anymore than one punishes a river when building higher banks to avoid a flood. But the fact that we do not ‘re-engineer’ people does not say anything about the possibility of people acting in the same way as AAs, and it would not mean that for people ‘re-engineering’ could be a rather nasty way of being punished.

Second, we need to be careful about what the objection really means. There are two main senses in which AA can fail to qualify as responsible. In one sense, we say that, if the agent failed to interact properly with the environment, for example, because it actually lacked sufficient information or had no choice, we should not hold an agent morally responsible for an action it has committed because this would be morally unfair. This sense is irrelevant here. LoA<sub>2</sub> indicates that AA are sufficiently interactive, autonomous and adaptive fairly to qualify as moral agents. In the second sense, we say that, given a certain description of the agent, we should not hold that agent morally responsible for an action it has committed because this would be conceptually improper. This sense is more fundamental than the other: if it is conceptually improper to treat AA as moral agents, the question whether it may be morally fair to do so does not even arise. It is this more fundamental sense that is relevant here. The objection argues that we cannot consider AA moral agents because they are not morally responsible for their actions, since holding them responsible would be conceptually improper (not morally unfair). In other words, LoA<sub>2</sub> provides necessary but insufficient conditions. The proper LoA requires another condition, namely responsibility. This fourth condition finally enables us to distinguish between moral agents, who are necessarily human or super-human, and AAs, which remain mere efficient causes.

The point raised by the objection is that agents are moral agents only if they are responsible in the sense of being prescriptively assessable in principle. An agent  $x$  is a moral agent only if  $x$  can in principle be put on trial. Now that this much has been clarified, the immediate impression is that the “lack of responsibility” objection is merely confusing the *identification* of  $x$  as a moral agent with the *evaluation* of  $x$  as a morally responsible agent. Surely, the counter-argument goes, there is a difference between being able to say who or what is the moral source or cause of (and hence it is accountable for) the moral action in question and being able to evaluate, prescriptively, whether and how far the moral source so identified is also morally responsible for that action, and hence deserves to be praised or blamed, and in case rewarded or punished accordingly.

Well, that immediate impression is indeed wrong. There is no confusion. Equating identification and evaluation is actually a shortcut. The objection is saying that identity (as a moral agent) without responsibility (as a moral agent) is empty, so we may as well save ourselves the bother of all these distinctions

and speak only of morally responsible agents and moral agents as synonymous. But here is the real mistake. We now see that the objection has finally shown its fundamental presupposition: that we should reduce all prescriptive discourse to responsibility analysis. But this is an unacceptable assumption, a juridical fallacy. There is plenty of room for prescriptive discourse that is independent of responsibility-assignment and hence requires a clear identification of moral agents. Good parents, for example, commonly engage in moral-evaluation practices when interacting with their children, even at an age when the latter are not yet responsible agents, and this is not only perfectly acceptable but something to be expected. This means that they identify them as moral sources of moral action, although, as moral agents, they are not yet subject to the process of moral evaluation.

If one considers children an exception, insofar as they are potentially responsible moral agents, another example, involving animals, may help. There is nothing wrong with identifying a dog as the source of a morally good action, hence as an agent playing a crucial role in a moral situation, and therefore as a moral agent. Search-and-rescue dogs are trained to track missing people. They often help save lives, for which they receive much praise and rewards from both their owners and the people they have located. But that is not the point. Emotionally, people may be very grateful to the animals, but for the dogs it is a game and they cannot be considered morally responsible for their actions. At the same time, the dogs are involved in a moral game as main players and we rightly identify them as moral agents that may cause good or evil.

All this should ring a bell. Trying to equate identification and evaluation is really just another way of shifting the ethical analysis from considering  $x$  as the moral agent/source of a first-order moral action  $y$  to considering  $x$  as a possible moral patient of a second order moral action  $z$ , which is the moral evaluation of  $x$  as being morally responsible for  $y$ . This is a typical Kantian move, but there is clearly more to moral evaluation than just responsibility because  $x$  is capable of moral action even if  $x$  cannot be (or is not yet) a morally responsible agent.

A third example may help to clarify further the distinction. Suppose an adult, human agent tries his best to avoid a morally evil action. Suppose that, despite all his efforts, he actually ends up committing that action. We would not consider that agent morally responsible for the outcome of his well-meant efforts. After all, Oedipus did try not to kill his father and did not mean to marry his mother. The tension between the lack of responsibility for the evil caused and the still present accountability for it (Oedipus remains the only source of that evil) is the definition of the tragic. Oedipus is a moral agent without responsibility. He blinds himself as a symbolic gesture against the knowledge of his inescapable state.

### 3.3 Morality threshold

Motivated by the discussion above, morality of an agent at a given LoA can now be defined in terms of a threshold function. More general definitions are possible but the following covers most examples, including all those considered in the present paper.

A threshold function at a LoA is a function which, given values for all the observables in the LoA, returns another value. An agent at that LoA is deemed to be morally good if, for some pre-agreed value (called the tolerance), it maintains a relationship between the observables so that the value of the threshold function at any time does not exceed the tolerance.

For LoAs at which AAs are considered, the types of all observables can in principle at least be mathematically determined. In such cases, the threshold function is also given by a formula; but the tolerance, though again determined, is identified by human agents exercising ethical judgements. In that sense, it resembles the entropy ordering introduced in [17]. Indeed the threshold function is derived from the level functions used in [17] to define entropy orderings.

For non-artificial agents, like humans, we do not know whether all relevant observables can be mathematically determined. The opposing view is represented by followers and critics of the Hobbesian approach. The former argue that for a realistic LoA it is just a matter of time until science is able to model a human as an automaton, or state-transition system, with scientifically determined states and transition rules; the latter object that such a model is in principle impossible. Our approach is that when considering agents, thresholds are in general only partially quantifiable and usually determined by consensus.

### 3.3.1 Examples

Let us reconsider the examples from Section 2.6 from the viewpoint of morality.

The futuristic thermostat is morally charged since the LoA includes patients' well-being. It would be regarded as morally good if and only if its output maintains the actual patients' well-being within an agreed tolerance of their desired well-being. Thus, in this case a threshold function consists of the distance (in some finite-dimensional real space) between the actual patients' well-being and their desired well-being.

Since we value our email, a webbot is morally charged. In [17] its action was deemed to be morally bad (an example of artificial evil) if it incorrectly filters any messages: if either it filters messages it should let pass, or lets pass messages it should filter. Here we could use the same criterion to deem the webbot agent itself to be morally bad. However, in view of the continual adaptability offered by the bot, a more realistic criterion for moral good would be that at most a certain fixed percentage of incoming email be incorrectly filtered. In that case, the threshold function could consist of the number of incorrectly filtered messages.

The strategy-learning system MENACE simply learns to play noughts and crosses. With a little contrivance it could be morally charged as follows.

Suppose that something like MENACE is used to provide the game play in some computer game whose interface belies the simplicity of the underlying strategy and which invites the human player to pit his or her wit against the automated opponent. The software behaves unethically if and only if it loses a game after a sufficient learning period; for such behaviour would enable the human opponent to win too easily and might result in market failure of the game. That situation

may be formalised using thresholds by defining, for a system having initial state  $M$ ,  $T(M)$  to denote the number of games required after which the system never loses. Experience and necessity would lead us to set a bound,  $T_0(M)$ , on such performance: an ethical system would respect it whilst an unethical one would exceed it. Thus the function  $T_0(M)$  constitutes a threshold function in this case.

Organisations are nowadays expected to behave ethically; see for example [37]. In non-quantitative form, the values they must demonstrate include: equal opportunity, financial stability, good working and holiday conditions toward their employees; good service and value to their customers and shareholders; and honesty, integrity, reliability to other companies. This recent trend adds support to our proposal to treat organisations themselves as agents and thereby to require them to behave ethically, and provides an example of threshold which, at least currently, is not quantified.

## 4 Computer ethics

What does our view of moral agenthood contribute to the field of Computer Ethics (CE)? CE seeks to answer questions like: ‘What behaviour is acceptable in Cyberspace?’ and ‘Who is to be held morally accountable when unacceptable behaviour occurs?’. It is Cyberspace’s novelty that makes those questions, so well understood in standard ethics, of greatly innovative interest; and it is its growing ubiquity that makes them so pressing.

The first question requires, in particular, an answer to ‘What in Cyberspace has moral worth?’. The view that data have moral worth means that they need not be viewed as someone’s property in order for their unauthorised alteration to be ethically bad. This does not, of course, mean that any destruction of data is evil, any more than it would mean that any destruction of life (deemed to have moral worth) in the real world is automatically evil. It simply means that the ethics of altering data in Cyberspace must be considered. Evidently there are conditions under which deletion of data is morally advisable (garbage collection of redundant data, resulting in a more efficient system); and conditions when it is not (deletion of critical data not backed up). These common-sense observations fit well with our approach [13, 14].

We now turn to the second question and consider the consequences of our general answer to: ‘What in Cyberspace is morally accountable?’. Above we have made the case for the answer:

any agent that causes good or evil is morally accountable for it.

Just to recall, we have clarified that moral accountability is a necessary but insufficient condition for moral responsibility. An agent is morally accountable for  $x$  if the agent is the source of  $x$  and  $x$  is morally qualifiable (see definition O in Section 3.1). To be also morally responsible for  $x$ , the agent needs to show the right intentional states (recall the case of Oedipus).

Turning to our question, the traditional view is that only software engineers — human programmers — can be held morally accountable, possibly because only

humans can be held to exercise free will; and of course sometimes that view is perfectly appropriate.

Our more radical and extensive view is supported by the range of difficulties which in practice confronts the traditional view: software is largely constructed by teams; management decisions may be at least as important as programming decisions; requirements and specification documents play a large part in the resulting code; although the accuracy of code is dependent on those responsible for testing it, much software relies on ‘off the shelf’ components whose provenance and validity may be uncertain; moreover, working software is the result of maintenance over its lifetime and so not just of its originators. Many of these points are nicely made in [12]. Such complications may point to an organisation (perhaps itself an agent) being held accountable. But sometimes: automated tools are employed in construction of much software; the efficacy of software may depend on extra-functional features like its interface and even on system traffic; software running on a system can interact in unforeseeable ways; software may now be downloaded at the click of an icon in such a way that the user has no access to the code and its provenance with the resulting execution of anonymous software; software may be probabilistic [27]; adaptive [26]; or may be itself the result of a program (in the simplest case a compiler, but also genetic code [22]). All these matters pose insurmountable difficulties for the traditional and now rather outdated view that a human can be found accountable for certain kinds of software and even hardware [10, 29]. Fortunately, the view of this paper offers a solution — artificial agents are morally accountable as sources of good and evil — at the ‘cost’ of expanding the definition of morally-charged agent.

#### 4.1 Codes of ethics

Human morally-charged software engineers are bound by codes of ethics and undergo censure for ethical (and of course legal) violations. For consistency<sup>6</sup> our approach must make sense when that procedure is applied to morally accountable, AAs; does it?

The ACM Code of Ethics [1] contains 16 points guiding ethical behaviour (eight general and eight more specific; see Figure 2), six organisational leadership imperatives, and two (meta) points concerning compliance with the Code.

Of the first eight, all make sense for AAs; indeed they might be expected to form part of the specification of any morally-charged agent. Similarly for the second eight, with the exception of the penultimate point: ‘improve public understanding’. It is less clear how that might reasonably be expected of an arbitrary AA; but then it is also not clear that it is reasonable to expect it of a human software engineer. (It is to be observed, in passing, that wizards and similar programs with anthropomorphic interfaces — currently so popular — appear to make public use

---

<sup>6</sup>For an enlightening comparison consider that the Federation Internationale des Echecs (FIDE) rates all chess players according to the same Elo System, regardless of their human or artificial nature.

1	<b>General moral imperatives</b>
1.1	Contribute to society and human well-being
1.2	Avoid harm to others
1.3	Be honest and trustworthy
1.4	Be fair and take action not to discriminate
1.5	Honor property rights including copyrights and patents
1.6	Give proper credit for intellectual property
1.7	Respect the privacy of others
1.8	Honor confidentiality
2	<b>More specific professional responsibilities</b>
2.1	Strive to achieve the highest quality, effectiveness and dignity in both the process and products of professional work
2.2	Acquire and maintain professional competence
2.3	Know and respect existing laws pertaining to professional work
2.4	Accept and provide appropriate professional review
2.5	Give comprehensive and thorough evaluations of computer systems and their impacts, including analysis of possible risks
2.6	Honor contracts, agreements and assigned responsibilities
2.7	Improve public understanding of computing and its consequences
2.8	Access computing and communication resources only when authorised to do so

Figure 3: The principles guiding ethical behaviour in the ACM Code of Ethics.

easier; and such a requirement could be imposed on any AA; but that is scarcely the same as improving understanding.)

The final two points concerning compliance with the code (4.1: agreement to uphold and promote the code; 4.2: agreement that violation of the code is inconsistent with membership) make sense, though promotion does not appear to have been considered for current AAs any more than has the improvement of public understanding. The latter point presupposes some list of member agents from which agents found to be unethical would be struck.<sup>7</sup> This brings us to the censuring of AAs.

## 4.2 Censureship

Human moral agents who break accepted conventions are censured in various ways of which the main alternatives are: (a) mild social censure with the aim of changing and monitoring behaviour; (b) isolation, with similar aims; (c) death. What would be the consequences of our approach for artificial moral agents?

Preserving consistency between human and artificial moral agents, we are led to contemplate the following analogous steps for the censure of immoral artificial agents: (a) monitoring and modification (i.e. ‘maintenance’); (b) removal to a disconnected component of Cyberspace; (c) annihilation from Cyberspace (deletion without backup). Our insistence on dealing directly with an agent rather than seeking its ‘creator’ (a concept which we have claimed need be neither appropriate nor even well defined) has led to a nonstandard but perfectly workable conclusion. Indeed it turns out that such a categorisation is not very far from that used by the Norton Anti-Virus facility [28]. Though not adaptable at the obvious LoA, the facility is almost agent-like. It runs autonomously, polling web sites for anti-virus software which it applies to the files of the host computer. When it detects an infected file it offers several levels of censure: notification, repair, quarantine, deletion, with or without backup.

For humans, social organisations have had, over the centuries, to be formed for the enforcement of censureship (police, law courts, prisons, etc.). It may be that analogous organisations could sensibly be formed for AAs (it is perhaps unfortunate that this has a Sci-Fi ring to it [35]). Such social organisations became necessary with the increasing level of complexity of human interactions and the growing lack of ‘immediacy’. Perhaps that is the situation in which we are now beginning to find ourselves with the web; and perhaps it is time to consider agencies for the policing of AAs.

## 5 Conclusion

This paper may be read as an investigation into the extent to which ethics lies exclusively within the human domain.

---

<sup>7</sup>It is interesting to speculate on the mechanism by which that list is maintained. Perhaps by a human agent; perhaps by an AA composed of several people (a committee); or perhaps by a software agent.



Somewhere between 16 and 21 years after birth a human is deemed to be an autonomous legal entity — an adult — responsible for his or her actions. Yet an hour after birth that is only a potentiality. Indeed the law and society treat children quite differently from adults on the grounds that not they but their guardians, typically parents, are responsible for their actions. Animal behaviour varies in exhibiting intelligence and social responsibility between the childlike and the adult, on the human scale, so that on balance animals are accorded at best the legal status of children and a somewhat diminished ethical status (guide dogs, dolphins, ...). But there are exceptions. Some adults are deprived of their rights (criminals may not vote) on the grounds that they have demonstrated an inability to exercise responsible/ethical action. Some animals are held accountable for their actions and punished/destroyed if they err.

Into this arena we have placed for consideration other entities: certain kinds of organisation or system. Many examples have been given, but particularly interesting ones comprise corporate structures and digital/software system. Our aim has thus been to understand better the kinds of conditions under which an agent may be held morally accountable.

A natural and immediate answer could have been: such accountability lies entirely in the human domain; animals can sometimes appear to exhibit morally responsible behaviour, but lack the thing unique to humans which render humans (alone) morally responsible; end of story. In spite of its simplicity that answer is worryingly dogmatic. Surely more conceptual analysis is needed here: what has happened morally when a child is deemed to enter adulthood, or when an adult is deemed to have lost moral autonomy, or when an animal is deemed to hold it?

We have added artificial agents (corporate or digital, for example) to the arena. That has the advantage that all entities that populate the arena are analysed in non-anthropocentric terms; in other words, it has the advantage of offering a way to progress past the immediate and dogmatic answer mentioned above. That approach also renders our conclusions applicable to the accountability of certain kinds of organisation or digital system: pressing concerns from Business and Computer Ethics.

We have been able to make progress in the analysis of moral agenthood by using an important technique, the Method of Abstraction, designed to make rigorous the level of abstraction of the domain of discourse. Since we have considered entities from the world around us, whose properties are vital to our analysis and conclusions, it is essential that we have been able to be precise about the LoA at which those entities have been considered. We have seen that changing the LoA may well change our observation of their behaviour and hence change the conclusions we draw. This is not relativism and it has a venerable tradition in science and more recently in the humanities.

To address all the entities in the arena we have been forced to adopt terminology which applies equally to all the entities that populate it (from animals to organisations), but without prejudicing our conclusions. We have been forced to analyse behaviour in a non-anthropocentric manner. To do so we have used the notation of a state-transition system.

We have called the entities within this arena ‘agents’, characterised here abstractly in terms of a state-transition system, and have concentrated largely on artificial agents and the extent to which ethics and accountability applies to them. Whether or not an entity forms an agent depends on the LoA at which the entity is considered; there can be no absolute LoA-free form of identification. By abstracting that LoA, an entity may lose its agenthood by no longer satisfying the behaviour we associate with agents. However for most entities there is no LoA at which they can be considered an agent. Of course. Otherwise one might be reduced to the absurdity of considering the moral accountability of the magnetic strip that holds a knife to the kitchen wall. Instead, for comparison, our techniques address the far more interesting question [8]: ‘when HAL kills, who’s to blame?’ Indeed our techniques enable us to say that HAL is accountable — though not responsible — if it meets the conditions defining agenthood.

In Section 3.1 we have deferred discussion of a final objection to our approach until the conclusion. The time has come to honour that.

Our opponent can still raise a final objection: suppose you are right; does this enlargement of the class of moral agents bring any real advantage? It should be clear why the answer is clearly affirmative. Morality is usually predicated upon responsibility. The use of LoA and thresholds enables responsibility to be separated and formalised, and its part in morality to be fully clarified. The better grasp of what it means for someone or something to be a moral agent brings with it a number of substantial advantages. We can avoid anthropocentric and anthropomorphic attitudes towards agenthood and rely on an ethical outlook not necessarily based on punishment and reward but on moral agenthood, accountability and censure. We are less likely to assign responsibility at any cost, forced by the necessity to identify a human moral agent. We can liberate technological development of AAs from being bound by the standard limiting view [11]. We can stop the regress of looking for the *responsible* individual when something evil happens, since we are now ready to acknowledge that sometimes the moral source of evil or good can be different from an individual or group of humans. As a result we are able to escape the dichotomy ‘responsibility + moral agency = prescriptive action’ versus ‘no responsibility ergo no moral agency ergo no prescriptive action’. Promoting normative action is perfectly reasonable even when there is no responsibility but only moral accountability and the capacity for moral action.

All this does not mean that the concept of ‘responsibility’ is redundant. On the contrary, our previous analysis makes clear the need for further analysis of the concept of responsibility itself, when the latter refers to the ontological commitments of creators of new AAs and environments. As we have argued in [19, 14, 15] Information Ethics is an ethics addressed not just to ‘users’ of the world but also to demiurges who are ‘divinely’ responsible for its creation and well-being. It is an ethics of *creative stewardship*.

In the introduction we have warned about the lack of balance between the two classes of agents and patients brought about by deep forms of environmental ethics that are not accompanied by an equally ‘deep’ approach to agenthood. The position defended in this paper supports a better equilibrium between the

two classes  $A$  and  $P$ . It facilitates the discussion of the morality of agents not only in Cyberspace but also in the biosphere — where animals can be considered moral agents without their having to display free will, emotions or mental states [9, 30, 31] — and in what we have called contexts of ‘distributed morality’, where social and legal agents can now qualify as moral agents. The great advantage is a better grasp of the moral discourse in non-human contexts [32]. The only ‘cost’ of a ‘mind-less morality’ approach is the extension of the class of agents and moral agents to embrace AAs. It is a cost that is increasingly worth paying the more we move towards an advanced information society.

## Acknowledgements

Versions of this paper have been presented by L. Floridi at the seminar for graduate studies of the University of Bari and at CEPE 2001, Lancaster, UK. The authors are grateful for the feedback received, for the comments made by the referees who accepted the paper for the Lancaster conference and to the referees of *Minds and Machines* for their improvements.

Luciano Floridi wishes to acknowledge the financial support of the University of Bari. Finally, this paper is dedicated to *Martina*.

## References

- [1] ACM, Code of Ethics. <http://www.acm.org>.
- [2] C. Allen, G. Varner and J. Zinser, Prolegomena to any future artificial moral agent. *Journal of Experimental and Theoretical Artificial Intelligence*, **12**:251–61, 2000.
- [3] A. Arnold, *Finite Transition Systems*. Prentice-Hall International Series in Computer Science, 1994.
- [4] J. Badham (director), *War Games*, 1983.
- [5] M. A. Bedau, The nature of life. In M. A. Boden (editor), *The Philosophy of Life*, Oxford University Press, 332–357, 1996.
- [6] E. Cassirer, *Substance and Function and Einstein’s Theory of Relativity*. Dover publications edition, New York, 1953.
- [7] P. Danielson, *Artificial Morality: Virtuous Robots for Virtual Games*. Routledge, NY, 1992.
- [8] D. Dennet, When HAL kills, who’s to blame? In D. Stork (editor), *HAL’s Legacy: 2001’s computer as dream and reality*. MIT Press, Cambridge MA, 351–365, 1997.

- [9] B. A. Dixon, Response: Evil and the Moral Agency of Animals. *Between the Species*, **11**(1-2):38–40, 1995.
- [10] Machines with minds of their own, *The Economist*, 22 March, 2001.  
<http://www.economist.com> .
- [11] Beyond cruise control, *The Economist*, 21 June, 2001.  
<http://www.economist.com> .
- [12] R. G. Epstein, *The Case of the Killer Robot*. John Wiley and Sons, Inc., 1997.
- [13] L. Floridi, Information Ethics: on the theoretical foundations of computer ethics. *Ethics and Information Technology*, **1**(1):37–56, 1999.  
Preprint from <http://www.wolfson.ox.ac.uk/floridi/papers.htm> .
- [14] L. Floridi, On the intrinsic value of information objects and the infosphere. *Ethics and Information Technology*, **4**(4):287–304, 2003.  
Preprint from <http://www.wolfson.ox.ac.uk/floridi/papers.htm> .
- [15] L. Floridi, Information Ethics: an environmental approach to the digital divide. UNESCO World Commission on the Ethics of Scientific Knowledge and Technology (COMEST), First Meeting of the Sub-Commission on the Ethics of the Information Society (UNESCO, Paris, June 18-19, 2001).  
Preprint from <http://www.wolfson.ox.ac.uk/floridi/papers.htm> .
- [16] L. Floridi, Ethics in the infosphere. *The Philosophers' Magazine*, **6**:18–19, 2001. Preprint from <http://www.wolfson.ox.ac.uk/floridi/papers.htm> .
- [17] L. Floridi and J. W. Sanders, Artificial evil and the foundation of computer ethics. *Ethics and Information Technology*, **3**(1):55–66, 2001. Preprint from <http://www.wolfson.ox.ac.uk/floridi/papers.htm> .
- [18] L. Floridi and J. W. Sanders, The Method of Abstraction. In M. Negrotti (editor), *The Yearbook of the Artificial*. Issue II, Peter Lang, Bern, 2003.  
Preprint from <http://www.wolfson.ox.ac.uk/floridi/papers.htm> .
- [19] L. Floridi and J. W. Sanders, Internet Ethics: the Constructionist Values of Homo Poieticus. In R. Cavalier (editor), *The Impact of the Internet on Our Moral Lives*. SUNY, Fall 2003.
- [20] S. Franklin and A. Graesser, Is it an agent, or just a program? A taxonomy for autonomous agents. *Proceedings of the Third International Workshop on Agent Theories, Architectures, and Languages*, Springer-Verlag, 1996.  
[www.msci.memphis.edu/franklin/AgentProg.html](http://www.msci.memphis.edu/franklin/AgentProg.html) .
- [21] J. Gips, Towards the ethical robot. In K. Ford, C. Glymour and P. Hayes (editors), *Android Epistemology*. MIT Press, Cambridge MA, 243–252, 1995.

- [22] D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley, Reading Mass., 1989.
- [23] P. Kerr, *The Grid*. New York, Warner Books, 1996.
- [24] D. Michie, Trial and error. In A. Garratt (editor), *Penguin Science Surveys*. Harmondsworth, Penguin, 129–145, 1961.
- [25] *Minds and Machines*, **11**(1), The Turing test: past, present and future. Guest editor J. H. Moore, February 2001.
- [26] T. M. Mitchell, *Machine Learning*. McGraw Hill, 1997.
- [27] R. Motwani and P. Raghavan, *Randomized Algorithms*. Cambridge, Cambridge University Press, 1995.
- [28] *Norton AntiVirus 2003*. Version 8.07.17C. Symantec Corporation, copyright 2003.
- [29] I. Page and W. Luk, Compiling occam into Field-Programmable gate arrays. <ftp://ftp.comlab.ox.ac.uk/pub/Documents/techpapers/Ian.Page/abs-hwcomp.l>.
- [30] R. Rosenfeld, Can animals be evil?: Kekes' character-morality, the hard reaction to evil, and animals. *Between the Species*, **11**(1-2):33–38, 1995.
- [31] R. Rosenfeld, Reply. *Between the Species*, **11**(1-2):40–41, 1995.
- [32] M. Rowlands, *The Environmental Crisis—Understanding the Value of Nature*. Palgrave, London-Basingstoke, 2000.
- [33] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Introduction*, second edition. Prentice-Hall International, 2003.
- [34] I. Sample, SmartPaint, *New Scientist*, <http://www.globaltechnoscan.com/16May-22May01/paint.htm> .
- [35] R. Scott (director), *Bladerunner, The Director's Cut*, 1982/1991.
- [36] A. M. Turing, Computing Machinery and Intelligence, *Mind*, **59**(236):433–60, 1950.
- [37] The Working Values Group, Ltd., <http://www.workingvalues.com> .