2014

Eccentric Positionally as a Precondition for the Criminal Liability for Artificial Life Forms

Mireille Hildebrandt, *Radboud University Nijmegen*

Draft

Please refer to the final text in:

Jos de Mul (ed.), *Plessner's Philosophical Anthropology. Perspectives and Prospects*. Amsterdam: Amsterdam University Press, 407-424.

Eccentric positionality as a precondition for the criminal liability of artificial life forms

Mireille Hildebrandt[*] (Erasmus University Rotterdam, Vrije Universiteit Brussel)

Abstract

This contribution explores Plessner's distinction between animal centricity and human eccentricity as "a difference that makes a difference" for the attribution of criminal liability to artificial life forms (ALFs). Building on the work of Steels and Bourgine & Varela on artificial life and Matura & Varela's notion of autopoiesis I will reason that even if ALFs are autonomous in the sense even of having the capacity to rewrite their own program, this in itself is not enough to understand them as autonomous in the sense of instantiating an eccentric position that allows for reflection on their actions as their own actions. Evidently this also means that to the extent that ALFs do develop some sort of conscious self-reflection they would in principle qualify for the censure of the criminal law.

ALFs: automatic and autonomous "agents"

Building on Maturana & Varela (1998), Bourgine & Varela (1992:xi) have defined artifical life (AL) as "a research program concerned with autonomous systems, their characterization and specific modes of viability". Their working definition stands in contrast to mainstream attempts to define AL in terms of the extraction of necessary and sufficient conditions of biological life forms, which are then applied to non-biological entities, hoping to thus create life. In opting for a focus on the autonomy of living systems they emphasize autonomy as the most salient feature of life, which they further define by the constitutive capacities for viability, abduction and adaptability. *Viability* regards the capacity to respond to unpredictable changes in the environment in a manner that allows the system to maintain its organisational identity (implying operational closure), for instance by changing its internal structure (often entailing structural coupling with other living systems within the environment). To anticipate changes and to respond to unanticipated change, living systems need to function as *abduction machines*, producing sets of responses that sustain the unity of the system. *Adaptability* implies that the internal restructuring adequately fits with the challenges produced by the external environment, without annihilating the organisational identity of the system. Clearly, Bourgine & Varela do not consider centricity let alone eccentricity a necessary condition. It is important to note that for them neither natural nor artificial life forms presume eccentricity or even centricity to qualify as living systems, whereas they do require a measure of autonomy (even if this may not be a sufficient condition).[i] In terms of Plessner one could associate the way Bourgine & Varela as well as Maturana & Varela define living systems with Plessner's notion of the border or boundary and with his concept of positionality. In both cases the difference between on the one hand the border between a non-living thing and its environment and on the other hand the border between a living system and its environment is that in the first case the border is *not* part of either the environment or the thing, whereas in the second case the border is actively created and maintained by the living system of which it is also a part. Interestingly, both Plessner and Varela consider living entities to be systems, meaning that the identity of the entity derives from the productive interrelations between its components. Last but not least both speak of the capacity for selfregulation (*Selbstregulierbarheit*, Plessner 1975:160-165; autopoiesis, Maturana & Varela 1998:47-8) as crucial for the living, even if Plessner may understand this as a transcendental category rather than as an observable.

To further investigate the notion of autonomy we can use Steels' distinction between automatic and autonomous agents. For a start Steels defines agents as systems, meaning "a set of elements which have a particular relation amongst themselves and with the environment" (Steels 1995:1). Second, Steels defines agents as performing a particular function for another agent or system, and third, he stipulates that agents are systems capable of maintaining themselves. Agents thus come close to living systems, which operate on the basis of two mechanisms (Steels 1995:2):

> + They continuously replace their components and that way secure existence in the face of unreliable or short-lived components. The individual components of the system therefore do not matter, only the roles they play.[ii]
>
> + The system as a whole adapts/evolves to remain viable even if the environment changes, which is bound to happen.

Steels then continues to explain the meaning of autonomy in terms of its etymological background, which stems from the self-government of the Greek city states: *autos* (self) and *nomos* (rule or law). Those who live in accordance with their own law are autonomous, whereas those obedient to an external law are not. He contrasts this with the term automatic, which derives from *cybernetic* or self-steering. This implies that "automatic systems are self-regulating, but they do not make the laws that their regulatory activities seek to satify" (Steels 1995:4).[iii] Autonomous systems are self-governing as well as self-regulating, whereas automatic systems are merely self-regulating. Old style artificial intelligences (AIs) are automatic to the extent that they cannot step outside the boundaries of their original design, whereas living systems are autonomous because they have not been built and programmed by others and are capable of self-organisation, development, adaptation and learning in order to sustain their viability in a changing environment. The difference between AI and AL, therefore, seems to be that between automatic systems and autonomous systems. In speaking of ALFs instead of Als I wish to prevent equating AL with humanoid robots or other imitations of human beings, since I expect that artificial life to develop from and into life forms entirely different from our own. ALFs might for instance emerge from distributed polymorphous multi-agent systems that would present us with novel life forms that are not necesseary contained within the "skin" of a robot, and it may be difficult at this point to even imagine "them" as identifiable entities.

Autonomic computing

Having differentiated between automatic and autonomous systems I will now introduce a further distinction, namely that of autonomic systems. The reason for this is that the autonomy of living systems, as defined by Varela and by Steels, does not equate with the notion of human autonomy that is assumed by our legal framework. If criminal liability requires a measure of autonomy we will need a further or other definition of autonomy, next to or over and above the one that adheres in all living systems.

IBM has introduced the notion of autonomic computing (Kephart and Chess 2003), using the autonomic nervous system as a metaphor, to describe computing systems capable of self-management (self-repair, self-maintenance, self-configuration). It strikes me as not altogether improbable that by designing a system that is capable of managing itself, a self may actually emerge in the process of defining itself and of actively maintaining its borders. This would imply that even if these systems are programmed to achieve goals that we have set for them, the goal of self-management may at some point gain prominence. This relates to the third criterion that Steels introduced in his definition of autonomous agents; to be an agent rather than merely an instrument, a system must have an own interest in maintaining itself. To the extent that autonomic computing systems achieve such a level of autonomy, they could thus qualify as ALFs.[iv]

In 1995 Steels concluded that no robots at that point could be said to be autonomous in the sense he described. To decide whether autonomic computing systems are autonomous in the above sense will require empirical investigation. Since autonomic computing is still in its infancy and a vision rather than a reality, empirical evidence at this moment cannot provide very precise answers. However, it makes sense to anticipate the consequences of autonomic systems in the case that they do become autonomous in the above sense, for instance because alternative designs of these systems may generate different legal and ethical consequences. For all practical purposes I will understand "truly" autonomic computing systems as a novel type of ALF, different from the usual suspects such as robots or webbots.

One of the consequences of computing systems that function as an agent for their human users by sustaining their own identity, will be a fundamental unpredictability. This unpredictability is "caused" by the complexity of the interactions of the components (autonomic systems will definitively be multi-agent systems), as well as by their polymorphous character (to achieve their goals they may change their structure). The uncertainty brought about by the emergent properties of multi-agent interactions relates to the *behaviour* of the emergent system, while the uncertainty brought about by structural changes of the components of the system relates to the *identity* of the system that is behaving. In both cases serious problems may occur if the behaviour of the system results in harm or damage: first, it may be difficult to attribute causality to either individual nodes of the system or to its designer (thus requesting us to focus on the behaviour of the ALF) and second, it may be difficult to identify the system as the object of attribution (how can we identify a particular ALF?). A third, even more challenging issue, is the question of fault: can we censure the system, hold the ALF accountable in its own right? Within the framework of the criminal law, liability depends on wrongfulness and culpability and this seems a bridge too far even if causality could be attributed to an identifiable autonomous computing system.

This raises many questions, with regard to causality, wrongfulness and culpability. In this contribution I will focus on the issue of what it would take to actually blame an ALF for its behaviour. At which point does an ALF qualify as an autonomous system that can be held accountable in a court of law? I believe that Plessner's distinction between a centric and an eccentric positionality provides a salient conceptual tool to make the difference between autonomy as a characteristic of all autonomic systems and the autonomy that is – so far - specific for human agency. But before embarking on this crucial difference it is important to pay attention to the uncentric positionality of plants, if only to keep in mind that life forms are not even necessarily centric.

Within the realm of living organisms Plessner distinguishes between three life forms: plants, animals and humans. Life forms are characterized by their positionality, which constitutes the manner in which they relate to their spatial and temporal environment and to their self. In categorizing these life forms Plessner does not deny their empirical overlappings, he rather emphasizes the need for analytical distinctions that allow one to better understand what kind of transitions are at stake between different life forms. The extent to which he adheres to a transcendental perspective is not what interests me here, though it may be interesting to investigate the implications of his interpretation of Kant at this point.

According to Plessner, plants have an open form of organisation, responding to their environment in a way that is less mediated by the internal complexities of the organism. Their inner workings are not centralised as in animals with a nervous cord or central nervous system. He therefore coins plants as dividuals instead of individuals. There is a kind of equivalence between the nodes of the plant that is transformed to a more hierarchical system of components in animals that allows for a greater differentiation between different parts of the animal, but also makes for a less flexible or open structure. The most salient feature of plants as compared to animals is the lack of a centre of organisation. In plants, perception is not enacted via a set of specific organs (eyes, ears, skin, nose) that produce a unified experience thanks to the way the central nervous system responds to the imprint of events in the environment. In fact the whole idea of enaction – coined as such by Varela, but present in Plessner's work – does not apply to plants (Plessner 1975:225):[v]

> Experience and action (i.e. centrally mediated movements, which can be modified by associations) go against the grain of the open form.

Animal life forms entail a centralised organisation of action/perception, afforded by the central nervous system that creates an awareness of an embodied self that is both the material body (*Körper*) and its central representation as an organised unity (*Leib*).[vi] The transition from plant to animal seems to afford the birth of a "unified" self that confronts the environment in a frontal manner instead of confronting it on all sides in all its components. The difference is not that a plant does not sustain an identifiable organisation with a specific repertoire of strategies to deal with environmental change. The difference is that the organisation of plants is not hierarchical in the sense that their perceptions and their decisions to act are not mediated in a central point, making it hard to even speak of a plant's actions when referring to their movements (eg growth). To read a centralized plan into the development of a plant would be a mistake (Plessner 1975:226):

> It would be a betrayal of the essence of plants (as it would be a betrayal of nature), to understand them as symbolic, as the embodiment of a principle that expresses itself in them, as the articulation of a force, a soul, a reality, that is no longer themselves (my translation).[vii]

In exploring the idea of artificial life forms we must take this warning to heart. Apart from the fact that smart technologies may not qualify as life forms at all, for instance because they are heteronomous in their design and mechanistic in their operation (self-regulating but not self-governing), smart technologies could theoretically display an uncentric positionality comparable to plants, rather than providing a virtual core that springs from a centralized organisation of perception and action.[viii]

According to Plessner, animal positionality is defined by centricity and frontality. Crucially, the animal has a body and is a body, but this dualism is not understood in a Cartesian manner (discriminating two substances). Plessner speaks of a double aspectivity that creates a distance between the *Körper* (the body an animal experiences as having) and the *Leib* (the body an animal experiences as being), suggesting that this distance is productive in allowing for a dynamic representation of the self (*Leib*) in relation to its body (*Körper*) in its environment. The productive split between these two bodies is brought about by the centric positionality of animals and pertains equally to human beings, who share the dynamics of becoming an individual in front of an environment that is constituted as such in the con-frontation with a unified self. The representation that is made possible by the centric organisation of animals, which is also what allows them to learn from past experience, aligning their findings in the past with anticipation of future occurrences (Plessner: 277-287). One can explain this by highlighting that representation does not necessarily refer to symbolic representation, as it can merely denote the imprint that is made on the perceptive/enactive structure of the animal. In other words, as Steels (1995:7) observes:

> Representations are physical structures (for example electro-chemical states) which have correlations with aspects of the environment and thus have a predictive power for the system. These correlations are maintained by processes which are themselves quite complex and indirect, for example sensors and actuators which act as transducers of energy of one form into energy of another form. Representations support processes that in turn influence behavior. *What makes representations unique is that processes operating over representations can have their own dynamics independently of the dynamics of the world that they represent* [my italics, mh].

In Plessner's analysis, the difference between animals and humans resides in the fact that humans have developed an awareness of the distance between *Leib* and *Körper*. In using language, which allows a person to address what is not here and what is not now, human beings have developed the capacity to decenter their position in space and time, "liberating" themselves from the here and now that holds together the animal self (Plessner 1975:239):

> To the animal his here-and-now character is not given, not present, it emerges in him and in that way carries the hidden barrier against his own individual Existence. Indeed it (the *Leib*) is present to itself (the whole), but the whole is not present to it. Present to it is the outer world and the *Körperleib* (my translation).[ix]

Other than the animal the human life form is not enclosed in the *Umwelt*-channel of the here and now. The animal is conscious of its environment and of its own body *as* its own body, but it is not conscious of being conscious because it cannot escape what is present – it cannot leave its specific *Umwelt-tunnel* (Cheung 2006:321). In phenomenological terms there is intentionality (consciousness of something), but this intentionality regards the outside world as it appears in the act of perception.[x] The animal is a self but not an "I" (Plessner:238), because the "I" depends on an eccentric positionality, which is made possible only in the use of human language.[xi] I sometimes use the example of a child to illustrate this crucial point. If I address a small infant by pointing at her and saying "you are Charlotte" and pointing back at myself, saying "I am Mireille", the child will initially imitate me and - pointing to herself – repeat: "you Charlotte", and – pointing to me – repeat: "I Mireille". My response ("no, you Charlotte, I

Mireille") will not catch on until the day that Charlotte suddenly realises that she is "you" for me, whereas she is "I" when speaking out for herself. This capacity to turn back on herself from the position of the other is afforded by language and constitutes the birth of the eccentric positionality in a human person. Language also affords one to present what is absent (in time or space) and this opens the door for the imagination, providing an entirely new "infrastructure" for the representation of our selves and our environment, creating the possibility to reinvent our selves as well as each other. The absence, lack or emptiness that animals cannot conceive of,[xii] is the precondition for humans to move from the *Umwelt* of ostensive reference to the *Welt* brought about by the use of language (Ricoeur 1976: chapter 4).

Eccentricity thus refers to a centric self that is consciously aware of itself (*Selbstdistanz*), looking back at it-self from the position of the other(s), thus decentering the self and introducing the position of the observer. This doubling of the double aspectivity is what enables the constitution of an outside world, an inner world and a middle world (*Mitwelt*).[xiii] The *Umfeld* is replaced by an outside world that is filled with things that are not merely perceived from the centric and frontal position of a self, but can be observed from the position of other selves or from that of the same self in another time and place. The body – one's own body – can now appear as one of these things that crowd an outside world. At the same time the self looks back upon itself via the gaze of the other (Mead 159/1934), thus instantiating the "I" that is performing this act of constitution, creating a view of the self as another (Ricoeur 1992, Waldenfels 2004) and thus instauring an inner world (Plessner 1975:295):

> In the distance to himself the living being is given as its inner world (my translation).[xiv]

Finally, taking the position of the other (Mead 1959, Merleau-Ponty 1945, Ricoeur 1992, Butler 2005) enables *a double anticipation* (Hildebrandt, Koops et al. 2008); it enables us to anticipate what others expect from us; it enables us to anticipate how others will interpret our behaviours. In a way it precedes the birth of the outside world and the inner world; to be consciously aware of oneself as a self one has to be addressed as such by another and to construct one's identity as a particular person one must be drawn into a web of meaning that can be molded into one that signifies such identification (Plessner 1975:303):

> The Mitwelt carries the person, while at the same time it is carried and built by him (my translation).[xv]

The position of the observer and human autonomy

From this analysis Plessner moves on to the articulation of three anthropological Constitutions: the law of natural artificiality; the law of mediated immediacy and the law of the utopian position. The double aspectivity that is typical for the human person involves an inescapable need to reconstruct the self, the world and the others, thus entailing an unavoidable artificiality of the *Körperleib* that grounds us. This natural artificiality is not lamentable, it should be celebrated because it is the result of a freedom that is productive of and produced by the human life form. But is also stands for a rootlessness, an uncertainty and an awareness that we have no immediate access to either the self, the world or our fellows. That is how we can envy the animal that is caught up in the immediacy of its *Umfeld*: it may be sad but it lacks the capacity to reflect upon its own sadness. Such reflection creates a distance between us and our feelings that seems alien to a centric position. The upside is that such reflection allows us to think in terms of possibilities, to imagine an outside world that is not (yet) present, to initiate a measure of novelty in the web of meaning that constitutes the *Mitwelt* and to experiment with different selves (roles), thus enlarging the repertoire of strategies to cope with the expected as well as the unexpected (Lévy 1998). Possibility, however, does not equate with unbounded freedom, it refers to a disposition that springs from a reiterated de-centering, a capability to look at things, people and the self from a variety of positions and to thus bring forth a world that is tested from a plurality of points of view, instead of taking the first (own) point of view for granted. At the same time the re-creative nature of language, which allows one to speak of what is not, makes possible genuine novelty - though within the constraints of the particular language used.

Language thus *affords* the human life form to take the position of the observer (the third person singular, or what Mead called "the generalized other"), which – it seems to me - lies at the root of Plessner"s anthropological Constitutions. Being thrown into a language that generates the position of the third person singular humans are forced to be artificial, mediated and utopian; there is no way back to a natural, unmediated access to the here and now. The constitution of the human self thus coincides with its fundamental splitting image: that of the observer who observes herself while observing herself *usw*. Looking at the manner in which Maturana and Varela introduce the position of the observer, I see important differences. First, they connect the position of the observer with the predictability of the behaviour of systems under observation (Maturana & Varela 1998:122-125). This – in itself - does not depend on an eccentric position, since all living systems need to anticipate the behaviour of their environment. Second, they explain that such prediction depends on the interpretation *by the observer* of past and present behaviours of other systems and their respective environments or niches (Maturana & Varela 1980:8-9). The observer's predictions, based on her descriptions of an entity as a unity of interactions with its environment, must not be confused with the way that the system under observation – itself - anticipates changes in its environment.[xvi] This will depend on its own organisation, whereas the observation depends on the observer's organisation. The interactions lie in the cognitive domain of the observed entity, while the causal or other relations between an entity and its niche as observed by the observer lie in the cognitive domain of the observer! In fact, Maturana & Varela's observer is a human person, whose description is addressed to another observer, which she may be herself. It seems that language is what not only enables observation, but what equals observation: "Anything said is said by an observer" (Maturana & Varela 1980:8). The most direct reference to what Plessner coins as eccentricity can be found in their statement (Maturana & Varela 1980:8):

> The observer can define himself as an entity by specifying his own domain of interactions; he can always remain an observer of these interactions, which he can treat as independent entities.

What is important here is the claim that an observer – thanks to the use of language – can distance herself from her self and treat her own interactions as if they are disentangled from her self:[xvii] she can take an eccentric position with regard to her self. Some observers would describe this as second order statements about first order interactions. This can be an accurate description if we acknowledge that the second order statement does not contain a view from nowhere; it does not represent a given reality "out there" and cannot make sense if it does not connect to the regularities we encounter (Maturana & Varela 1998:241):

> Again we must walk on the razor's edge, echewing the extremes of representationalism (objectivism) and solipsism (idealism). Our purpose in this book has been to find a *via media*: to understand the regularity of the world we are experiencing at every moment, but without any point of reference independent of ourselves that would give certainty to our descriptions and cognitive assertions. Indeed, the whole mechanism of generating ourselves as describers and observers tells us that our world, as the world which we bring forth in our coexistence with others, will always have precisely that mixture of regularity and mutability, that combination of solidity and shifting sand, so typical for human experience when we look at it up close.

It seems apparent that the position of the observer defines human autonomy as distinct from the autonomy that defines all living systems. The observer entails the difference that makes a difference. In the words of Katherine Hayles (1999:43):

> Although the observer's perceptions construct reality rather than passively perceive it, for Maturana this construction depends on positionality rather than personality. In autopoietic theory, the opposite of objectivism is not subjectivism but relativism.

Though Hayles is not referring to Plessner's usage of the term positionality, her observation is interesting because one could say that Plessner's eccentric positionality similarly does not refer to a psychological analysis but to the fact that the human life form is capable of taking a second and third person perspective. Connecting Plessner and Maturana & Varela one could say that this position accounts for the fact that humans can bring forth a world that (1) reflects the contraints they encounter in their domain of interactions and (2) opens up a plurality of alternative interactions. The connection between being an observer and being capable of prediction implies that the individual observer has a *measure* of choice in how to act, working out the potential consequences of different courses of action as well as anticipating how her fellow-observers will "read" alternative actions. This foresight can be based on Plessner's utopian position that allows an observer to look back from the future, consciously anticipating how her actions will be understood by other observers. The domain of the observer assumes and generates Plessner's *Mitwelt* that carries and supports (*trägt*) the individual person, who is born in the process. There is another interesting analogy between Plessner's work and that of Maturana and Varela. In both cases the human organism is capable of creating a shared world, called *Mitwelt* in the case of Plessner and society in the case of Maturana & Varela. Whereas systems theory, especially in the case of Luhmann, tends to reify this *Mitwelt* or society as a third order unity that effects operational closure in a manner similar to first- and second order organisms (cells and metacellulars), neither Plessner nor Maturana & Varela fall in this trap. In the case of Plessner the *Mitwelt* is underdetermined by the anthropological constitutions, saturated in the ambiguities of the double anticipations that nourish it. In the case of Maturana & Varela (1998:199) they devoted an entire chapter to Social Phenomena, to explain that whereas

> an organism restricts the individual creativity of its component unities, as these unities exist for that organism. The human social system amplifies the individual creativity of its components, as that system exists for these components.

This is why the domain of the observer cannot effect the kind of operational closure that biological organisms must perform to sustain their identity, and that is also why the human life form entails a rootlessness, ambiguity and uncertainty that is at the same time its freedom to perform as an outlier. In that sense human autonomy differs fundamentally from the autonomic nervous system that enables its emergence, as well as from autonomic computing systems that cannot reflect on the meaning of their interactions.

One last point must be made. Maturana & Varela's self-observing observer can be understood as a mechanism that produces second order beliefs about its own first order beliefs. To the extent that autonomic computing entails self-management one could argue that the system generates second order beliefs about its first order beliefs. This, one could argue, implies that autonomic computing systems indeed qualify as eccentric and autonomous systems. The double anticipation that is implied in the eccentric position can be interpreted as an observer's attribution of certain intentions, allowing the observer to adequately infer and predict the behaviour of others. Siding with Dennett (2009), for example, one could claim that the question of whether people or autonomic computing systems "really" have second order beliefs and intentions is a misguided question, since there is no way to anchor the difference. In that case the point is not whether autonomic computing systems develop an eccentric position, but whether we can better anticipate their behaviour if we understand them as rational agents capable of having first and second order beliefs and desires. This is what Dennett has coined as the intentional stance: assuming that another acts on the basis of intentionality because that better explains her behaviours. For may reasons, however, Dennett does not fit well with Plessner. Though I do not think that Plessner would reject the possibility of non-biological life forms developing an eccentric positionality, I do think that he would not be satisfied with ALFs merely "displaying" an intentional stance.[xviii] From the perspective of his semi-transcendental position it makes no sense to attribute intentions – and thus eccentricity – to a system that has no conscious self-awareness. Dennett's behaviourism and physicalism seem not to accord with Plessner's philosophical anthropology.

However, if we "(mis)read" Dennett's position in line with that of behaviourists like Ryle and G.H. Mead I would think that we could end up close to Plessner: what counts is not some metaphysical theory about the essential nature of human beings, but the paradoxical first hand experience of an eccentricity that allows us a permanent distantiation and concurrent reconfiguration of inner, outer and middle world (*Mitwelt*). This is what forces us to develop second order intentions and this allows us to address those we "read" as taking a similar stance in life as responsible for their actions. Taking Dennett seriously in alignment with Plessner I would suggest that endorsing the intentional stance versus an ALF will be an experiment, a way of finding out whether they indeed live up to the expectations that spring from interactions with a person who is capable of developing a mind of her own.

Criminal liability for ALFs?

The difference between centricity and eccentricity is important from the perspective of legal philosophy because it relates to the type of human agency that is presumed in legal notions of accountability, especially in the case of criminal liability. To hold a person to account in a court of law for having committed a crime she must be capable of wrongful action and of being culpable. Before embarking on these requirements we must, however, first establish that ALFs such as autonomic computing systems could cause criminal harm or endanger values and interests that are protected by the criminal law. One could argue that it is never an ALF that commits such crimes but rather its designers or users, the ALF being merely an instrument. However, if we assume that ALFs are autonomous in the sense of being capable of taking decisions that neither the designer (programmer) nor the user could have foreseen it seems obvious that they can in fact cause harm or damage that would fall within the scope of the criminal law if committed by a legal person (a human agent, or for example, an association, company or trust fund to the extent that positive law allows this). It seems to make no sense to attribute the liability to those who could not have foreseen (and thus not have prevented) this action. Note that ALFs will be created precisely because of their capacity to find novel solutions without human intervention.[xix] ALFs will mostly be created as agents that perform specific tasks for an organisation or person who is keen on delegating tasks to such agents; they can be automatic in the case of simple straightforward tasks but when we speak of ALFs we refer to more complex autonomous systems that are capable of coming up with unexpected solutions. The relevant comparison here is not the liability of the producer or the user of a product, but the liability of the owner of an animal for the harm or damage it causes due to its own initiative. This in fact demonstrates that the mere fact that autonomous agents entail a measure of unpredictability does not imply that we can hold them responsible under the criminal law.[xx] Another salient comparison is the notion of "acts of god" like earth quakes, flooding or vulcanic outbursts. Though they cannot be attributed to a particular human person or organisation we cannot call the earth, sea or mountain to account in a court of law.

This is related, as in the case of animals, to the fact that holding a person to account for a criminal offence assumes that she should have been aware of the wrongfulness of her action and moreover can be blamed for having violated the law. In this sense we do not consider animals as persons.[xxi] My use of the word person is somewhat provocative, because we tend to assume that only human animals qualify for personhood. Following Plessner, however, we can understand personhood as the individuality typical for the eccentric position, emerging at the nexus of the inner, outer and middle world (*Mitwelt*), though without any essentialist foundation or final determination: "His existence is truly built on nothing" (Plessner 1975:293, my translation).[xxii] To be a person means to be capable of knowing and willing (idem, also Frankfurt 1971), that is, capable of differentiating between one's urges and one's second order desires about these urges, thus forming an intention and acting upon it. For an ALF to be liable under the criminal law this would imply that four conditions but must fulfilled: first, it must be identifiable as a unity of perception and action (centricity), second, the harm or danger must be attributable to its action, third, it must be foreseeable by the ALF when performing the incriminated action, and fourth, the ALF must be capable of the double anticipation that allows it to anticipate how others will "read" its behaviours (eccentricity). The first condition regards what is called *actus reus* in the criminal law, the second (overlapping with the third) regards the requirement of causality, whereas the third and fourth regard wrongfulness and culpability.

If we want to punish an ALF for actions that fall within the scope of the criminal law, it must also be in a position to exercise its rights of defence; under the rule of law punishment takes place after a fair trial (due process) has taken its course. This is the crucial difference between punishment and discipline, as Hegel saliently suggested when differentiating between punishing a person and training a dog. In a liberal democracy punishment implies more than the *expression* of strong disapproval, requiring instead the *communication* of censure (Duff 2001), which is a two-way process. Punishment – other than discipline or manipulation – assumes an appeal to the double anticipation or whoever has been singled out for its address. This means that a defendant can contest the accusation and reject the charge, not only by denying that she committed the incriminating action, but also by opening a dialogue about the meaning of the legal norm she allegedly violated and the meaning of her action in the light of this norm. Due process and a fair trial thus again assume an eccentric positionality. Punishment is not merely about imposing suffering but also about defining which behaviour counts as criminal. For this reason, finally, a democracy entitles a person who is subject to criminal sanction to being part of the constituency that defines which actions fall within the scope of the criminal law (self-government).

Conclusions

Plessner's notion of human eccentricy has turned out to make a difference when it comes to the attribution of criminal liability. This notion indeed allows us to discriminate between automatic devices, autonomous machines and the type of human autonomy that is preconditional for accountability under the criminal law. Moreover, it enables us to investigate whether artificial life forms could at some point qualify for the censure of the criminal law, as well as the exercise of due process rights and – eventually – participation in defining the contents of the criminal law.

Bibliography

Bourgine, Paul and Francisco J. Varela 1992. Towards a Practice of Autonomous Systems, in Varela, Francisco J. and Paul Bourgine, eds. *Towards a Practice of Autonomous Systems. Proceedings of the First European Conference on Artificial Life.* Cambridge, MA: MIT Press: xi-xviii.

Butler, Judith. 2005. *Giving an account of oneself.* New York: Fordham University Press.

Cheung, T. 2006. The language monopoly: Plessner on apes, humans and expressions. *Language & Communication* 26: 316-30.

Dennett, D. 2009. Intentional Systems Theory in *Oxford Handbook of the Philosophy for Mind*. Oxford: Oxford University Press: 339-50.

Duff, R.A. 2001. *Punishment, Communication, and Community*. Oxford: Oxford University Press.

Floridi, Lucia and Mariarosaria Taddeo. 2009. Turing's Imitation Game: Still an Impossble Challenge for All Machines and Some Judges - An Evaluation of the 2008 Loebner Contest. *Mind and Machines*. 19 (1): 145-50.

Frankfurt, Harry G. 1971. Freedom of the Will and the Concept of a Person. *The Journal of Philosophy* 68 (1): 5-20.

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

Karnow, C.E.A. 1996. Liability for distributed artificial intelligences. *Berkely Technology Law Journal* 11: 148-204

Lévy, Pierre. 1998. *Becoming Virtual. Reality in the Digital Age*. Trans. Robert Bononno. New York and London: Plenum Trade.

Maturana, H.R. and F.J. Varela. 1991. *Autopoiesis and Cognition: The Realization of the Living.* Dordrecht: Reidel.

--- 1998. *The Tree of Knowledge. The Biological Roots of Human Understanding*. Boston & London: Shambhala.

Mead, G.H. [1934] 1959. *Mind, self & society. From the standpoint of a social behaviorist. edited*, edited and with introduction by Charles W. Morris. Chicago - Illinois: The University of Chicago Press.

Plessner, Helmuth. [1928] 1975. *Die Stufen des Organischen under der Mensch. Einleitung in die philosophische Anthropologie*. Berlin: Walter de Gruyter.

Steels, Luc. 1995. When are robots intelligent autonomous agents? *Robotics and Autonomous Systems* 15: 3-9.

Teubner, G. 2007. Rights of non-humans? Electronic agents and animals as new actors in politics and law. In *Max Weber Lecture Series* 2007/04: European University Institute.

Waldenfels, Bernhard. 2004. Bodily experience between selfhood and otherness. *Phenomenology and Cognitive Sciences* 3: 235-248.

[·] Mireille Hildebrandt, LLM, holds a PhD in philosophy of law, she is Associate Professor of Jurisprudence at the Erasmus School of Law in Rotterdam and senior researcher at Law Science Technology and Society (LSTS) Vrije Universiteit Brussels. For some of her publications see: http://works.bepress.com/mireille_hildebrandt/.

[i] Maturana & Varela (1992), at 48 emphasize that living systems are not the only autonomous entities. For instance, societies are autonomous systems, but they differ from organisms in the degree of autonomy of their components (idem, at 198-9). In a similar vein one could suggest that autonomic computing systems could be autonomous in the sense outlined by Maturana & Varela, without necessarily being alive.

[ii] This observation is similar to Maturana and Varela's distinction between organisation and structure: whereas the organisation of a system refers to 'those relations that must exist among the components of a system for it to be a member of a specific class', its structure refers to 'the components and relations that actually constitute a particular unity and make its organization real' (Maturna and Varela 1992, at 47).

[iii] Steels (1995:5) refers to a personal communication from Tim Smithers in September 1992).

[iv] Note that in *The tree of knowledge* (1992:48) Maturana & Varela admit that not all autonomous entities are living beings. Though in *Autopoeisis and cognition* (1980:82) they still held that autopoiesis is both a necessary and a sufficient condition for life, their claim in *The tree of knowledge* is more modest. This would mean that

autonomic computing systems *could* qualify as ALFs, depending on what finally determines whether a system is or is not alive. Note also that consciousness, let alone self-consciousness is not a requirement for an entity to be alive.

[v] "Empfindung und Handlung (d.h. durch Assoziationen modifizierbare, zentral vermittelte Bewegungen) widersprechen dem Wesen offener Form" (my translation).

[vi] Stufen:230-1: "Er [animal organism, mh] ist die uber die einheitliche Repräsentation der Glieder vermittelte Einheit des Körpers, welcher eben dadurch von der zentralen Repräsentation abhangt. Sein Körper ist sein Leib geworden, jene konkrete Mitte, dadurch dat Lebenssubjekt mit dem Umfeld zusammenhangt".

[vii] "Es ist nun einmal ein Verrat am Wesen der Pflanze (wie es ein Verrat am Wesen der Natur ist), sie symbolisch zu nehmen, als Verkorperung eines in ihr sich aussprechenden Prinzips, als Ausdruck einer Kraft, einer Seele, einer Wirklichkeit, die nicht mehr sie selbst ist."

[viii] I expect that smart technologies that qualify as autonomous according to Steels' definition will exhibit the kind of distributed intelligence that is explained by connectionist models of the brain. This raises the question which entity we are talking about when referring to an ALF: the autonomic computing system itself (the brains) or the system that it embodies, nourishes and produces (the entire body).

[ix] "Dem Tier is sein Hier-Jezt-Charakter nicht gegeben, nicht gegenwärtig, es geht noch in him auf und trägt darin die ihm selbst verborgene Schranke gegen seine eigene individuele Existenz. Wohl ist es (als Leib) ihm (dem Ganzen), nicht aber das Ganze sich gegenwärtig. Ihm gegenwärtig ist Außenfeld und Körperleib."

[x] Cf. Waldenfels (2004:237-239) on intentionality as consciousness *of something as something* ("nothing is given which is not given *as such*").

[xi] On Plessner's language concept in relation to the human life form, see Cheung (2006).

[xii] Stufen:271. Plessner speaks of 'ihm verschlossene Anschauungsmöglichkeiten'. I prefer the term conceive, because dogs and cats seem very capable of missing a close companian, whether another pet or a human person. They cannot, however, thematise, objectify, modify or otherwise conceptualize the sense of loss they experience. They are caught up in it, just like they are caught up in their *Umwelt-tunnel*.

[xiii] Plessner situates the *Seele* at the level of the *Innenwelt*, the *Bewustsein* at the level of the *Aussenwelt* and the *Geist* at the level of the *Mitwelt*. The term *Mitwelt* would translate as 'with-world', which is not the same as 'middle world' or 'shared world'. For this reason I mostly prefer to use the German term.

[xiv] "In der Distanz zu ihm selber ist sich das Lebewesen als Innenwelt gegeben."

[xv] "Die Mitwelt trägt die Person, indem sie zugleich von ihr getragen und gebildet wird."

[xvi] Making the distinction between the operational closure of the observed system and that of the observer is itself an observation that belongs to the domain of the observer. Hayles (1999:145) rightly points some epistemological paradoxes in Maturana's position here.

[xvii] On the role of semantic description, language and human consciousness, see Maturana & Varela 1998, chapter 9: Linguistic Domains and Human Consciousness.

[xviii] This could bring us to a discussion of Searle's Chinese room argument against the Turing-test. See on this point Hayles (1999: xi-xiv and 289-290). Maybe the point is that Searle's fear that machines could in fact act as if they know Chinese, meaning that once they convince a Chinese of such knowledge over an extended period of time they will in fact have developed the capacity to attribute meaning to the signifiers of the Chinese language. This is not to deny that their meaning will not differ substantially from ours, due to difference in embodiment and historicity. See also Floridi, Taddeo (2009).

[xix] Cf. eg Karnow (1996) who points out the precarious legal implications of this planned unpredictability.

[xx] For this reason Teubner's example of animal liability in medieval times seems mistaken to the extent that it aims to demonstrate that it makes sense to attribute legal personhood to non-humans. See Teubner 2007.

[xxi] Nevertheless animals can be said to have a specific personality: fearful, trusting, secretive, unscrupulous, loyal. This probably means that personhood entails more than just moral and legal responsibility; it would be interesting to investigate to what extent having a personality is a necessary but not a sufficient condition for moral or legal personhood.

[xxii] "Seine Existenz ist wahrhaft auf Nichts gestellt."