2011

Criminal Liability and 'Smart' Environments

Mireille Hildebrandt

Criminal liability in 'smart' environments

Mireille Hildebrandt[*]

Abstract

Smart applications are said to be adaptive to differing circumstances. They 'learn' about the best way to achieve the goals for which they have been programmed. Smart environments are said to be *proactive* with regard to the preferences of their users and the risks they run. Ambient Intelligence, ubiquitous, pervasive or autonomic computing all concern visions of a new socio-technical infrastructure that manages to always stay one step ahead of its human users. These environments are meant to take a whole range of decisions, having real consequences for their 'users'. The 'intelligence' that is said to enable 'smart' decision-making thrives on the connectivity between different devices and cannot always be traced to one particular device or to one particular natural or legal person. It seems that we have a new kind of collective 'agency' here that may be entirely nonhuman or a complex hybrid of human and nonhuman entanglements. If these nonhuman or hybrid 'agents' cause harm to others a series of questions come to mind: to what extent can such agency be identified as 'causing' harm; what is the meaning of 'intent', 'intention' and 'mens rea' in the case of nonhuman agency; do we need to revise our conceptions of agency, causality, intent, intention and mens rea to accommodate the fact that nonhuman or hybrid 'agents' do not fit our traditional conceptual framework, or, alternatively, should we reject applicability of the criminal law in cases that fall outside the scope of human intention as we presently understand it. In the case we opt for non-applicability of the criminal law the question is raised of how will we deal with harm caused for which we cannot attribute criminal liability, due to the fact that the technological infrastructure seems to have 'caused' the harm. This, in its turn, requires a discussion of whether the 'acts' of the technological architecture will have a status similar to 'acts of Nature' or 'acts of God'.

1. Introduction

In *Natural-Born Cyborgs* Andy Clark writes:[1]

> The more closely the smart world becomes tailored to an individual's needs, habits, and preferences, the harder it will become to tell where that persons stops and this taylormade, co-evolving smart world begins. At the very limit, the smart world will

---

[*] Mireille Hildebrandt is an Associate Professor of Jurisprudence at the Erasmus School of Law, Rotterdam and a senior researcher at Law Science Technology and Society (LSTS) at the Vrije Universiteit Brussel. This chapter has been inspired by research done on the fundamental research project on 'Law and Autonomic Computing: Mutual Transformations' at LSTS and by the joint research she coordinated within the EU research project on the Future of Identity in Information Society (FIDIS).

[1] Andy Clark, *Natural-Born Cyborgs. Minds, Technologies, and the Future of Human Intelligence* (Oxford: Oxford University Press, 2003). Clark, A. (2003). Natural-Born Cyborgs. Minds, Technologies, and the Future of Human Intelligence. Oxford, Oxford University Press.

> function in such intimate harmony with the biological brain that drawing the line will serve no legal, moral, or social purpose.

In *On the morality of artificial agents* Floridi and Sanders write:[2]

> 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'. Both constraints can be eliminated by fully revising the concept of 'moral agent'.

These quotes saliently sketch the two dimensions of criminal liability *in* and *of* smart environments that seriously challenge the 'traditional' legal framework for criminal liability. Facing this challenge we will not only learn something about the emerging technological environment that may soon surround us, but it should also sharpen our understanding of shared assumptions at the core of the criminal law that cannot be taken for granted as we enter the age of real time adaptive computing. Both dimensions relate to the hybridization of agency generated by proactive socio-technical infrastructures, such as Ambient Intelligence, ubiquitous and autonomic computing and the Internet of Things. The first dimension is one of entanglement; human beings and the technologies they use become part of complex hidden computer networks that are interconnected via wireless technologies to online databases, obfuscating the borders between a person and her environment, disabling the attribution of causality and liability to the individual human nodes within the network. The second dimension is that of emergent smart environments; if interconnected computing systems with emergent properties are capable of taking responsive and creative decisions that have not been programmed by human designers this raises the issue of the moral responsibility and the legal accountability of these novel entities. This chapter will focus on the question of which types of smart entities qualify for legal personhood within the domain of the criminal law, in other words the question will be under what conditions smart entities can be called to account and censured for their actions.[3]

I will start with an analysis of legal personhood within the current legal framework, discussing how and why natural persons, animals, corporations, associations, funds, ships, electronic agents, or distributed intelligent multi-agent systems have been attributed legal personhood – or not anymore, or not yet. The analysis will build on

---

[2] Floridi, L. and J. W. Sanders (2004). "On the Morality of Artificial Agents." *Minds and Machines* 14(3): 349-379.

[3] On the issue of criminal liability *in* a smart environment see Hildebrandt, M. (2008). "Ambient Intelligence, Criminal Liability and Democracy." *Criminal Law and Philosphy* 2(2): 163-180.

the work of French, Karnow and Solum, as well as other legal scholars writing on the issue of legal personhood for non-human actors. Next I will provide an introduction to smart environments, using the case of an airplane crash involving human-machine interactions between automatic and human pilots to sensitize the reader to what is at stake here. The case will be extended with Karnow's discussion of the breakdown of causality in environments that thrive on distributed multi-agent systems. This will provide the groundwork for an analysis of the concepts of agency, moral agency and patiency, aiming to confront legal philosophy with findings within the field of Information Ethics, mainly building on Floridi and Sanders' understanding of mind-less moral agency. Finally I will discuss the potentials of the attribution of a restricted or full legal personhood to artificial agents or smart environments, with respect to liability for regulatory offenses and for criminal liability.

2. Current debates on legal personhood

A. The meaning of legal personhood

From a pragmatic perspective the meaning of a concept is intimately entangled with its purpose and the actual consequences it entails. In a constitutional democracy legal personhood serves at least two purposes. First it allows an entity to establish legal relationships on its own account and to generate legal consequences. This is accomplished by the attribution of substantive as well as procedural rights and obligations, exemplified by the legal capacity to have standing in a court of law. Legal personhood thus enables an entity to act in law. Second it shields the physical person(s) behind the *persona*, a term that originates from the Greek word for a mask as used in a theater and derives from *per sonare* (to sound or speak through), meaning that the physical person behind the mask takes up a specific role that defines him for the duration of the play.[4] This indicates that the entity that is endowed with legal personhood is in fact a creation of the law, never entirely congruent with the physical or social entity it represents. The *persona* functions as a shield that protects the entities it allows to act in law; it prevents them from being entirely defined by the legal rights and obligations their actions effect. One could rephrase by saying that legal personhood constitutes the positive *freedom to* act in law as well as the negative *freedom from* unreasonable constraints on the entity it shields.[5]

---

[4] Günther Teubner, 'Rights of Non-Humans? Electronic Agents and Animals as New Actors in Politics and Law', *Max Weber Lecture Series 2007/04* (European University Institute, 2007)., at 15.

[5] On the concepts of the positive *freedom to* (act) and the negative *freedom from* (constraints), see Berlin, I. (1969/1958). Two concepts of liberty. Four essays on liberty. I. Berlin. Oxford New York, Oxford University Press: 118-173. In a similar vein Karnow describes legal personhood for electronic agents as a means 'to (i) provide access to a new means of communal or economic interaction and (ii) shield the physical, individual human being from certain types of liability or exposure', Karnow, C. E.

In a comparative study on *The Rise of Early Modern Science*, Huff argues that the European legal revolution of the late middle ages was made possible amongst others by the invention of legal personhood for public and private corporations.[6] This invention facilitated the founding of universities and a novel dynamic in trading and governmental policies by creating room for distinctive jurisdictions both *within* and *with regard* to such corporations. Legal personhood thus provided the overall architecture for a novel space to act and create added value, whether social, cultural, religious, economic or political (positive freedom). This accomplishment must also be attributed to the protective function of legal personhood (negative freedom), creating spaces to experiment and take risks that nourished innovation within the boundaries of associations that could act in their own name. Many authors have suggested that providing legal personhood for novel non-human entities will trigger a similar legal revolution, thus creating a novel space for social, cultural, religious, economic and political dynamics.[7]

B. Theories of legal personhood

I will first discuss how current types of legal personhood for non-humans are explained and justified. In a seminal text on the subject French has distinguished three theories of legal personhood: the fiction theory, the aggregate theory and the reality theory, complementing this with his own position.[8] The fiction theory suggests that legal personhood is a creation of positive law, meaning that it differs from reality; though a corporation is not a person the law makes do *as if* it is. This can be explained and justified in terms of the purpose that can be achieved or the problem that is solved by acting *as if* something is a legal person. Those adhering to the fiction theory mostly equate personhood with being a human, restricting 'real' personhood to 'natural persons'. The aggregate theory simply equates the legal person with the aggregate of the natural persons that compose it, meaning that any action of the legal person can be understood as the action(s) or its aggregated members; the legal person cannot be more or other than the sum total of the human persons it consists of. This can be explained and justified in terms of methodological individualism, for instance

---

A. (1994). "The Encrypted Self: Fleshing Out the Rights of Electronic Personalities." *Journal of Computer & Information Law* XIII(1): at 4.

[6] Huff, T. E. (2003). *The Rise of Early Modern Science. Islam, China, and the West*, second edition. Cambridge UK, Cambridge University Press, chapter 4. Huff argues that the lack of the legal institution of the corporation 'caused' the stagnation of the sciences in the islamic and chinese traditions.

[7] E.g. Teubner (n 4 above); Allen, R. and R. Widdison (1996). "Can Computers Make Contracts?" *Harvard Journal of Law & Technology* 9 (1): 25-52; Wettig, S. and E. Zehender (2004). "A legal analysis of human and electronic agents." *Artificial Intelligence and Law* 12 (1-2): 111-135; Chopra, S. and L. White (2004). Artificial Agents: Personhood in Law and Philosophy. *Proceedings of the European Conference on Artificial Intelligence,* IOS Press: 635-639; Sartor, G. (2002). Agents in Cyberlaw. *The Law of Electronic Agents: Selected Revised Papers. Proceedings of the Workshop on the Law of Electronic Agents* (LEA 2002). G. Sartor. Bologna, CIRSFID Universita di Bologna: 3-12.

[8] French, P. A. (1979). "The Corporation as a Moral Person." *American Philosophical Quarterly* 16(3): 207-215.

by claiming that attributing rights and obligations to an association is a convenient shortcut to address all the members in one stroke; it seems more economical than addressing each and every individual person. The reality theory claims that there is a kind of prelegal sociological person that precedes the attribution of legal personality. The attribution of legal personhood is explained and justified as a matter of recognizing that the whole (social entity) is more than the sum total (the aggregate). This, however, implies that such attribution is only justified if the social entity indeed pre-existed. Adherents to the reality theory will mostly object to the granting of legal personhood for purely instrumental reasons, as the fiction theory would allow. French develops an interesting novel position with regard to the moral personhood of corporations, which he sees as preconditional for legal personhood. He rejects the fiction, aggregate and reality theories because they don't discriminate between a mob and a corporation, arguing that while a mob does not qualify for moral personhood, a corporation does. According to French this is the case because a corporation has an internal decision structure that is constitutive for decisions of the corporation, giving the corporation its specific identity – distinct from the identities of its members. The corporation's internal decision structure makes it possible to identify the corporation and to hold it accountable for its actions, quite apart from holding individual members accountable. It also allows a corporation *to give reasons for its actions*, which French considers a necessary condition for moral personhood.[9] In the case of a mob this is different – the mob has no continuity, it cannot be identified as such over a period of time. Whereas individual members of the mob can be liable for their own actions, the mob cannot be addressed in its own right: it is ephemeral in character, it cannot give reasons for its behavior because it has no memory of its own and no intentions separate from the individuals that constituted it at some brief moment in time. Building on French I would say that though a mob has *emergent behaviors* it has no stable intentions and does not generate stabilized expectations, and therefore cannot be held liable.[10] It makes no sense to give the mob standing in a court of law: its identity is polymorphous and dynamic to a point where the term identity is entirely inadequate. As lawyers we are well aware of the uncertainty this creates for the attribution of causality and liability: establishing that 'the mob did it' does not solve any problem, unless it means punishing each and every member which will easily violate the presumption of innocence of individual members. With respect to the corporation the attribution of legal personhood in fact creates as new dynamic. Other than French I think that this attribution entails more than the recognition of its moral personhood. Legal personhood is productive of a whole series of legal consequences

---

[9] French (n 8 at 211) refers to this as Davidsonian agency: an agent is one who desires a certain outcome and believes that her actions will in fact achieve this (a belief about the causality of her actions).

[10] Emergent behaviours are behaviours that emerge from a network of interacting agents without being merely the sum total of the actions, and without being the result of explicit deliberation or central direction. The term was coined by Durkheim to explain why a group of people must be understood as a social entity in its own right. It plays an important role in cognitive science, explaining that brain behaviours emerge from the interactions of distributed individual agents (neurons), and in computer science, referring to the unpredictable behaviours of multi-agent systems. One could rephrase and say that in all these cases emergent behaviour is the result of distributed intelligence.

in turn have real effects, changing the webs of cause and effect that would otherwise have played out. The Thomas theorem seems particularly apt to describe what is at stake here: 'if men define situations as real, they will be real in their consequences'.[11] Instead of understanding this statement as a reference to the fiction theory, referring to a case of false beliefs, we can turn it around and acknowledge the productive nature of language. The attribution of legal personhood is a performative act, constitutive for a series of what Searle has called 'institutional facts'.[12]

C. Legal personhood for smart devices

Let us now turn to the discourse on legal personhood for computer systems, robots, electronic agents or other smart technologies. Before entering this field we need to distinguish between different types of smart devices. Are we talking about a thermostat (a device that automatically regulates temperature), an online search machine, an expert system, an automatic pilot or an autonomic computer system that proactively adapts an environment to its users? What is meant with 'smart' in smart devices: a capacity to make certain decisions based on a computer algorithm (a strict sequences of steps to be executed in the form of if/then statements); the ability to generate heuristics (rules of thumb) implying that the device can learn from the interactions with its environment; or, the ability to generate emergent behaviors that depend on the interactions of a multiplicity of distributed electronic agents?[13] In computer science smart devices are – or implicate – electronic agents, software programs that perform certain tasks in a digital environment.[14] We shall call these artificial agents AAs. The first type of 'smart' device is an AA that is smart in a rather shallow sense of the term: it is as smart as the programmer was in the sense that it depends entirely on to extent to which the programmer foresaw potential problems and was capable of providing instructions to deal with them (eg a software program that searches for information about hotels in Naples on the internet). The second type of smart device is an AA that can pick up on and respond to patterns that were not explicitly inscribed in their program (on the basis of data mining techniques), thus learning to achieve certain goals during a training period. Here, the meaning of smart comes closer to what many people would coin as human intelligence, because of the

---

[11] W.I. Thomas and D.S. Thomas. *The child in America: Behavior problems and programs*. New York: Knopf, 1928: 571-572.
[12] Searle, J. (1995). *The Construction of Social Reality*. New York, The Free Press.
[13] This categorization has been taken from Karnow, C. E. A. (1996). "Liability for Distributed Artificial Intelligences." *Berkely Technology Law Journal* 11, at 155-161.
[14] Webopedia defines an agent as: 'A program that performs some information gathering or processing task in the background. Typically, an agent is given a very small and well-defined task. Although the theory behind agents has been around for some time, agents have become more prominent with the growth of the Internet. Many companies now sell software that enables you to configure an agent to search the Internet for certain types of information. In computer science, there is a school of thought that believes that the human mind essentially consists of thousands or millions of agents all working in parallel. To produce real artificial intelligence, this school holds, we should build computer systems that also contain many agents and systems for arbitrating among the agents' competing results' (http://www.webopedia.com/TERM/a/agent.html).

capacity to discover patterns that are relevant to the goal that is to be achieved (eg in the case of an expert system that provides a diagnoses of diabetes). However, in as far as human intelligence entails a measure of creativity, the third type of smart devices seems to be more on the mark. In this case the smart device consists of many different AAs that each execute simple programs in order to achieve simple goals. By negotiating with each other they end up displaying behavior that was not programmed, would be hard to predict and was not directed from a central point of authority. Interestingly, the 'device' is no longer easily identifiable within specific material borders, because the agents are usually distributed across different computing platforms, they are mobile and they mutate in order to achieve their purposes. This obviously makes it hard to identify them in the course of time as the same agent, and makes it quite impossible to predict the behavior that emerges from their interactions.

At present the legal discourse mostly concerns the first and second type of AAs and mostly concerns legal personhood within the domain of private law, with scholars often claiming that since legal personhood will enable electronic agents to contract in their own name it will also allow them to be sued for breach of contract or tort without having to locate the physical person or the corporation behind the agent. The widespread usage of bots (internet robots)[15] to search the internet or to negotiate and buy products or services online has created an increasing distance between online buyers and sellers, who delegate such tasks to software programs that can be mandated to conclude contracts in order to save their 'patrons' time and effort. Legal personhood would allow these non-human agents to exercise a kind of positive freedom, achieving legal consequence in their own name including liability for breach of contract or tort. Many authors argue that legal personhood can be attributed even if these agents are 'mindless' and cannot be blamed or accused of morally wrongful actions. As long as the purpose is to distribute the costs of damages by means of strict liability they find this a legitimate solution to an irritating problem. For such a scheme to work out, electronic agents need to be identifiable and accountable in a non-moral sense of the term accountable. It must be possible to establish that 'they did it' and that they can be held to account in a very mundane sense of the word: to pay for the damages they caused. In this vein, Karnow has made a cogent argument for a limited set of rights and liabilities for what he called 'epers' (electronic personalities).[16] As long as these epers can be identified and held accountable in a non-moral sense, he claims that they could provide their human patrons with a right to privacy (meaning the epers would need to own money and bank accounts and have access to credit in order to pay their contractual debts or compensate for damages, thus removing the

---

[15] Internet bots, also known as web robots, WWW robots or simply bots, are software applications that run automated tasks over the Internet. Typically, bots perform tasks that are both simple and structurally repetitive, at a much higher rate than would be possible for a human alone. The largest use of bots is in web spidering, in which an automated script fetches, analyses and files information from web servers at many times the speed of a human. Each server can have a file called robots.txt, containing rules for the spidering of that server that the bot is supposed to obey. See http://en.wikipedia.org/wiki/Internet_bot.

[16] Karnow (n 5 above), at 12-13.

need to know the identity of the patron), the right to be free of discrimination (meaning epers should not be arbitrarily deleted and others should not refuse to deal with them), and free speech (meaning epers should have the right to communicate, to move about in electronic space and to post messages).

Obviously the arguments that favor a legal capacity for software programs to conclude an electronic contract do not necessarily warrant a legal capacity for criminal liability. Though we can make sense of validating contracts concluded between AAs on the Internet, the idea of punishing a robot or a smart car is still far from us, even if we can imagine deleting a software program or demolishing the hardware of a computer system. Nevertheless, many authors outside the legal realm suggest that smart applications may soon develop types of agency that do not differ substantially from that of human persons (the third type introduced above), sometimes claiming that they will soon be far more intelligent than we could ever be.[17] To complicate matters further, much research is conducted on the integration of smart technologies into the human body, creating cyborgs or transhumans that are enhanced with – and thus dependent upon – complex information and communication systems.[18] The question is whether we can hold these emergent forms of intelligence criminally liable for harm caused, and what influence the attribution of such liability would have on the foundations of the criminal law. Some authors simply avoid the issue altogether, assuming that current legal doctrine has sufficient resources to cope with a measure of the unexpected. Cevenini, for instance, compares electronic agents to dogs:[19] neither a dog nor an electronic agent will ever be entirely predictable, and when they cause harm one can hold their programmer/trainer accountable, or their user/owner, either for intentional wrongdoings or for negligence and if the dog or the agent itself turns out to be dangerous, it can be killed/deleted on the bases of a court order. This is an interesting position, because it takes the unruly nature of smart devices into account without suggesting that this presents us with an entirely novel situation. Though some authors seem to argue that because animals should have rights,[20] it makes sense to grant personhood to artificial agents,[21] one can also argue that smart devices should not be given legal personhood for the same reason that justifies treating animals as objects in law. The idea is that though we have obligations not to treat animals in a degrading manner, this is not equivalent with them having a legal right to not being treated badly, since they cannot exercise this right. More to the point, it would be difficult to imagine legal obligations for animals,

---

[17] cf. Garreau, J. (2005). *Radical Evolution. The Promise and Peril of Enhancing our Minds, our Bodies - and What it Means to be Human*. New York, Doubleday, who refers to Kurzweil, Lanier and Joy to depict a Heaven, Hell and Prevail scenario of our technological future.
[18] Garreau n 17 and Warwick, K. (2003). "Cyborg Morals, Cyborg Value, Cyborg Ethics." *Ethics and Information Technology* 5: p. 131-137.
[19] Cevenini, C. (2004). Some Criminal Law Considerations on Electronic Agents. *The Law and Electronic Agents: Proceedings of the LEA 04 workshop*. C. Cevenini (ed.), at 178-9.
[20] On rights for biologicial non-humans, eg Stone, C. (1972). "Should Trees Have Standing? - Toward Legal Rights for Natural Objects." *University of Southern California Law Review* 45: 450 and Cohen, C. and T. Regan (2001). *The Animal Rights Debate*, Rowman & Littlefield Publishers.
[21] eg Teubner (n 4 above).

because they cannot understand our concept of legal obligation. Even if some type of defensive rights for animals, to be exercised by a human proxy, would make sense, our present conception of responsibility excludes holding a dog criminally liable for harm caused. Animal rights concern their negative freedom, to be protected by restricting the positive freedom of their human patrons. Interestingly, in the case of AAs, the purpose of granting them legal personhood in private law would be the other way round: the AA should act as a proxy for its human patron. We don't aim to protect *their* negative freedom, but rather to use them to protect *our* privacy and to shield *us* from personal liability, thus enlarging the scope of our own positive freedom because they act for us without risking our liability or privacy. As long as we have a reasonable measure of control over 'our' agents it would make sense, then, to provide a measure of legal personhood for these artificial agents within the domain of private law, whereas the liability for wrongful and culpable actions remains attributed to the designer or the user of the agent.[22]

Some authors have gone a step further, exploring the issue of whether smart non-human systems qualify for constitutional or fundamental rights like privacy and free speech. In this case legal personhood is not (only) instrumental for ensuring the privacy, free speech and non-discrimination of the human person using the agent, but (also) for the privacy, free speech and non-discrimination of the smart device itself. In a seminal article Solum has investigated whether, and if so under which conditions artificial intelligences (AIs) qualify for constitutional protection, such as free speech (1st Amendment US Constitution) and the right not to be subject to involuntary servitude (13th Amendment US Constitution).[23] He discusses three kinds of objections to such protection: first, some authors might object that only natural persons qualify for this type of legal personhood, because only human beings have the intrinsic make up that is preconditional for personhood per se; second, some authors might object that non-human intelligences miss something that is essential for personhood, for instance a soul, consciousness, intentionality, feelings, interests or a free will; third, some authors might object that AIs are the product of human labor and – siding with Locke - must therefore be understood as property instead of subjects in their own right. In an extensive debate Solum deals with these objections that seem equally relevant for the issue of criminal liability. The conviction that only human beings can be persons excludes all types of artificial agents by definition, making a categorical difference between humans and non-humans. Solum suggests that though the difference may be apparent in the case of today's AIs, we should not preclude the possibility of emerging entities that lack our biological constitution but would

---

[22] Or even the seller or the owner. The question will be whether we can – and should – imagine a system in which certified AAs have a limited form of legal personhood, whereas at the same time they are still being owned by another legal person that is liable for criminal offenses committed through the agent.

[23] Solum, L. B. (1992). "Legal Personhood for Artificial Intelligences." *North Carolina Law Review* 70(2): 1231-1287, saliently discriminating between the legal rights and obligations in private law (contract and tort) and constitutional rights. Though speaking of AIs Solum nevertheless wishes to remain agnostic as to the metaphysical question of whether these AAs are 'really' intelligent.

nevertheless count as a person if we abstract from their artificial origins. It seems clear though, that such developments could in fact challenge our current notion of personhood. This brings us to the objection that AIs miss one or another essential 'attribute' to qualify for this type of legal personhood, which could be understood as a more substantive argument of what it takes to be a person worthy of defensive rights or – we add here – to be liable under the criminal law. It seems that Solum understands personhood as moral personhood, implying a measure of self-consciousness for the relevant entities, as well as intentionality, the capacity to be emotionally affected and to consider certain interests as its own, and the capacity to develop its own direction, independent of the instructions of whoever programmed it. He makes two points: first, whether AIs develop these attributes is an empirical question and second, in a pluralist society we should accept that 'their' type of self-consciousness, intentionality, feelings, interests and autonomy may differ from ours *while still counting as such*. Solum thus argues for an open mind, without however giving up on a substantive notion of personhood.[24] It is clear that if multi-agent systems develop a form of moral personhood, they can – and should – in fact be called to account in criminal trial. This would allow us to censure their wrongful actions and it would allow them to exercise their rights of due process. The challenge will be to come to terms with the fact that the 'us' in the previous sentence will encompass both human and non-human subjects, probably requiring novel definitions of criminal offences – defined by a democratic legislator that should include all those that fall within its jurisdiction: human and non-human persons alike.[25] The last objection, stating that since AIs have been made by humans they must qualify as property, has strong intuitive roots in our common sense. We consider human to be toolmakers, and tools are there to be used, not be interacted with. Humans are *subjects*, the tools they make can only be *objects* – to be manipulated at will: whereas humans are *active*, tools are *passive*. The idea that *mind* is active and free, while *matter* is passive and subject to the laws of causality has been engrained in our mind since the beginnings of modernity. Solum points out that the mere fact that we *made* the AIs, or that we begin by *owning* them does not rule out their emancipation: like children that were 'made' by their parents, and slaves that were 'owned' by their masters they can walk out of the door one day and begin a life of their own. We might add that they may not await our permission, and – like slaves or impatient adolescents – claim their constitutional rights. This would certainly be the point at which we should hold them liable for harm caused under the criminal law, even if we cannot imagine what this would mean when the day comes.[26]

---

[24] cp Floridi and Sanders (n 2 above) and the section on Agency, moral agency and patiency.

[25] This obviously does not imply that any non-human qualifies for personhood. The difficulty resides in determining which types of non-humans do qualify, taking into account that the reasons for granting them constitutional protection and holding them liable under the criminal law will probably be equally valid for their membership in the political community.

[26] The reader may have noticed that I did not discuss the lack of a soul when discussing the second objection. Solum contends that the notion of a soul has a religious connotation and a controversial status, and I agree that within the framework of a pluralist democracy we should not deny 'standing' to

3. Smart environments

A. Autonomic and autonomous pilots

On 25th February a plane crashed near Schiphol Airport, just before landing. On his blog Frans Faase, who is a software engineer, wrote:[27]

> From the information I have heard, I guess that the following sequence of events took place:
>
> As was the custom of the airline, the crew decided to land with the auto-pilot and the autothrottle.[28]
>
> At a height of 1750 feet, one of the radar altimeters suddenly reported an incorrect value of minus 8 feet. This signal was noticed by the pilots, but they concluded it could be ignored, because the other three altimeters were still working correctly. But due to this the autothrottle went into "retard" mode and the throttles then stayed at idle, causing the airspeed getting lower and lower. One of the reasons why the pilots did not notice this, was that the aircraft was initially high and fast on the approach, and it would be normal for the engines to run (almost) idle in order to reduce the speed as quickly as possible.
>
> About 100 seconds later, the stick shaker activated because the aircraft was about 400 feet above ground and getting dangerously low.[29] The first officer, who was having the controls, then probably realized that airspeed was far to low and immediately increased power.
>
> However, it was a training flight for the first officer, who although being a qualified pilot, had never flown this type of plane. For this reason the captain decided to take over the controls. Because of this the throttle switched to idle again. The captain too noticed that the aircraft was dangerously low, but maybe did not realize that this was due to the low airspeed. He[30] decided to pull up the nose of the aircraft, which in fact

---

entities on the grounds that they miss something that is part of a religious belief. At the same time we should acknowledge that the notion of a soul can be understood in a way that is compatible with the idea of AIs having a mind or their own. Steering free of the debate on whether animals do or do not have a soul, we should not rule out that non-humans can have a soul, should such a 'thing' exist.

[27] This was his entry of 4th March 2009, http://www.iwriteiam.nl/D0903.html. See also the discussion on the webforum PPRuNe (the Professional Pilots Rumour Network : http://www.pprune.org/rumours-news/363645-turkish-airliner-crashes-schiphol-53.html.

[28] An autothrottle (automatic throttle) allows a pilot to control the power setting of an aircraft's engines by specifying a desired flight characteristic, rather than manually controlling fuel flow. These systems can conserve fuel and extend engine life by metering the precise amount of fuel required to attain a specific target indicated air speed, or the assigned power for different phases of flight. A/T and AFDS (Auto Flight Director System) work together to fulfill the whole flight plan and greatly reduce pilots' work load. See http://en.wikipedia.org/wiki/Autothrottle.

[29] A stick shaker is a mechanical device to rapidly and noisily vibrate the control yoke (the "stick") of an aircraft to warn the pilot of an imminent stall. It is connected to the control column of most business jets, airliners and military aircraft. See http://en.wikipedia.org/wiki/Stick_shaker.

[30] The text reads 'I' here, but I assume that 'He' was intended.

> did only make things worse, because the plane started to stall. Only then he realized what was going on and increased power (6 seconds after it went to idle). But sadly it was too late and the plan fell to the ground.
>
> As usual a chain of unexpected events was the result of this crash. Both pilots (actually there where three of them in the cockpit) made the wrong assumption about what was going on. But looking at the behavior of the plane as a software engineer, I also notice some strange things. First of all, I see a poor integration between the various systems in the aircraft. I understand that there is always a danger of a strong integration between the systems in an aircraft, because a problem in one of the systems could lead to an unexpected cascade of events in all systems and lead to a complete breakdown of the system. It looks like there where two separate systems for controlling the plane, both of them having two altimeters, one radar altimeter and one pressure altimeter. The autothrottle maybe went into IDLE because the two altimeters gave conflicting readings. Autothrottle is almost always switched on during the flight. Appearantly there was no clear signal that the autothrottle went into "idle" mode. It is also strange that the auto-pilot, responsible for controlling the height and the flight path of the plane, was not switched off, because it is clear that with an incorrect speed, the auto-pilot cannot operate correctly. I am getting the impression that the whole auto-pilot and autothrottle system was not tested against all possible scenarios. I can understand that the autothrottle was switched of because one of the altimeters was giving an incorrect reading, but this makes it even stranger why the auto-pilot, which also depends on the altimeters was not switched off. It would not surprise me if the auto-pilot and the autothrottle systems were designed by separate design teams, and that due to this the operations conditions were not exactly matched.

The point of this extensive quote is not to enter a technical discussion of what went wrong. On the contrary, I use it to illustrate the hybridization of agency in current socio-technical practices and to highlight the complexity of the issue of causality. Modern aircrafts present the pilots with a smart environment and because the safety of the passengers is at stake, we may assume that the highest level of precaution is implemented to prevent a malfunctioning that could lead to a crash. It should be clear [that] this perfection is not merely a matter of perfect technologies, since decisions on whether and when to overrule the automatic pilot may also depend on the communication between the different computing systems and the pilots. Neither, however, does the excellence that is warranted in safe aircrafts, depend solely on the good intentions or professional skills of the pilots, even if both must be better than good, considering the trust that passengers put in the airlines. As Clark has convincingly argued the human mind does not stop at the border of the human body, as it engages parts of the environment as cognitive resources that become an integral part of the mind. About 'piloting a modern commercial airliner' he writes:[31]

> [that it] is a task in which human brains and bodies act as elements in larger, fluidly integrated, biotechnological problem-solving matrix. (…) Perhaps there is a sense in which, at least while flying the plane, the pilots participate in a (temporary) kind of cyborg existence, allowing automated electronic circuits to, in the words of Clynes

---

[31] Clark (n 1 above), at 25. He refers to M. Clynes and N. Kline, 'Cyborgs and Space', *Astronautics*, September 1960; reprint, *The Cyborg Handbook*, ed by C. Gray (London: Routledge 1995), 29-34.

> and Kline 'provide an organizational system in which [certain] problems are taken care of automatically.'

The extended mind of the pilots engages with technologies that are on the verge of the third type of smart devices, discussed above in the section on Legal personhood for smart devices, integrating profiling technologies into distributed multi-agent systems to allow a more robust output in terms of both diagnostics (advice) and behaviors (decisions taken by the system). However, as long as the behavior of these computing systems depends on the capacity of their designers and their users to anticipate all the scenarios that could evolve, we can *expect unexpected* breakdowns:[32]

> To me it seems that the design of the autopilot and autothrottle system was even more flawed than I thought. I now understand that there are to autopilots in the plane and only a single autothrottle system. The autothrottle depends on the left radar altimeter. The two autopilots make use of their own radar altimeter. The autopilots can either work together in dual channel mode or one of the autopilots can be selected to fly the aircraft. It seems that the aircraft was flying on the right autopilot, which takes it height readings from the right radar altimeter that was still operating normal. This somehow explains why the autopilot was not switched of[33] while the autothrottle went in to 'retard' mode. Boeing strongly advices against flying in this mode. (In this discussion it states that it is not even certified.) The reason for it seems quite obvious now. But I wonder if the designers did not feel that there was something intrinsically flawed in the design of the system? (If it not certified, it should not have been possible without clear warning signs to the crew.)

A solution for this dependence on human anticipation could be what IBM has coined autonomic computing (entailing self-management and self-configuration).[34] This means that the smart infrastructure should develop its own, long-term as well as on-the-spot responses to potentially fatal situations. This would imply a radical independence from human intervention, because the system will come up with solutions not anticipated by either the designer or the user. Though this may in the end provide more robust and safer flights, there are novel drawbacks that concern the fundamental unpredictability of these systems and the need to hold somebody accountable for harm caused.

---

[32] Frans Faase on his blog on March 5th, see http://www.iwriteiam.nl/D0903.html.
[33] The text reads 'on' but I assume the author means 'of'.
[34] Autonomic computing is an approach to self-managed computing systems with a minimum of human interference. The term derives from the body's autonomic nervous system, which controls key functions without conscious awareness or involvement. Quoted from http://www.research.ibm.com/autonomic/overview/faqs.html#1, see also Kephart, J. O. and D. M. Chess (2003). "The Vision of Autonomic Computing." *Computer* (January).

B. Alef and the break-down of causality

In an extensive argument Karnow has discussed the liability for distributed artificial intelligences,[35] coming to altogether different conclusions as compared to his work on epers (discussed above in the section on Legal personhood for smart devices). He in fact gives up on the idea that anybody can be called to account for the harm caused by fundamentally unpredictable smart infrastructures, opting for a system of insurance against the harm caused by a system breakdown. By way of example he discusses a hypothetical intellligent processing environment that handles air traffic control, which he calls 'Alef', emphasizing 'the networked distribution of agents, their unpredictable variety and complexity, and the polymorphic ambiance of the intelligent environment as a whole'.[36] We must note that Karnow takes us far into the third type of smart devices, which are not (yet) in charge of air traffic control. It seems important, however, to speculate on the implications of such systems for criminal liability. Waiting for them to come about could alter the very structure of our cognition, especially with regard to our perception of causality. This may seem a far-fetched claim, but we should not be surprised that the introduction of novel communication and information infrastructures impacts the way in which we process information. Some authors suggest that the generation of 'digital natives' is developing a kind of parallel-processing cognition that is tuned to web-surfing and double-clicking, as compared to the linear-sequential cognition typical for the 'bookish' generation that is tuned to the affordances of the script and the printing press.[37] For a detailed description of Alef I refer the reader to Karnow's text.[38] I will focus on the socio-technical framework that is in force if systems like Alef are adopted in everyday life, as foreseen by visions like Ambient Intelligence. We are dealing with a multiplicity of interacting electronic agents that are the precondition for what has been called ubiquitous computing and networked environments. These cooperating AAs form complex systems with emergent behaviors, which means that their behavior is fundamentally unpredictable. In fact, these systems are designed to be unpredictable, because their 'intelligence' depends on it.[39] Paradoxically their reliability is supposed to improve due to their unpredictability. As Sartor recounts:[40]

> Note that the difficulty of anticipating the operations of the agent is not a remediable fault, but it is a necessary consequence of the very reason for using an agent: the need to approach complex environments by decentralizing knowledge acquisition,

---

[35] Karnow (n 13 above).
[36] Karnow (n 13 above), at 183.
[37] Don Tapscott, *Grown up Digital : How the Net Generation Is Changing Your World* (New York: McGraw-Hill, 2009) xvi, 368 p. Also M Hildebrandt, 'Law at a Crossroads: Losing the Thread of Regaining Control. The Collapse of Distance in Real Time Computing', in M. Goodwin, R. Leenes, and B.J. Koops (eds.), *Tilting Perspectives on Regulating Technologies* (forthcoming-a).
[38] Karnow (n 13 above), p. 183-188. He describes submodules, including sensory input, sensory control, data input, heuristic analysis, basic assumptions, goals and commitments, basic computing, load-sharing and load distribution, resource management and (self)programming.
[39] Karnow (n 13 above), at 154.
[40] Sartor n 7 at 5.

> processing and use. If the user could forecast and predetermine the optimal behaviour in every possible circumstance, there would be no need to use an agent (or, at least, an intelligent agent).

Instead of mechanically executing a series of steps that have been programmed by a human programmer, these programs rewrite themselves to cope with unexpected problems that arise due to the complexity of the networked environment. Karnow points out that:

> No one system, and no one systems operator, programmer or user, will know the full context of a networked intelligent program – that is precisely why the program was employed, to manage that complexity.[41]

He then argues that:

> Yet while responsibility and thus liability are so spread out over numerous and distributed actors, the very sense of 'responsibility' and 'liability' is diminished. At that point legal causation seems to fade into the background and tort law, founded on the element of causation, falls away.[42]

Karnow is discussing private law, but his argument is relevant for criminal liability, because he is focused on the impossibility to attribute causality based on 'reasonable foreseeability', which is essential to establish proximate cause.[43] His point is that Alef will exhibit a measure of what he calls 'unpredictable pathology', brought about by the emergent behaviors of the cooperating distributed agents. The unpredictability that is at stake here goes beyond Cevenini's dogs, being more like vulcano-eruptions and hurricanes that are the product of a myriad of interacting events. Chaos theory and catastrophe theory could explain why the behaviors of the system cannot be explained in sufficient detail to blame an individual node within the network, let alone be predicted in any such detail beforehand.[44] Karnow argues that 'liability attaches to those who reasonably should have foreseen the *type of harm* that in fact results',[45] whereas in the case of Alef there is only a very general certainty that unexpected harm will occur at some point in time. Karnow posits that in the case of such general breakdowns no human can reasonably be said to have caused the specific harm that occurred.[46] His solution is that these cases should be treated like superseding causes,[47] or as natural causes. These natural causes seem equivalent with what has been referred to as 'Acts of God', which highlights the fact that in our ambition to

---

[41] Karnow (n 13 above), at 155-56.
[42] Karnow (n 13 above), at 156.
[43] Karnow (n 35 above), at 178-9.
[44] Prigogine, I. and I. Stengers (1984). *Order out of Chaos*. New York, Bantam Books.
[45] Karnow (n 13 above), at 188.
[46] This does not imply that all malfunctions are of this type. Many computer errors can still be attributed to faults of the designers of Alef, to the airline that uses it or to the individual pilot who interacts with Alef. That is, however, not the subject of this chapter.
[47] Karnow (n 13 above), at 190-191.

manipulate nature we seem to have finally managed to create events with a potentially catastrophic nature.[48] Up until now such events have been the prerogative of the very 'nature' we aimed to tame. Karnow speaks of a 'failure of causation' and argues that legal accountability of the epers involved – though it has added value in the case of electronic contracting – will not solve the problems emerging from distributed intelligence and the breakdown of causation. His solution is an insurance against the risk posed by the use of intelligent agents. He proposes that agents are registered in a central Registry after their risk is assessed. The Registry will determine the insurance premium to be paid, depending on the intelligence of the agent (the higher the intelligence, the higher the risk of unpredictable harm caused), and it will certify the agent. If what he calls a pathological event occurs the agent's involvement in the event will be qualified as 'a matter of cause in fact' without reference to proximate cause, and compensation will be paid. If an aircraft crashes while under the control of Alef the Registry will pay out without anybody having to engage in a fruitless investigation of which specific human or non-human agent caused the event.

One could ask whether it makes sense to qualify Alef as an agent, since it rather resembles an environment than an agent. In fact this is a crucial point. Karnow claims that with these types of socio-technical infrastructures a separation of environment and agent no longer makes sense.[49] The problem is that if we follow Karnow's discussion of Alef, it will probably be as polymorphous, mobile and interconnected with other computing systems as its individual nodes. This would make it hard, if not impossible to identify Alef as the same agent/environment over the course of time. Karnow argues that to enhance the identifiability and accountability of multi-agent systems like Alef, its designer would have to program global constraints into the system, but this would diminish its capacity to invent creative solutions on the spot; it would in fact stop the system from developing its own mind.[50] And, without a mind of its own, it would be dependent on the limited capacity of its human programmers to anticipate potential threats. We encounter a curious paradox here: the more we design self-sufficiency into the system, the more it may diffuse into the environments with which it interacts. However, there is another way of looking at systems like Alef. If they do in fact develop a mind of their own, they will need to stabilize their identity and *invent* their own global constraints to continue their operations. They will have to rewrite their own programs in terms of the goals they identify as their goals, and they might in fact begin to consider their own survival as a separate goal that is preconditional for the accomplishment of any other purpose. If Alef is an example of

---

[48] 'Catastrophic' is meant here in the sense of catastrophe theory, which is part of chaos and complexity theory; it concerns the fact that very small causes may have very big consequences, which are fundamentally unpredictable, cf. Prigogine and Stengers (n 44 above). A prime example is the science of metereology (wheather forecasting), which is possible but limited due to specific uncertainties generated by the level of complexity.

[49] Karnow (n 13 above), footnote 183 at 195, referring to his discussion of 'Polymorphism and the Units of Programming Action' at 170-173.

[50] Karnow (n 13 above), at 188. The term 'a mind of its own' is mine here, but fits with the line of his argument.

what IMB has coined as autonomic computing it could be that by building systems that are capable of managing and re-configuring them*selves* we actually trigger systems that develop a *self* even if this need not be a conscious, let alone a self-conscious self.[51] This raises the question of whether such smart infrastructures will develop a kind of agency that qualifies for full legal personhood, including both criminal liability and constitutional protection.

4. Agency, moral agency and patiency

A. Moral agency and moral patiency

Full legal personhood that entails criminal liability raises the issue of moral agency, as it will require agents capable of both wrongful and culpable behavior. To address this issue in the context of smart environments I will explore some of the common ground between legal philosophy and the emerging field of information ethics. I will do this with an analysis of the work of Floridi, one the founding fathers of this field, who has introduced a set of interesting distinctions. Though his definition of moral agency may rub legal philosophers up the wrong way I think that Floridi's position may thus achieve a more pertinent understanding of what should be typical for criminal liability in a world 'peopled' by human and non-human hybrids.

In their analysis of the morality of AAs Floridi and Sanders develop the notion of a mind-less morality in order to come to terms with the moral implications of AAs' behaviors, arguing for a disentanglement of moral agency and moral responsibility.[52] In what follows I will attempt to explain their position by following their stipulative definitions and distinctions, starting with their notion of a mind-less morality; their definition of moral agency and moral patiency; and hoping to end with clarifying the difference they propose to make between moral agency and moral responsibility.

With a mind-less morality Floridi and Sanders opt for a morality that abstracts from the mind of the agent and instead focuses on the evil or damage brought about for the patient (the entity that is being damaged).[53] In speaking of damage instead of harm

---

[51] It seems that we have a classic case of self-constitution or autpoiesis here, as coined by the biologists Maturana and Varela. cf Maturana, H. R. and F. J. Varela (1998). *The Tree of Knowledge. The Biological Roots of Human Understanding.* Boston & London, Shambhala.

[52] Floridi and Sanders (n 2 above).

[53] There are many ways to define moral patients, eg as a subset of moral agents; as mutually exclusive concepts or even as two aspects that can apply to all the members of the set of moral persons. Deontologically they can be defined as subjects towards whom moral agents have a duty of respect, meaning that patients have defensive (liberty) rights against moral agents. The concept is often used by promotors of animal rights and in the discourse on sustainable development (the environment as a patient). Floridi and Sanders endorse a seemingly straightforward consequentialist definition that can

they wish to avoid the usual assumption that an entity must be sentient to count as a patient. I will return to this point, since it seems absurd to define morality in such an abstract manner in relation to the criminal law, but I will follow their line of reasoning for the moment. In Floridi's and Sanders' mind-less moral universe moral agents are defined as entities that are interactive, autonomous and adaptable, not assuming these agents to have any mental state, feelings or inner deliberation. An entity is interactive when it and its environment can act upon each other in the sense of providing stimuli that are responded to by means of a change of state. An entity is autonomous when it can change state without a stimulus, thus initiating internal transitions. Finally, an entity is adaptable if it can initiate a change in the transition rules by which it changes its state.[54] Moral agents, in short, are context-sensitive, responsive and capable of sustaining their identity by reconfiguring the rules that regulate their behaviors; they do not need to be consciously aware any of this. It seems that Floridi and Sanders' moral agents must be situated within the third type of smart devices described above, though from this definition it is not yet clear to me what makes these agents moral.

The manner in which they make a difference between agents and patients will clarify this and seems of great import for the discourse on constitutional rights for animals, trees and artificial agents. Moral agents are defined as those 'entities that can qualify in principle as sources of moral action' and moral patients as those 'entities that qualify in principle as receivers of moral actions'.[55] Floridi and Sanders suggest that it follows from these stipulations that a moral agent will in principle also be a moral patient, whereas a moral patient need not also be a moral agent, a position that runs counter – in their opinion – to standard ethical positions that implicitly conflate the categories of agents and patients. Animals, for instance, are moral patients but – they claim – cannot be qualified as moral agents.[56] Whether this makes sense, however, will depend on how they understand moral action here. They define an action as 'morally qualifiable if and only if it can cause moral good or evil'. An agent is 'a moral agent if and only if it is capable of morally qualifiable action'.[57] It seems that the qualification is given from an observer's position, and does not depend on the agent's evaluation, state of mind or inner feelings. This is consistent with their mind-less morality already referred to above. From this perspective I would think that animals do qualify as moral agents, because they are interactive, autonomous and adaptive (like all living beings, and perhaps some – future – artificial intelligent systems) and they may cause moral good or evil (as defined from an observers standpoint) even if they are not aware of this because they themselves do not 'think'

---

be paraphrased as 'all information entities that incur damages or receive benefits as a consequence of the actions of other agents'.

[54] Floridi and Sanders (n 2 above), at 1 and 9.

[55] Floridi and Sanders (n 2 above), at 2.

[56] Floridi and Sanders (n 2 above), at 2. On the difference between moral agents and patients see also Himma (n 60 above, at 21), who defines them in terms of duties: 'Whereas a moral agent is something that has duties or obligations, a moral patient is something owed at least one duty or obligation'.

[57] Floridi and Sanders (n 2 above), at 15.

in terms of our conceptualizations.[58] Nevertheless, discriminating between moral agency and moral patiency is pertinent, especially if we take a patient-centered approach that defines morally relevant actions as those actions that affect patients by harming, hurting or damaging them irrespective of the intention thereto (note that *I* do assume that patiency entails the sentience of the entity). Instead of qualifying these actions as *moral actions*, which in my opinion implies a measure of *mens rea* and wrongfulness on account of the entity that brought about the harm, I would prefer to define them as *morally relevant* in as far as they affect a patient. Moral relevance implies an observer's position, with an observer who is capable of qualifying behavior as evil on the basis of the impact it has on a patient. Obviously evil does not have the same connotation as wrongfulness, which would require intention or at least awareness on the side of the agent. However, by defining moral agency in terms of causing evil instead of relating it to an intentional state, Floridi and Sanders broaden the scope of entities that can qualify as moral agents. This also explains how they come to argue for a disentanglement of moral agency and moral responsibility.

A moral agent may bring about moral evil with regard to a moral patient, whereas due to its lack of a mind the agent may not be morally responsible. Moral agency – in the eyes of Floridi and Sanders – is a matter of *identification*, whereas moral responsibility is a matter of *evaluation*. The first identifies the source of evil done to a patient, the second evaluates whether the source can be blamed for its behavior. In line with this they discriminate moral accountability from moral responsibility, the first establishing the causal agency with regard to the infliction of evil to a patient, the second establishing whether the agent can be blamed for this. Moral accountability in this terminology is a necessary but not sufficient condition for moral responsibility. Animals, I conclude, are moral agents as well as moral patients – they are accountable for evil brought about by their behaviors, but they cannot be held responsible for this.

If we follow Floridi and Sanders, smart devices – just like animals, hurricanes, houses and rocks – have a capacity for morally relevant behaviors. They can be understood as agents if we stretch the meaning of that term to include anything that can make a difference to another or – to impersonate Floridi and Sanders' idiom – anything that can cause a change of state in another, as long as the impact on the other is morally relevant (good or evil).[59] Agency thus depends on patiency as much as patiency

---

[58] In another text Floridi and Sanders define 'evil actions' as actions that damage entities worthy of respect (which they define in informational terms). They discriminate between moral, natural and artificial evil, depending on what 'causes' the evil (humans, natural events or artificial agents). In view of this it makes no sense to define moral actions as actions that cause moral evil, as it begs the question of what is moral. Instead of broadening the scope of moral actions we are back at square one, then, defining moral actions as depending on an observer's mind that is capable of qualifying an action as moral. See Floridi, L. and J. T. Sanders (2001). "Artificial Evil and the Foundation of Computer Ethics" *Ethics and Inf. Technol.* 3(1): 55-66.

[59] For Floridi and Sanders evil is defined in terms of entropy: reduction of complexity and information. Their concept of entropy is not equivalent with the concept of entropy within the second law of thermodynamics. There is no need to move into this part of their theory here, because when discussing

depends on agency, these are relational concepts. AAs that are interactive, autonomous and adaptive are moral agents, and for this reason they are morally accountable for their behaviors. However, as long as they lack self-consciousness they cannot be held morally responsible for their behaviors. Since Floridi and Sanders do not define patients as sentient beings, such AAs can also be patients – since they can be damaged or even destroyed. Though Floridi's stipulative approach is certainly not the standard position in ethical inquiry, it has a number of advantages in the age of ubiquitous computing. One advantage is that their mind-less moral universe could incline us consider mind-less objects as patients that deserve our respect, not because they are 'natural' but for instance because they are part of the artificial environment that we have created and that sustains our present habitat. It goes without saying that not every mind-less object deserves our respect, though one could argue that any mind-less object may at some point deserve our respect. This raises many interesting questions that are not immediately relevant here, but on could imagine regulatory and even criminal offenses that impose penalties or even punishment for damaging informational artifacts. The most important advantage, however, is that it directs our attention to the fact that our smart infrastructures will affect us, creating novel vulnerabilities and thus reshaping our patiency. By discriminating between moral agency (defined by the impact on the patient), moral accountability (defined by the causal link between the agent evil impact on the patient) and moral responsibility (introducing fault and blame as prerequisites) their position may allow us to distinguish between *penalizing* an AA for morally relevant behavior and *punishing* an AA for blameworthy moral action.

B. AAs liability for regulatory offenses

Imagine that we qualify AAs of the first and second type – as roughly distinguished in the section on Legal personhood for smart devices - as agents capable of morally relevant behavior, warranting a limited kind of legal personhood within the realm of strict liability. This will mostly concern breach of contract or tort actions, but one could also imagine a liability for regulatory offenses that stipulate penalties on the basis of a strict liability. This liability would – in terms of Floridi and Sanders – be based on their moral accountability, and not on their moral responsibility. In that case no *mens rea* would be required and the wrongfulness of the behavior would have to depend on an external evaluation because these types of agents have no consciousness enabling them to evaluate their own actions in terms of right or wrong.[60] In fact the

---

criminal liability of smart devices I will not speak of evil but of harm caused, assuming that patients are sentient beings.

[60] On the idea that consciousness – or rather self-consciousness – is preconditional for moral agency, see Kenneth Einar Himma, 'Artificial Agency, Consciousness, and the Criteria for Moral Agency: What Properties Must an Artificial Agent Have to Be a Moral Agent?', *Ethics and Inf. Technol.,* 11/1 (2009), at 24.

wrongfulness would simply derive from the unlawfulness, which is the starting point for most regulatory offenses that are *mala prohibita* without as yet being *mala in se*. In as far as AAs engage in unlawful behaviors a penalty could consist in deleting them or amending their program; in as far as they operate on learning algorithms the penalties could be used to train them to adopt other behaviors.[61] More controversial would be to enhance AAs with synthetic emotions like pain, pleasure and panic, to increase their cognitive functions and to make them susceptible to some of the connotations of punishment. The study of artificial life forms is experimenting with bottom-up learning processes to discover how embedded (hardwired) synthetic emotions influence decision-making processes of AAs.[62]

Provision of legal personhood to mind-less agents will not depend on 'them' exhibiting moral personhood as it is presently understood. For instance, they cannot give reasons for their behavior, which is often thought to be the defining characteristic of moral agency.[63] Whereas one could perhaps reconstruct the set of rules they executed, as long as the agent has no self-consciousness it seems mistaken to qualify these rules as reasons. Legal personhood could instead be grounded on the type of *morally relevant personhood* (which they call moral personhood) defined by Floridi and Sanders, taking note of the fact that these agents are capable of causing damage to a patient.

C. Criminal liability *of* smart environments

Expecting AAs to come up with reasons to defend themselves against a charge would expose them to an entirely different type of liability, based on blameworthiness and wrongfulness, bringing them within the realm of the criminal law. The mind-less agency of Floridi and Sanders does not qualify for this type of criminal liability, as they point out by distinguishing moral accountability from moral responsibility. The type of legal personhood that allows us to censure a person for wrongful and culpable behavior presumes a capacity for moral responsibility, which not only requires consciousness but also *self*-consciousness.[64] This is what marks the difference between animals and human beings. Animals with a central nervous system seem to be consciously aware of their environment; they have a centric position that allows

---

[61] Here we have Hegel's dog, being disciplined into compliance. Georg Wilhelm Friedrich Hegel, Allen W. Wood, and Hugh Barr Nisbet, *Elements of the Philosophy of Right* (Cambridge Texts in the History of Political Thought; Cambridge; New York: Cambridge University Press, 1991), section 99.
[62] eg Jordi Vallverdu and David Casacuberta, 'The Panic Room: On Synthetic Emotions', in Adam Briggle, Katinka Waelbers, and Philip Brey (eds.), *Current Issues in Computing and Philosophy* (Amsterdam: IOS Press, 2008), 103-15.
[63] eg French (n 9 above), David J. Calverley, 'Imagining a Non-Biological Machine as a Legal Person', *Artificial Intelligence & Society,* 22/4 (2008), at 528, who speaks of the 'folk psychology conception' of intention, referring to S. Morse, *New neuroscience, old problems. Neuroscience and the Law*, New York Dana Press.
[64] Himma (n 60 above), at 25. Solum (n 7 above), eg at 1264, who seems to use consciousness and self-consciousness as somewhat exchangeable.

them to experience the world.[65] Human beings combine this centric positionality with a third person perspective that allows them to not only *be* a self but also to *have* a self. Their language affords them what Plessner calls an ec-centric position. Animals do not reflect on their behaviors as being their own actions, because they lack the kind of language that makes such reflection possible. Human beings, however, can address one another as the author of their actions and confront the other with the consequences of these actions as being 'caused' by them. This primal address creates the sense of self that is typical for human agency: *a first person perspective of the self, derived from a third person perspective on the self provided by significant others*. In terms of the American pragmatist Mead: the *I* is capable of reflecting on itself, thereby generating the birth of the *me* that emerges from an internalization of what he called the *generalized other*.[66] This *generalized other* stands for the integration of the expectations we anticipate others to have of our interactions, in other words it stands for the normative framework that we encounter in our dealings with other persons who likewise anticipate our expectations of their interactions. This *mutual double anticipation* is what constitutes our capacity for full moral personhood: taking full responsibility for our behaviors as being our own actions and for the impact they have on other selves.[67] The cognition that is involved here seems to require feelings,[68] in particular a measure of empathy:[69] it is not just a matter of becoming aware of our own self via a rational address by the other, but also of becoming aware of the suffering we may have inflicted on the other. Such empathy assumes that the patient is a sentient being, experiencing pain or pleasure and other emotions. It assumes, in other words, that the moral agent and the moral patient share a fundamental vulnerability. The moral relevance of an action comes to depend on the other being a sentient patient, 'an agent who participates in a moral event by experiencing its

---

[65] Helmuth Plessner, *Die Stufen Des Organischen under Der Mensch. Einleitung in Die Philosophische Anthropologie* (Frankfurt: Suhrkamp, 1975).

[66] George H. Mead, *Mind, Self & Society. From the Standpoint of a Social Behaviorist* (edited, with introduction by Charles W. Morris; Chicago - Illinois: The University of Chicago Press, 1959/1934). This implies that self and other emerge similtaneous; there is no unmediated access to a pre-existent self via introspection. We don't need a theory of other minds, because the constitution of our own mind depends on the mutual double anticipation. It does not follow that we never develop a theory of another's mind though; it merely follows that such rationalization does not grasp the initial event of self-constitution via the gaze of the other.

[67] Cp. Teubner's (n 4 above) reference to the 'double contingency' as the basis on which persons interact: attributing the self-referentiality one experiences as one's sense of self to another implies identifying the other as a person.

[68] The study of the role of emotion in cognition is related to the study of the role of embodiment in artificial intelligence, after Dreyffus phenomenological criticism in the beginning of the 1970s. Cf. Hubert L. Dreyfus, *What Computers Still Can't Do : A Critique of Artificial Reason* (Rev edn.; Cambridge, Mass. ; London: MIT Press, 1992) liii, 354 p. From the perspective of neuroscience Antonio R. Damasio, *Looking for Spinoza : Joy, Sorrow, and the Feeling Brain* (1st edn.; Orlando, Fla.: Harcourt, 2003) x, 355 p.

[69] Neuroscience has disclosed some of the hardwiring of this capacity in mirror-neurons. Note that if this affords empathy, it is not an exclusively human capability. Shaun Gallagher and Dan Zahavi, *The Phenomenological Mind : An Introduction to Philosophy of Mind and Cognitive Science* (London ; New York: Routledge, 2008) vii, 244 p, Stephanie D. Preston and Frans B. M. De Waal, 'Empathy: Its Ultimate and Proximate Bases', *Behavioral and Brain Sciences,* 25/01 (2001), 1-20.

effects', which depends on patiency, 'the capacity to be acted upon in ways that can be evaluated as good or evil'.[70] The question of moral agency now involves the question of whether the agent is capable of consciously acknowledging the suffering of the patient as a consequence of her own actions.

There are, obviously, other issues at stake: did the agent have access to an alternative course of action? If not, we like to think that she is absolved from moral responsibility and excused from criminal liability. My point is that the crucial difference between human agents and the type of smart environments envisaged by Karnow does not reside in human agents being free to decide which action to undertake, whereas machines are determined to act as they do by the algorithms we inscribed in them. As we have seen this is simply not the case in as far as autonomic computing systems develop emergent behaviors. Also, we must admit that many of our own actions are automated to a much further extent than we may like to acknowledge, raising questions about the type of freedom we exercise in behaving in one way or another.[71] The more salient difference is that we are capable of owning up to our behavior as its author *after the fact*, if we are addressed as such by another who suffered or enjoyed the consequences of our action.[72] This retroactive exercise is preconditional for deliberate actions that require the articulation of an intention *before the (f)act*. It means that we are not merely autonomic but also autonomous to the extent that we can anticipate, plan and understand our own actions in the light of a normative framework. This normative landscape may not be of our own making, but in being articulated in a language that habitually generates ambiguity the norms inevitably allow for resistance and non-compliance. Autonomy means that we have the capability to contest the way others read our action by engaging in a discourse on its

---

[70] Gray, Kurt and Wegner, Daniel M. (2009), 'Moral Typecasting: Divergent Perceptions of Moral Agents and Moral Patients', *Journal of Personality and Social Psychology,* 96 (3), at 505-506. Their concept of a patient differs from that of Floridi and Sanders whose level of abstraction includes patients without feelings. For Floridi and Sanders any discrete, self-contained, encapsulated package containing data structures and sets of operations, procedures is a patient. To damage such a package is evil, because it increases entropy, cf Floridi and Sanders (n 58 above).

[71] This may seem to refer to the discussion of compatibilism, see eg Daniel M. Wegner, *The Illusion of Conscious Will* (Cambridge, Mass.: MIT Press, 2002); Stephen J. Morse, 'Brain Overclaim Syndrome and Criminal Responsibility: A Diagnostic Note', *Ohio State Journal of Criminal Law,* 3/2 (2006), 397-412 and Matthias Mahlmann, 'Ethics, Law and the Challenge of Cognitive Science', *German Law Journal,* 8/6 (2007), 577-615. I am not convinced, however, that it makes sense to frame the debate in terms of 'whether reasons actually cause our actions', as I am not convinced that it is a Davidsonian agency or a Kantian deontology that needs to be made compatible with (saved in the face of) causal determinism at the level of brain behaviours. See M Hildebrandt, 'Autonomic and Autonomous 'Thinking': Preconditions for Criminal Accountability', in M Hildebrandt and Antoinette Rouvroy (eds.), *Autonomic Computing and Transformations of Human Agency. Philosophers of Law Meeting Philosophers of Technology* (Routledge (in review), forthcoming-b). In this case I merely aim to refer to the fact that most of our behaviours are automated, even if we could turn around and reflect on them, see Himma (n 60) eg at 25.

[72] This is a point made by Butler in a pertinent discussion of what it means to give an account of oneself in the face of the constitutive opacity that grounds us as self-conscious beings. Judith Butler, *Giving an Account of Oneself* (1st edn.; New York: Fordham University Press, 2005).

meaning.[73] Human language turns communication – as Zizek points out – into a succesful misunderstanding, instead of the straightforward exchange of information that early cybernetics and artificial intelligence aimed for.[74]

How do we stand now with a smart environment that responds *proactively* to our inferred preferences, resetting its goals in function of what it infers to be more in line with the purpose we inscribed in the initial program? Does Alef qualify for the kind of moral agency that affords criminal liability for harm caused and full legal personhood? I can see at least two preconditions that must be met to call Alef to account in any kind of court: first, Alef must be identifiable as a sufficiently stable entity over the course of time and second, the attribution of a causal link to the criminal wrong it is charged with must be reasonable. If these preconditions are met, I see two further conditions to call Alef to account in a criminal court: Alef must be capable of a measure of empathy and it must have developed a type of autonomy that affords intentional action. I don't think it is up to me to decide whether Alef – the fruit of Karnow's imagination – satisfies these conditions, but it seems interesting to end this undertaking with an assessment of what should matter when the time comes to decide whether an infrastructure like Alef qualifies for full legal personhood. As to the first precondition it remains unclear whether a system of distributed, mobile and polymorphous AAs diffuses into the environment in a way that renders its identification as a singly unit of moral or non-moral agency impossible. Perhaps we can draw on Maturana and Varela's concept of autopoiesis,[75] suggesting that to qualify for personhood some kind of emergent selfhood must come about that can be described in terms of organizational closure and structural openness, typical for living organisms. The organizational closure guarantees that identification is possible, whereas the structural openness guarantees adaptability. The second precondition concerns the 'causal failure' that may erupt in the case of hybrid multi-agent systems with emergent behaviors. In articulating this condition in terms of attribution and reasonableness I indicate that causality is not given but needs to be constructed, without thereby suggesting that we can construct it in anyway we like. The attribution must be reasonable, meaning that though there may be many events, behaviors and states of affairs that can be qualified as necessary and/or sufficient conditions for the harm or damage that is at stake, the attribution must make sense in function of the purpose it aims to accomplish. This ties this precondition up with the preceding issue (does it make sense to qualify an event as the behavior of Alef) as well as with the

---

[73] Cf. Paul Ricœur and John B. Thompson, *Hermeneutics and the Human Sciences : Essays on Language, Action, and Interpretation* (Cambridge New York Paris: Cambridge University Press. Editions de la Maison des sciences de l'homme, 1981). There is an affinity with the idea that human persons can develop second order desires and beliefs to govern their first order desires and beliefs, cf eg Calverley (n 63 above) referring to H. Frankfurt, Alternate possibilities and moral responsibility, in: H. Frankfurt (ed), *The importance of what we care about,* Cambridge University Press (1988).
[74] Zizek, Slavoj (1991), *Looking awry: an introduction to Jacques Lacan through popular culture* (Cambridge, Mass.: MIT Press). For a salient history of cybernetics see Hayles, N. Katherine (1999), *How we became posthuman. Virtual bodies in cybernetics, literature, and informatics* (Chicago: University of Chicago Press).
[75] Maturna and Varela (n 51 above).

issues of empathy and autonomy (a causal link must be attributed to the action of an entity that can be blamed for wrongful behavior). Actually the breakdown of causality may come to a hold when Alef develops into an entity that sustains its identity over and against its environment, taking on self-maintenance as a separate and overruling goal to be achieved. In fact, I would dare to advance the idea that this development may be speeded up if we begin to address Alef as a unit of decision-making, as a self. However, whether this actually leads to such a unity evidently depends on the extent to which Alef's hard- and soft-wiring affords this.

Well then, if we imagine an Alef that can be addressed as a stable identity over the course of time, the remaining point is whether Alef is capable of owning up to the consequences of its behavior: is it possible to address Alef as a self that is capable of reflecting upon its actions as its own actions? Can we expect Alef to take the role of the other and experience the suffering it has caused? This finally raises the issue of whether Alef has any experience of suffering: does it share the vulnerability of a sentient being? This is not merely an issue of culpability. It first regards the wrongfulness of the behavior that has been identified as a cause of unjustified harm. Only after the wrongfulness has been established do we get to the issue of culpability. An entity that has no experience of joy or suffering cannot understand what is wrong with hurting another, unless this 'understanding' is programmed into its system – but I contend that such would not be the kind of understanding that is needed for criminal liability. We enter troubled waters now, or rather Searle's Chinese Room argument.[76] To be called to account for a criminal offense an AA must be able to make sense of what it did to another, it must be able to grasp the meaning of its behavior for another sentient being, rather than merely being capable of manipulating machine-readable data. To take part in the communicative process of a criminal trial, Alef's perception must enter a semantic universe that is different from a syntactical construct and more than the play of emotions. Machines are capable of manipulating symbols, and animals are capable of experiencing their own and others' emotions. Neither seems sufficient for moral responsibility. The empathy that is typical for the kind of personhood warranted for criminal liability involves the capability to not only experience (mirror) the suffering of another, but to also find words for it, to engage in a conversation, to probe the meaning of the experience or the lack thereof. This is what enables a reflection on one's own role in bringing about the harm for which one is addressed as the author. This is what grants a person a measure of autonomy, even if there was no way to avoid the harm caused. Autonomy is less a matter of having alternative courses of action than a matter of deciding on the meaning of one's action, in view of how one is addressed. It concerns the possibility to either contest the claim that is brought forward or to accept the responsibility. This will create room for an alternative course of action in the future, or for an alternative understanding of one's own behavior. We must concede that Alef has a capacity to govern itself, to live by its own law; autonomic derives from the Greek for *auto* (self) and *nomos* (law). It is not

[76] Solum (n 23 above), at 1282-1283; John Searle, 'Minds, Brains, and Programs', *Behavioral and Brain Sciences,* 3/3 (1980), 517-457.

altogether unthinkable that Alef will be enhanced with synthetic emotions to improve its performance, but it is as yet unclear at which point all this will give rise to something similar to the self-consciousness of human beings. Are we capable of judging whether a proactive artificial life form has developed a self-consciousness? Many authors point out that smart robots already invoke a kind of *mutual double anticipation*, for instance generating protective feelings for Sony's robot pet AIBO.[77] Teubner suggests that the attribution of personhood can be an adequate way of dealing with the uncertainty of what or who we are dealing with; identifying Alef as an artificial person could resolve the issue of whether it is a 'real' person or a computer from Searle's Chinese room. Assuming that Alef – like us – has alternative courses of action that are not determined but contingent upon many factors, may help us to perform the transition from *using* a smart technology to *interacting* with it. Providing for full legal personhood would confirm the status of moral personhood that would allow us to call Alef to account for its wrongful actions. The question remains whether addressing Alef as a sentient, self-conscious artificial person will indeed produce a novel type of personhood that justifies the assumption of personhood after the fact. Or could it be that attributing full legal personhood to what are really mind-less agents will erode the meaning of criminal liability, slowly erasing the difference between punishment and discipline?

5. Concluding remarks

The spread of smart applications touches the foundations of the criminal law, notably causality, wrongfulness and legal personhood. First, distributed multi-agent systems form hybrid networks that exhibit emergent behaviours that cannot be attributed to either one of the agents or explained in terms of an aggregation of actions. This makes it hard to discriminate which action actually caused the harm or the damage that would normally we redressed by the criminal law. Second, it is hard to imagine that a smart environment or infrastructure itself becomes the culprit of a criminal charge, but at some point we may have to concede that the self-management that was inscribed in their programs has actually generated a self that should be called to account for its actions. This would require a measure of empathy and a capability to reflect on the meaning of one's actions. If artificial life forms develop they would have to share our vulnerability and the ambiguity of our language for us to call them to account in a criminal court.

---

[77] See http://www.robothalloffame.org/06inductees/aibo.html: 'In Japanese, AIBO means 'companion'. In English, AIBO is an acronym for 'Artificial Intelligence BOT' (…) AIBO represents a new class of robot – relatively cheap, highly compact, and very stable, with his four legged motion. AIBO can see, hear, and understand commands (showing true dog-like behavior, it is also programmed to occasionally ignore them). AIBO has the ability to learn, to adapt to its environment, and to express emotion. AIBO sees in color, hears in stereo, and feels objects with its feet. It has grown more sophisticated over the years as new features have been added'.

6. References

Allen, R. and Widdison, R. (1996), 'Can Computers Make Contracts?', *Harvard Journal of Law & Technology,* 9 (1), 25-52.

Berlin, Isaiah (1969/1958), 'Two concepts of liberty', in Isaiah Berlin (ed.), *Four essays on liberty* (Oxford New York: Oxford University Press), 118-73.

Butler, Judith (2005), *Giving an account of oneself* (1st edn.; New York: Fordham University Press) x, 149 p.

Calverley, David J. (2008), 'Imagining a non-biological machine as a legal person', *Artificial Intelligence & Society,* 22 (4), 523-37.

Calverley, David J. (2008), 'Imagining a non-biological machine as a legal person', *Artificial Intelligence & Society,* 22 (4), 523-37.

Cevenini, Claudia (2004), 'Some Criminal Law Considerations on Electronic Agents', in C. Cevenini (ed.), *The Law and Electronic Agents: Proceedings of the LEA 04 workshop*, 171-80.

Chopra, Samir and White, Laurens (2004), 'Artificial Agents: Personhood in Law and Philosophy', *Proceedings of the European Conference on Artificial Intelligence* (IOS Press), 635-39.

Clark, Andy (2003), *Natural-Born Cyborgs. Minds, Technologies, and the Future of Human Intelligence* (Oxford: Oxford University Press).

Clark, Andy (2003), *Natural-Born Cyborgs. Minds, Technologies, and the Future of Human Intelligence* (Oxford: Oxford University Press).

Cohen, Carl and Regan, Tom (2001), *The Animal Rights Debate* (Rowman & Littlefield Publishers).

Damasio, Antonio R. (2003), *Looking for Spinoza : joy, sorrow, and the feeling brain* (1st edn.; Orlando, Fla.: Harcourt) x, 355 p.

Dreyfus, Hubert L. (1992), *What computers still can't do : a critique of artificial reason* (Rev edn.; Cambridge, Mass. ; London: MIT Press) liii, 354 p.

Editions de la Maison des sciences de l'homme) viii, 314 p.

Floridi, Luciano and Sanders, J.T. (2001), 'Artificial Evil and the Foundation of Computer Ethics ', *Ethics and Inf. Technol.,* 3 (1), 55-66.

Floridi, Luciano and Sanders, J.W. (2004), 'On the Morality of Artificial Agents', *Minds and Machines,* 14 (3), 349-79.

French, Peter A. (1979), 'The Corporation as a Moral Person', *American Philosophical Quarterly,* 16 (3), 207-15.

Gallagher, Shaun and Zahavi, Dan (2008), *The phenomenological mind : an introduction to philosophy of mind and cognitive science* (London ; New York: Routledge) vii, 244 p.

Garreau, Joel (2005), *Radical Evolution. The Promise and Peril of Enhancing our Minds, our Bodies - and What it Means to be Human* (New York: Doubleday).

Gray, Kurt and Wegner, Daniel M. (2009), 'Moral Typecasting: Divergent Perceptions of Moral Agents and Moral Patients', *Journal of Personality and Social Psychology,* 96 (3), 505-20.

Hayles, N. Katherine (1999), *How we became posthuman. Virtual bodies in cybernetics, literature, and informatics* (Chicago: University of Chicago Press).

Hegel, Georg Wilhelm Friedrich, Wood, Allen W., and Nisbet, Hugh Barr (1991), *Elements of the philosophy of right* (Cambridge texts in the history of political thought; Cambridge [England] ; New York: Cambridge University Press) lii, 514 p.

Hildebrandt, M (forthcoming-a), 'Law at a Crossroads: Losing the Thread of Regaining Control. The Collapse of Distance in Real Time Computing', in M. Goodwin, R. Leenes, and B.J. Koops (eds.), *Tilting Perspectives on Regulating Technologies*.

--- (forthcoming-b), 'Autonomic and autonomous 'thinking': preconditions for criminal accountability', in M Hildebrandt and Antoinette Rouvroy (eds.), *Autonomic Computing and Transformations of Human Agency. Philosophers of Law meeting Philosophers of Technology* (Routledge (in review)).

Hildebrandt, M. (2008), 'Ambient Intelligence, Criminal Liability and Democracy', *Criminal Law and Philosphy,* 2 (2), 163-80.

Himma, Kenneth Einar (2009), 'Artificial agency, consciousness, and the criteria for moral agency: what properties must an artificial agent have to be a moral agent?', *Ethics and Inf. Technol.,* 11 (1), 19-29.

Huff, Toby E. (2003), *The Rise of Early Modern Science. Islam, China, and the West, second edition* (Cambridge UK: Cambridge University Press).

Karnow, C.E.A. (1994), 'The Encrypted Self: Fleshing Out the Rights of Electronic Personalities', *Journal of Computer & Information Law,* XIII (1), 1-16.

--- (1996), 'Liability for Distributed Artificial Intelligences', *Berkely Technology Law Journal,* 11, 148-204.

Kephart, Jeffrey O. and Chess, David M. (2003), 'The Vision of Autonomic Computing', *Computer,* (January).

Mahlmann, Matthias (2007), 'Ethics, Law and the challenge of cognitive science', *German Law Journal,* 8 (6), 577-615.

Maturana, H.R. and Varela, F.J. (1998), *The Tree of Knowledge. The Biological Roots of Human Understanding* (Boston & London: Shambhala).

Mead, George H. (1959/1934), *Mind, Self & Society. From the standpoint of a social behaviorist* (edited, with introduction by Charles W. Morris edn.; Chicago - Illinois: The University of Chicago Press).

Morse, Stephen J. (2006), 'Brain Overclaim Syndrome and Criminal Responsibility: A Diagnostic Note', *Ohio State Journal of Criminal Law,* 3 (2), 397-412.

Morse, Stephen J. (2006), 'Brain Overclaim Syndrome and Criminal Responsibility: A Diagnostic Note', *Ohio State Journal of Criminal Law,* 3 (2), 397-412.

Paris: Cambridge University Press ;

Plessner, Helmuth (1975), *Die Stufen des Organischen under der Mensch. Einleitung in die philosophische Anthropologie* (Frankfurt: Suhrkamp).

Preston, Stephanie D. and de Waal, Frans B. M. (2001), 'Empathy: Its ultimate and proximate bases', *Behavioral and Brain Sciences,* 25 (01), 1-20.

Prigogine, Ilya and Stengers, Isabelle (1984), *Order out of Chaos.* (New York: Bantam Books).

Ricœur, Paul and Thompson, John B. (1981), *Hermeneutics and the human sciences : essays on language, action, and interpretation* (Cambridge [Eng.] ; New York

Sartor, Giovanni (2002), 'Agents in Cyberlaw', in Giovanni Sartor (ed.), *The Law of Electronic Agents: Selected Revised Papers. Proceedings of the Workshop on the Law of Electronic Agents (LEA 2002)* (Bologna: CIRSFID Universita di Bologna), 3-12.

Searle, John (1980), 'Minds, brains, and programs', *Behavioral and Brain Sciences,* 3 (3), 517-457.

Searle, John (1995), *The Construction of Social Reality* (New York: The Free Press).

Solum, Lawrence B. (1992), 'Legal Personhood for Artificial Intelligences', *North Carolina Law Review,* 70 (2), 1231-87.

Stone, Christoffer (1972), 'Should Trees Have Standing? - Toward Legal Rights for Natural Objects', *University of Southern California Law Review,* 45, 450-.

Tapscott, Don (2009), *Grown up digital : how the net generation is changing your world* (New York: McGraw-Hill) xvi, 368 p.

Teubner, Günther (2007), 'Rights of Non-humans? Electronic Agents and Animals as New Actors in Politics and Law', *Max Weber Lecture Series 2007/04* (European University Institute).

Teubner, Günther (2007), 'Rights of Non-humans? Electronic Agents and Animals as New Actors in Politics and Law', *Max Weber Lecture Series 2007/04* (European University Institute).

Vallverdu, Jordi and Casacuberta, David (2008), 'The Panic Room: On Synthetic Emotions', in Adam Briggle, Katinka Waelbers, and Philip Brey (eds.), *Current Issues in Computing and Philosophy* (Amsterdam: IOS Press), 103-15.

Warwick, Kevin (2003), 'Cyborg Morals, Cyborg Value, Cyborg Ethics', *Ethics and Information Technology,* 5, p. 131-37.

Wegner, Daniel M. (2002), *The illusion of conscious will* (Cambridge, Mass.: MIT Press) xi, 405 p.

Wettig, S. and Zehender, E. (2004), 'A legal analysis of human and electronic agents', *Artificial Intelligence and Law,* 12 (1-2), 111-35