Artificial Evil and the Foundation of Computer Ethics

L. L. Floridi
University of Oxford
Wolfson College, Oxford, OX2 6UD, UK, tel. 01865-274137
fax 01865275125, luciano.floridi@philosophy.oxford.ac.uk

J. W. Sanders
University of Oxford
Programming Research Group, OUCL, Wolfson Building,
Parks Road, Oxford, OX1 3QD, tel. 01865-273840, fax 01865-273839, jeff@comlab.ox.ac.uk

Abstract

Moral reasoning traditionally distinguishes two types of evil: moral (ME) and natural (NE). The standard view is that ME is the product of human agency and so includes phenomena such as war, torture and psychological cruelty; that NE is the product of nonhuman agency, and so includes natural disasters such as earthquakes, floods, disease and famine; and finally, that more complex cases are appropriately analysed as a combination of ME and NE. Recently, as a result of developments in autonomous agents in cyberspace, a new class of interesting and important examples of hybrid evil has come to light. In this paper it is called artificial evil (AE) and a case is made for considering it to complement ME and NE to produce a more adequate taxonomy. By isolating the features that have led to the appearance of AE, cyberspace is characterised as a self-contained environment that forms the essential component in any foundation of the emerging field of Computer Ethics (CE). It is argued that this goes some way towards providing a methodological explanation of why cyberspace is central to so many of CE's concerns; and it is shown how notions of good and evil can be formulated in cyberspace. Of considerable interest is how the propensity for an agent's action to be morally good or evil can be determined even in the absence of sentient participants and thus allows artificial agents not only to perpetrate evil (and for that matter good) but conversely to `receive' or `suffer from' it. The thesis defended is that the notion of entropy structure, which encapsulates human value judgement in a formal mathematical definition, is sufficient to achieve this purpose and, moreover, that the concept of AE can be determined formally, by mathematical methods. A consequence of this approach is that the debate on whether CE should be considered unique, and hence developed as a Macroethics, may be viewed, constructively, in an alternative manner. The case is made that whilst CE issues are not uncontroversially unique, they are sufficiently novel to render inadequate the approach of standard Macroethics such as Utilitarianism and Deontologism and hence to prompt the search for a robust ethical theory that can deal with them successfully. The name Information Ethics (IE) is proposed for that theory. It is argued that the uniqueness of IE is justified by its being non-biologically biased and patient-oriented: IE is an Environmental Macroethics based on the concept of information entity rather than life. It follows that the novelty of CE issues such as AE can be appreciated properly because IE provides a new perspective (though not vice versa). In light of the discussion provided in this paper, it is concluded that Computer Ethics is worthy of independent study because it requires its own application-specific knowledge and is capable of supporting a methodological foundation, Information Ethics.

Keywords

Agent, Artificial agent, Artificial evil, Cyberspace, Entropy, Environmentalism, Information ethics, Moral evil, Natural evil, Nonsubstantialism, Patient, Theodicean problem, Uniqueness debate

INTRODUCTION: THE NATURE OF EVIL

Evil is the most comprehensive expression of ethical disapproval and, as the reverse of moral good, a key concept in any axiology. Of the many conceptual clarifications available in the literature, three need to be recalled here to provide the essential background of the paper.[1]

Any action, whether morally loaded or not, has the logical structure of a variably interactive process, which relates a set of one or more sources (depending on whether we are working within a multiagent context), the agent *a*, which initiates the process, with a set of one or more destinations, the patient *p*, which reacts to the process.[2] To clarify the nature of *a* and *p* it is useful to borrow the concept of information 'object' from the object-oriented analysis paradigm (OOA).[3] The agent and the patient are discrete, self-contained, encapsulated[4] packages containing:

---

[1] The model follows but does not pressupose knowledge of Floridi (1998).

[2] The terms 'agent' and 'patient' are standard in Ethics and therefore will be maintained in this paper, however, it is essential to stress the interactive nature of the process and hence the fact that the patient is hardly ever a passive receiver of an action. A better way to qualify the patient in connection with the agent would be to refer to it as the 'reagent'.

[3] The article follows the standard terminology and the conceptual apparatus provided by James Rumbaugh et al. (1991).

[4] Encapsulation or information hiding is the technique of keeping together data structures and the methods (class-implemented operations), which act on them in such a way that the package's internal structure can be accessed only by means of the approved package routines. External aspects of an object, which are accessible to other objects, are thus separated from the internal implementation details of the object itself, which remain hidden from other objects.

- the appropriate data structures, which constitute the nature of the entity in question (state of the object, its unique identity, and attributes)
- a collection of operations, functions or procedures (methods[5]), which are activated (invoked) by various interactions or stimuli, namely messages (in this essay 'actions' is used with this technical meaning) received from other objects (message passing) or changes within itself, and correspondingly define (implement) how the object behaves or reacts to them.

In Leibnizian and more metaphysical terms, an object is a sufficiently permanent (a continuant) information monad, a description of the ultimate primal component of all beings. The moral action itself can be constructed as an information process, i.e. a series of messages (M), initiated by an agent *a*, that brings about a transformation of states directly affecting a patient *p*, which may interactively respond to M with changes and/or other messages, depending on how M is interpreted by *p*'s methods, that is $\exists a\ \exists p$ M (a, p).

When discussing the nature of evil, the following two clarifications are usually accepted as standard:

1) 'evil' is a second order predicate that qualifies primarily M.

Only actions are primarily evil. Sources of evil (agents and their intentional states) are evil in a derivative and often unclear sense: intentional states are evil if they (can) lead to evil actions, and agents are evil if the *preponderance* of their intentional states or actions is evil. The domain of intentional states or actions, however, is probably infinite, so the concept of 'preponderance' is based either on a limit in time and scope (*a* is evil between time t_1 and time t_n and as far as intentional states or actions y are concerned), or on a inductive/probabilistic projection (*a* is such that *a*'s future intentional states or actions are more likely to be evil than good). Obvious difficulties in both approaches reinforce the view that an agent is evil only derivatively;

2) the interpretation of *a* ranges over the domain of all agents, both human and nonhuman.

Evil actions are the result of human or nonhuman agency (e.g. natural disasters). The former is known as moral evil (ME) and it implies autonomy and responsibility, and hence a sufficient degree of information, freedom and intentionality. The latter is known as natural evil (NE). It is usually defined negatively, as any evil that arises independently of human intervention, in terms of prevention, defusing or control. A third clarification, although rather common, is less uncontroversial:

3) the positive sense in which an action is evil (*a*'s intentional harming) is parasitic on the privative sense in which its effect is evil (decrease in *p*'s welfare).

Contrary to 'responsibility'—an agent-oriented concept that works as a robust theoretical 'attractor', in the sense that standard Macroethics (e.g.

[5] A method is a particular implementation of an operation, i.e. an action or transformation that an object performs or is subject to by a certain class. An operation may be implemented by more than one method.

Consequentialism or Deontologism) tend to concentrate on it for the purpose of moral evaluations of the agent—'evil' is a perspicuously patient-oriented concept. Actions are ontologically dependent on agents for their implementation (evil as cause), but are evaluated as evil only in view of the degree of severe and unnecessary harm that they may cause to their patients (evil as effect). Hence, whether an action is evil can be decided only on the basis of a clear understanding of the nature and future development of the interacting patient.

Since an action is evil if and only if it harms or tends to harm its patient, evil, understood as the harmful effect that could be suffered by the interacting patient, is properly analysed only in terms of possible corruption, decrease, deprivation or limitation of *p*'s welfare, where the latter can be defined in terms of the object's appropriate data structures and methods. This is the classic, 'privative' sense in which evil is parasitic on the good and does not exist independently of the latter (evil as *privationem boni*). In view of this further qualification, and in order to avoid any terminological bias, it is better to avoid using the term 'harm'—a zoocentric, not even biocentric word, which implicitly leads to the interpretation of *p* as a sentient being with a nervous system—in favour of 'damage', an ontocentric, more neutral term, with 'annihilation' as the level of most severe damage.

According to the OOA approach endorsed in this paper, messages are processes that affect objects either positively or negatively. Positive messages respect or enhance *p*'s welfare; negative messages do not respect or damage *p*'s welfare. Evil actions are a subclass of negative messages, those that do not merely fail to respect *p* but (can) damage it (for an axiological analysis see Floridi (1998)). The following definition attempts to capture the clarifications introduced so far:

> (E) Evil action = one or more negative messages, initiated by *a*, that brings about a transformation of states that (can) damage *p*'s welfare severely and unnecessarily; or more briefly, any patient-unfriendly message.

(E) excludes both victimless and anonymous evil: an action is (potentially) evil only if there is (could be) a damaged patient, and there is no evil action without a damaging source, even if, in a multiagent and distributed context, this may be sufficiently vague or complex to escape clear identification (however, we shall argue below that this does not imply that evil cannot be gratuitous). In fact, because standard Macroethics tend to privilege agent-centred analyses, they usually concentrate on evil actions *a parte agentis*, by *presupposing* the presence of an agent and qualifying the agent's actions as evil, at least hypothetically or counterfactually. On the basis of these clarifications, it is now possible to develop five main theses:

1) IE (Information Ethics) can defend a deflationary approach to the existence of evil
2) ICT (information and communication technology) modifies the interpretation of some evils, transforming them from natural into moral
3) ICT extends the class of agents, generating a new form of artificial evil (AE)

4) ICT extends the class of patients, promoting a new understanding of evil as entropy
5) (1)-(4) contribute to clarify the uniqueness debate in computer ethics.

NONSUBSTANTIALISM: A DEFLATIONARY APPROACH TO THE EXISTENCE OF EVIL

The classic distinction ME vs. NE is sufficiently intuitive but may also be misleading. Human beings may act as Natural Agents (e.g. unaware and healthy carries of a disease) and natural evil may be the mere means of moral evil (e.g. through morally blameworthy negligence). But above all, the terminology may be misleading because it is the result of the application of first ('moral', 'natural') to a second order ('evil') predicate, which paves the way to a questionable hypostasization of evil. This substantialism reifies evil as if it were a 'token' transmitted by M from *a* to *p*, an oversimplified 'communication' model that is implausible, since *a*'s messages appear to generate negative states only by interacting with *p*'s methods, and do not seem either to be evil independently of them, or to bear and transfer some pre-packaged, perceivable evil by themselves.

To avoid the hypostasization of evil, a nonsubstantialist position (i) must defend a deflationary interpretation of evil's existence without (ii) accepting the equally implausible alternative represented by revisionism, i.e. the negation of the existence of evil *tout court*, which may rely, for example, on an epistemological interpretation for its elimination (evil as appearance). This can be achieved by (iii) accepting the derivative and privative senses of evil (evil as absence of good) to clarify that 'there is no evil' means that (iv) only actions, and not objects in themselves, can be qualified as primarily evil, and that (v) what type of evil x is should not be decided on the basis of the nature of the agent initiating x, since ME and NE do not refer to some special classes of entities, which would be intrinsically evil, nor to some special classes of actions *per se*, but they are only shortcuts to refer to a three-place relation between types of agents, actions and patients' welfare, hence to a specific, context-determined interpretation of the triple <*a*, *M*, *p*>.

The points made in (i)-(v) seem perfectly reasonable. Unfortunately, especially in ancient philosophy,[6] they have often been overinterpreted as a proof for the non-existence of evil. This because nonsubstantialism has been equated with revisionism through an ontology of things, i.e. the assumption that either x is a substance, something, or x does not exist. But since evil is so widespread in the world, any argument that attempts to deny its existence is doomed to be rejected as sophistic. So revisionism is hardly defensible and, through the equation, the consequence has been that the presence of evil in the world has often been taken as definitive evidence against nonsubstantialism as well and, even more generally, as a final criticism of any theory based on (1)-(3) and (i)-(v). It should be obvious, however, that this conclusion is not inevitable: nonsubstantialism is deflationary

[6] Especially in the Platonic tradition, see Plato, Proclus, Plotin, Augustine, but also Aristotle and in modern times Leibniz and Spinoza.

but not revisionist, and it is perfectly reasonable to defend the former position by rejecting the implicit reliance on an ontology of things. Actions-messages and objects' states, as defined in the OOA paradigm, do not have a lower ontological status than objects themselves. Evil exists not absolutely, *per se*, but in terms of damaging actions and damaged objects. The fact that its existence is parasitic does not mean that it is fictitious. On the contrary, in an ontology that treats interactions, methods (operations, functions and procedures) and states on the same level as objects and their attributes, evil could not be any more real. Once an ontology of things is replaced by a more adequate OOA ontology, it becomes possible to have all the benefits of talking about evil without the ontological costs of a substantialist hypostasization. This is the approach defended by IE (Floridi (1998)).

THE EVOLUTION OF EVIL AND THE THEODICEAN PROBLEM

Natural evil has been introduced as any evil that arises through no human action, either positive or negative: NE is whatever evil human beings do not initiate and cannot prevent, defuse or control.[7] Since the discussion on the nature of evil has been largely monopolised by the theodicean debate (whether it is possible to reconcile the existence of God and the presence of evil), contemporary Macroethics seem to have failed to notice that this definition entails the possibility of a diachronic transformation of what may count as NE because of the increasing power of design, configuration, prevision and control over reality offered by science and technology (sci-tech), including ICT. If a negative definition of NE, in terms of ¬ ME, is not only inevitable but also adequate, the more powerful a society becomes, in terms of its sci-tech, the more its members are responsible for what is within their power to influence. Past generations, when confronted by natural disasters like famine or flood, had little choice but to put up with their evil effects. Nowadays, most of the ten plagues of Egypt would be considered moral rather than natural evils because of human negligence.[8] A clear sign of how much the world has changed is that people expect human solutions for virtually any natural evil, even when this is well beyond the scientific and technological capacities of present times. Whenever a natural disaster occurs, the first reaction has become to check whether anyone is responsible for an action that might have initiated or prevented its evil effects.

The human-independent nature of NE and the powerfulness of science and technology, especially ICT, with its computational capacities to forecast future

[7] It is probably useful to conceive different kinds of NE as placed on a scale, from the not-humanly-initiated and not-preventable earthquake (only the evil effects of it can be a matter of human responsibility) to the not-humanly-initiated but humanly preventable plague to the humanly initiated and preventable mistake (human agents as natural causes).

[8] It may be interesting to stress that in the Old Testament the plagues have mainly an ontological value, as signs of total control and power over reality, rather than ethical. Several times the Pharaoh's magicians are summoned to deal with the extraordinary phenomena, but the point is always whether they may be able to achieve the same effects 'by their secret arts'—hence showing that there is either no divine intervention or equal divine support on the Egyptian side—not whether they can undo or solve the difficulties caused by the specific plague. They loose the 'ontic game' when 'the magicians tried by their secret arts to bring forth gnats, but they could not'.

events, determine a peculiar phenomenon of constant erosion of NE in favour of an expansion of ME. If anyone were to die from smallpox in the future this would certainly be a matter of ME, no longer NE. Witchcraft in theory and sci-tech in practice share the responsibility of transforming NE into ME and this is why their masters look morally suspicious. It is an erosion that is inevitable, insofar as science and technology can constantly increase human power over nature. It may also seem unidirectional: at first, it may appear that the only transformation brought about by the evolution of sci-tech is a simplification in the nature of evil. However, the introduction of the concept of artificial evil (AE) provides a corrective to this view (see next section). If, for the present purpose, it is simply assumed that, at least in theory, all NE can become ME but not vice versa, it is obvious that this provides an interesting approach to the classic theodicean problem of evil. The theist may need to explain only the presence of ME despite the fact that God is omniscient, omnipotent, and all-good, and it is known that a theodicy based on the responsibility that comes with freedom is more defensible, especially if connected with a nonsubstantialist approach to the existence of evil. In a utopian world, the occurrence of evil may be just a matter of human misbehaviour. What matters here, of course, is not to solve the theodicean puzzle, but to realise how ICT is contributing to make humanity increasingly accountable, morally speaking, for the way the world is.

ARTIFICIAL EVIL

More and more often, especially in advanced societies, people are confronted by visible and salient evils that are neither simply natural nor immediately moral: an innocent dies because the ambulance was delayed by the traffic; a computer-based monitor 'reboots' in the middle of surgery because its software is not fully compatible with other programs also in use, with the result that the patient is at increased risk during the reboot period. The examples could easily be multiplied. What kind of evils are these? 'Bad luck' and 'technical incident' are simply admissions of ignorance. Conceptually, they indicate the shortcomings of the ME vs. NE dichotomy. The problem is that the latter was formulated at a time when the primary concern was anthropocentric, human-agent-oriented and the main issue addressed was that of human and divine responsibility. Strictly speaking, the difference between human and Natural Agents is not that the former are not natural, but that they are autonomous, i.e. they can regulate themselves. So, following the standard approach, the correct taxonomy turns out to be a four-place scheme: forms of agency are either natural or artificial (non-natural) and either autonomous or heteronomous (non-autonomous). Although this is not the context to provide a detailed analysis of an agent, the following definition is sufficiently adequate to clarify these four basic forms of agency:

A) Agent = a system, situated within and a part of an environment, which initiate a transformation produces an effect or exerts power on it over time, as contrasted with a system that is (at least initially) acted on or responds to it (patient).

A Natural Agent is an agent that has its ontological basis in the normal constitution of reality and conforms to its course, independently of human beings' intervention.

Conversely, an Artificial Agent is an agent that has its ontological basis in a constructed reality and depends, at least for its initial appearance, on human beings' intervention. An autonomous agent is an agent that has some kind of control over its states and actions, senses its environment, responds in a timely fashion to changes that occur in it and interacts with it, over time, in pursuit of its own goals, without the direct intervention of other agents. And a heteronomous agent is simply an agent that is not autonomous. Given these clarifications, the taxonomy is:

Agent	**N**atural	**A**rtificial
Autonomous	**NAA**	**AAA**
Heteronomous	**NHA**	**AHA**

NAA = natural and autonomous agent, e.g. a person, an animal, an angel, a god, an extraterrestrial.
NHA = natural and heteronomous agent, e.g. a flood, an earthquake, a nuclear fission.
AAA = artificial and autonomous agent, e.g. a webbot, an expert system, a software virus, a robot.
AHA = artificial and heteronomous agent, e.g. traffic, inflation, pollution.
ME is any evil produced by a *responsible* NAA; NE is any evil produced by NHA and by any NAA that may not be held directly responsible for it; AE is any evil produced by either AAA or AHA. The question now is: is AE always reducible to (perhaps a combination of) NE or ME?

It is clear that AE is not reducible to NE because of the nature of the agent involved, whose existence depends on human creative ingenuity. But this leads precisely to the main objection against the presence of AE, namely that any AE is really just a ME under a different name. Human creators are morally accountable for whatever evil may be caused by their Artificial Agents, as mere means or intermediaries of human activities (indirect responsibility). The objection of indirect responsibility is based on an analogy with the theodicean problem and is partly justified. In the same way as a divine creator can be blamed for NE, so a human creator can be blamed for AE.

A first reply consists in remarking that even in a theodicean context one still speaks of 'natural' not of 'divine' evils, thus indicating the nature of the agent, not of the morally responsible source. But this, admittedly, would be a weak retort, for it misses the important ethical point: if NE is 'real' then this causes a problem precisely because it is reducible to 'divine' evil and, *mutatis mutandis*, this could apply to the relation between AE and ME. AE could be just the result of carrying on morally wrong actions by other means.

A better reply consists in clarifying the differences between the two cases. On the one hand, AE may be caused by AHA whose behaviour depends immediately and directly on human behaviour. In this case, the reduction AE = ME is reasonable. AHA are just an extension of their human creators, like tools, because the latter are both the ontological and the nomological source of the formers' behaviour. Human beings can be taken to be directly accountable for the artificial evil involved, e.g. pollution. On the other hand, AAA, whose behaviour is nomologically independent of human intervention, may cause AE. In this case, the interpretative model is not God vs. created universe, but parents vs. children. Although it is conceivable that the evil caused by the children may be partly blamed on their parents, it is also true that, normally, the sins of the sons will not be passed on to the fathers. Indirect responsibility can only be forward, not backward, as it were. Things are in fact even more complicated than this. For the 'creatures' are more like pets, agents whose scope of action is very wide, which can cause all imaginable evils, but which cannot be taken *morally* responsible for their behaviour, owing to their insufficient degree of intentionality, intelligence and freedom. It turns out that, like in a universe without God, in cyberspace evil may be utterly gratuitous: there may be evil actions without any causing agent being *morally* blameable for them. Digital Artificial Agents are becoming sufficiently autonomous to pre-empty the possibility that their creators may be nomologically in charge of, and hence morally accountable for their misbehaviour. And we are still dealing with a generation of agents fairly simple, predictable and controllable. The phenomenon of potential artificial evil will become even more obvious as self-produced generations of AAA evolve.

Of course there is no IT-heodicean problem because the creators, in this case, are fallible, only partly knowledgeable, possibly malevolent and may work at cross-purposes, so there is no need to explain how the presence of humanity may be compatible with the presence of AE. Unfortunately, like Platonic demiurges, fallible creators much less powerful than the Christian God, we may not be able to construct truly intelligent AAA, but we can certainly endow them with plenty of autonomy and freedom, and it is in this lack of balance that the risk lie. It is clear that something similar to Asimov's Laws of Robotics will need to be enforced for the digital environment (the infosphere) to be kept safe. Sci-tech transforms natural into moral evil but at the same time creates a new form of evil, AE. In a dystopian world like the one envisaged in the film *The Matrix*, there could be just AE and ME.

EXTENDING THE CLASS OF PATIENTS OF ARTIFICIAL EVIL

In the previous section we have made the case for an Artificial Agent to be the source of an evil action. To contrast that case with the standard one, in which evil applies to the actions of Natural Agents, let us call that position *Weak Artificial Evil* (WAE) (cf. weak AI, Searle 1980). *Strong Artificial Evil* (SAE) is the position that an Artificial Agent can be the patient (or reagent, recall the interactive nature of the action-relation between agent and patient) of Artificial Evil. In this section we revisit the previous argument and make the case for SAE.

SAE has been prefigured by the *deep ecology* of Environmental Ethics (Zimmerman 1993) in which the state of inanimate objects is taken into account when considering the consequences of an action (e.g. how is building a certain freeway going to impinge on the rockface in its path). However, in the form of SAE the concept can be taken further, due largely to the characteristic properties of cyberspace, i.e. the (eco)system of information acted on by digital agents. The information is stored as bits, but encompasses vast tracts of data in the form of databases, files, records and online archives. The agents are programs and so include operating systems and applications software. Cyberspace is spanned by the Internet, which provides the vacuous but connected space; it is populated by all that data and programs and is lent geometrical presence by the web. It is to be emphasised that it is not helpful, for present purposes and despite its name, to conceive of cyberspace spatially. The rapid search and communications that are part of the web ensure that only addresses matter. Indeed, the features of importance to us here are:

a) spatiality: completeness of the network (any site is available from any other: point-to-point connectivity); homogeneity (standardised addresses); robustness against failure (Cartesian multiplicity of links);
b) democracy: nonhierarchical; not policed; free where possible; user extensible;
c) real-time: fast synchronous access to sites and fast asynchronous email communication; high bandwidth;
d) digitised: standardised digital storage and communications (both interpreted consistently throughout cyberspace).

Features (a)-(d) seem to characterise interactions in cyberspace. For example ecommerce exploits (a), (b), (c); downloading free music exploits (b), (d).

The frontier of cyberspace is the human/machine interface; thus we regard humans as lying outside cyberspace. In his famous Test (Turing 1950), Turing posited a keyboard/screen interface to blanket human and computer. Half a century later, that very interface has become part of our everyday reality. Helped perhaps by the ubiquitous television and the part it has played in informing and entertaining us, we are coming to rely on that interface for communication (email), information (sites), business (ecommerce) and entertainment (computer games). The all-pervading nature of cyberspace seems at present to depend partly on the extent to which we accept its interface as integral to our reality; indeed we have begun to accept the virtual as reality. What matters is not so much moving bits instead of atoms—this is an outdated, communication-based interpretation of the information society that owes too much to mass-media sociology—as the far more radical fact that the very essence and fabric of reality is changing. The information society is better seen as a neo-manufacturing society in which raw materials and energy have been superseded by the new digital gold. Not just communication and transactions then, but the creation, design and management of information are the keys to its proper understanding.

Cyberspace supports a variety of agents: from routine service software (like communications protocols) through less routine applications packages (like cybersitters, webbots) to applets downloadable from remote web sites. The latter highlight a shift in the burden of responsibility of software engineers. Formerly, (and still, of course, in the bulk of situations today) there was a contract between software engineer and user; the software engineer was responsible for the performance of the software, defensible if necessary at law. That model suited the context in which computers, or local-area networks, were isolated from others, except by physical media (disks, CDROMs, etc). In the new model, promoted by cyberspace, there is no 'point of sale', since a program may be downloaded at one of a sequence of mouse clicks, with no clear responsibility or even specification attending its acquisition. So seamless is the interface that the user may not even be aware that a program has been downloaded and executed locally: (b), (c).

The autonomy (and hence seamlessness) of that interaction is further reinforced by Artificial Agents which employ randomisation in making decisions (the giver of a coin can hardly be held responsible for decisions made on the basis of tossing it, even if the coin is sold as a binary-decision-making mechanism); and Artificial Agents which are able to adapt their behaviour on the basis of experience (in only an indirect sense were the programmers of Deep Blue responsible for its win, since it 'learnt' by being exposed to volumes of games, (King 1997)).[9] Given the presence of such agents, and the tendency towards further autonomy, the only reasonable view seems to be that misfortune resulting from such programs is evil for which neither human nor nature is directly responsible. Such a situation does not appear in the physical world inhabited by mechanical artifacts because their physical presence renders such machines, and their behaviour, traceable to their origins. Were they autonomous and able to transform and adapt, in the way programs can, such machines would provide an analogous example of AE; but so far they seem to be no more than instruments of science fiction.[10]

Cyberspace and its interface support actions that may originate from humans (email from a colleague) or Artificial Agents (messages from a word processor or directives from a webbot). The claim is not that current software has passed the Turing Test. It is simply that, with the types of software mentioned above, there is scope for evil that lies beyond the responsibility of human beings or nature.

Our region of cyberspace is in general changed as a result of the autonomous execution of Artificial Agents: decisions are delegated to routine procedures, data are altered, settings changed and programs subsequently behave differently. Artificial Patients in cyberspace thus 'respond' or 'react', often interactively, to actions. Some actions seem benign: the *easter eggs* cuckoo-ed inside Macintosh and Palm software (Pogue 1999, p. 36) constitute such examples. It seems equally clear that certain actions on Artificial Patients are evil: viruses and the action of certain webbots, for example. But the case for an Artificial Agent being the recipient of evil (and in particular, Artificial Evil) depends on our being able to make the case for determining when the preponderance of

---

[9] Mitchell (1997) provides the following examples of adaptive software: 'data-mining programs that learn to detect fraudulent credit-card transactions, to information-filtering programs that learn users' reading preferences, to autonomous vehicles that learn to drive on public highways.'

[10] For mechanisms that adapt to terrain see http://www.parc.xerox.com/modrobots. For statistically adaptive reconfigurable logic arrays, see http://jisp.cs.nyu.edu/RWC/rwcp/activities/achievements/AD/nec/eng/home-e.html. In fiction adaptive robots occur in the work of James P Hogan (e.g. `Two faces of Tomorrow' (1979) in which a semi-intelligent system controls a production line as part of a space station and, under pressure of attack, designs and produces different kinds of robot) and the popular film *Terminator 2* (in which the shape-shifting cyborg, T-1000 is sent back from the future to kill John Connor before he can grow up to lead the resistance).

consequences—as far as the patient goes—are bad. For that, we rely on the digital nature of cyberspace and employ the notion of entropy. We summarise an argument begun in (Floridi 1998) and developed in (Floridi and Sanders 1999).

First, we observe that an action in cyberspace is not uncontroversially bad or good; some value judgement is required to evaluate its moral worth. Thus it is a matter of judgement and context whether we regard as good or bad the effect of running a program: it might delete useful data (as might a virus) and so be judged bad, or it might perform useful garbage collection by removing inaccessible data. In (Floridi and Sanders 1999) we have made the case for *entropy structures* as a means of evaluating an action in cyberspace that combines judgements about desirable features of cyberspace with its discrete, and hence unambiguously definable, nature. An entropy structure is an ordering on cyberspace defined to capture the notion of a bad state change. The state-after is worse than the state-before. The state S of cyberspace consists of the values of all data, including software. A bad action changes state S_1 into S_2, where S_2 is greater in the entropy ordering; a benign action decreases the entropy ordering. By (d) the effect of any action is characterised, as a state transformer, mathematically by the relationship (a predicate) between the state-before, the input and output, and the state-after (in the example above, state is partitioned into used and unused store and the action converts some used store into unused store). It is then a matter of proof or counterexample whether an action is bad (none of its transitions yields an after-state which is greater in the entropy ordering than its before-state) or evil (there is a before-state and a transition in which the after-state is greater in the entropy ordering). Furthermore, the formalism can be used to determine when one action is more, or less, evil than another. The increase of entropy has been chosen, of course, to match the standard view from thermodynamics. However, in that setting no judgement is required since any increase, leading as it does to an increase in global randomness, is deemed bad (for formal definitions, examples and further discussion see Floridi and Sanders 1999). In summary, it is reasonable to permit an Artificial Agent to be the patient of evil and thus to have a moral standing. We conclude that the interpretation of the relational and interactive structure, symbolised by the triple <agent, action, patient>, is one of the central component of any Information Ethics.

THE UNIQUENESS DEBATE

The informative 'uniqueness' debate (Johnson 1999 and Maner 1999) has aimed to determine whether the issues confronting CE are unique and hence whether, as a result, CE should be developed as an independent Macroethics. By concentrating on just the aspect of AE, the view presented above suggests that to concentrate on uniqueness may not be necessary in order to reach that conclusion. Although it is manifest in cyberspace and readily studied there, AE is not necessarily unique to CE. It may be apparent, for example, in Environmental Ethics and in the world of physical automata (if only potentially in the latter case). Because of its novelty and important position in Ethics, AE seems to demand further study in its own right. Because it embraces many of the current difficulties of CE, it should be studied in, amongst other places, an applied setting where appropriate policy decisions can be analysed. The setting of CE seems appropriate, then, for at least three reasons:

a) CE has a methodological foundation IE (Floridi 1999a) and so is able to support theoretical analysis;

b) CE contains domain-specific issues, including pressing practical problems, which can be used to 'test' the results of (a). Moreover, standard forms of evil are also present and so can be used for purposes of comparison;

c) cyberspace is digital, and the notion of state rigorous, so actions can be quantified entirely; hence all consequences of an action in cyberspace can in principle be determined mathematically (cf. entropy structures), by contrast with the world of interacting sentient beings.

It may be that, in the uniqueness debate, the justification of the study of CE as a Macroethics has focussed on uniqueness of the issues confronting it as a result of the *policy vacuum* approach (Moor 1985). Although that view has done much to isolate and promote the novelty of problems confronting CE, perhaps it has led us to think that such a vacuum is *required* to justify the study of CE. By considering just one special topic—that of AE—we hope to have made the case affirming the study of CE as an independent Macroethics in an alternative manner. [11]

[11] This paper was given at the conference entitled "Computer Ethics: Philosophical Enquiry 2000", held at Dartmouth College, July 14-16, 2000, see the Proceedings of CEPE 2000, edited by D. G. Johnson, J. H. Moor and H. T. Tavani. Pp. 142-156.

REFERENCES

Floridi L. (1998) "Does Information have a Moral Worth in Itself?", *Computer Ethics: Philosophical Enquiry (CEPE'98)*, London School of Economics and Political Science, (London, 14-15 December, 1998), http://www.wolfson.ox.ac.uk/~floridi/cepe.htm

Floridi L. (1999a), "Information ethics: On the philosophical foundation of computer ethics", *Ethics and Information Technology* (1), pp. 37-56

Floridi L. (1999b) (ed.), *Etica & Politica*, special issue on *Computer Ethics*, 2, http://www.univ.trieste.it/~dipfilo/etica_e_politica/1999_2/homepage.html.

Floridi L. and Sanders J. (1999), "Entropy as Evil in Information Ethics", in Floridi (1999).

Johnson D. G. (1999), "Sorting out the uniqueness of computer-ethical issues", in Floridi (1999b).

King D. (1997), *Kasparov v. Deeper Blue*. B T Batsford Ltd, London.

Maner W. (1999), "Is computer ethics unique?", in Floridi (1999b).

Mitchell T. M (1997), *Machine Learning*, McGraw Hill.

Moor J. H. (1985), "What is computer ethics?" *Metaphilosophy* (16.4), pp. 266-275.

Pogue D. (1999), *Palm Pilot: the ultimate guide*. 2nd edition. O'Reilly.

Rumbaugh J. et al. (1991). *Object-Oriented Modeling and Design*. Englewood Cliffs, NJ: Prentice Hall.

Searle John R. (1980), "Minds, Brains, and Programs", *The Behavioural and Brain Sciences*, vol. 3. Cambridge: CUP.

Turing A. M. (1950), "Computing Machinery and Intelligence", *Mind* 59 (236), pp. 433-60.

Zimmerman, M. et al. (eds.) (1993), *Environmental Philosophy: From Animal Rights to Radical Ecology*. Englewood Cliffs, NJ: Prentice Hall.

ABOUT THE AUTHORS

L L Floridi was educated in Rome “La Sapienza” (Laurea in philosophy) and then Warwick University (MPhil and PhD in philosophy). He is Research Fellow in Philosophy (Wolfson College), and Lecturer in Philosophy (Keble College), University of Oxford (UK), where he is a member of the Sub-faculties of Philosophy and of Computation. His most recent book is *Philosophy and Computing. An Introduction* (Routledge, 1999).
Email: luciano.floridi@philosophy.oxford.ac.uk
Web: http://www.wolfson.ox.ac.uk/~floridi

J W Sanders was trained as a mathematician (BSc (Pure Mathematics, Hons) Monash, PhD (Abstract Harmonic Analysis) Australian National University) and since 1986 has been University Lecturer and Tutorial Fellow in Computation at the University of Oxford and Lady Margaret Hall (UK). His interests are the application of mathematics to the study of computer systems (in particular the specification and derivation of systems in both hardware and software, concurrency, probabilism, atomicity, security, quantum computation and semantics) and more recently in ethical issues.
Email: jeff@comlab.ox.ac.uk
Web: http://web.comlab.ox.ac.uk/oucl/people/jeff.sanders.html