2nd International Conference on Autonomous Robots and Agents
December 13-15, 2004 Palmerston North, New Zealand

A Terminological Stance on Artificial Autonomy

Colin Schmidt
Sorbonne University & Le Mans University,
LIUM - FRE 2730 du CNRS, France
Colin.Schmidt@univ-lemans.fr

Felicitas Kraemer
Institute of Science and Technology Studies (IWT), Bielefeld University
P.O. Box 10 01 31, 33501 Bielefeld, Germany
felicitas.kraemer@gmx.net

Abstract

In the present enterprise we take a look at the meaning of Autonomy, how the word has been employed and some of the consequences of its use in the sciences of the artificial. Could and should robots really be autonomous entities? Over and beyond this, we use concepts from the philosophy of mind to spur on enquiry into the very essence of human autonomy. We believe our initiative sheds light upon the problems of robot (and other machines) design with respect to their relation with humans.

Keywords: Autonomy, terminology, machines, personhood, communication, emotion, Artificial Intelligence

1 Introduction: The misleading use of the term 'autonomy'

The term "autonomous agent" is amply used in current Artificial Intelligence (A.I.) and has become a standardized term. We wonder though whether this expression would remain valid if one was able to demonstrate that robots are not or could not be autonomous. One cannot blame the A.I. community for striving for full autonomy as we, and the general public now too, expect Science and Technology to make significant bounds forward; such a will for change is so strongly embedded in western culture that we tend to ask for the impossible or the inappropriate.

In this paper, we give reasons why the adjective 'autonomous' largely used in A.I. discourse to describe agents is utterly weak in its explanatory power. The notion of autonomy is an extremely vague term and requires further clarification. It will turn out that, for conceptual reasons, it is a *terminological mishap* to apply the notion of autonomy to artefacts because it would be a mistake to apply it to humans. If it does not apply to humans, can it apply to artefacts?

In the first part, we attempt to shed some light on the philosophical tradition of the term, defining autonomy as the capability to abide by self-imposed moral laws.

This presupposes that the artefact already has or successively develops its own moral standards and follows them independently from the moral standards of its creator.

An autonomous agent, by definition, must have a moral will of its own that proves its independence of the will of its creator.

The standards of the creator and the creature may even contradict each other (think of a "morally good creator"/ "evil artefact" and *vice versa*). The degree to which the artefact can break free from the intention of its (human) creator and find and employ its own moral standards is what demarcates its independency with respect to its creator.

This means, we are suggesting a strong and normative term of autonomy. We think it appropriate to (re-)introduce the original humanist meaning of autonomy to the current A.I. debate which was replaced by the merely functional use of the term. (Our stance on this matter would secondarily suggest that the core of A.I. is anti-humanistic in goal and character.)

The second part of the paper denies the possibility that artificial systems are able to fulfil the conditions required by our strong notion of autonomy. They can be called autonomous agents only in a narrow and derivative sense; and we doubt that in its weakened form, the notion make any sense.

The third and last part, although on a merely conceptual level, gives some examples to test the autonomy of selected artefacts to provide a tool for separating real autonomy from non-autonomy.

2 On what machines are not: towards a proper understanding of 'autonomy'

2.1 Moral implications and the creator-creature relationship

The word 'autonomy stems from the Greek 'autos' which means 'self' and 'nomos' which means 'law'. Taken in its original sense, an autonomous entity is able to impose or actually imposes a formal law on itself. This definition is close to the Kantian notion of autonomy and can serve as a working definition [1]. So we have to note that traditionally, the word autonomy is closely linked to morals and moral laws.

Furthermore, it stems from a theological background. There is a similar debate in the Enlightenment era, starting in the 17th century. According to *Deism*, there once was a creator, God, who has, like a watchmaker, created the universe and brought everything into being. But although he was the first cause of all this and the creator of man, he decided not to interfere any more in our affairs but to release us into freedom. So, according to prominent Deists like Voltaire, there is some scope within the given. It is possible for us to attain autonomy, or freedom of will during our lifetime, *although* we are God's creatures by origin.

Here one can easily draw a parallel to the artificial. Although artificial creatures are, of course, created by humans, it is still possible that they can develop into autonomous beings in the course of time. The fact that they are man-made does not necessarily imply that they cannot become autonomous some day. Thus, there is a fundamental problem lurking behind the notion of autonomy. All artefacts are endowed with a certain dispositional structure their creators had in mind. On this most basic level, these artefacts are non-autonomous as far as their origin, and as they are products and material implementations of someone else's intentions. But this does not necessarily imply preventing them from *developing* more or less independently. So the question is to be raised whether their being man-made condemns them to staying dependent forever, of whether there is some autonomy possible for them in the future, brought about by developing learning skills and adaptive behaviour.

In our secular age, human beings play a major role as creators, whereas the idea of God is of minor importance, and it is the concept of evolution that is replacing the idea of a world created by God. We will call this type of autonomy "possibility-controlled autonomy."

So the word autonomy establishes a relation between the creature and its creator. We can call this form of autonomy "origin-controlled autonomy". This raises the question whether the act of creation definitely preserves a being from becoming autonomous in the long run. There is evidence that, although beings or artefacts were created once by someone else, they can gradually attain some autonomy by learning and adaptive behaviour. This presupposes that the creator has endowed them with a capability and disposition to develop free will.

To sum this up, traditionally, autonomy is imposed *on* humans as moral agents and *by* humans with respect to God.

We strongly assert this original and traditional use of the word autonomy to maintain and re-vitalize its original meaning. In the current A.I. debate, the word autonomy is not used in a proper sense since its moral implications and humanistic background are neglected.

In our secular age, our notions of dependence and autonomy are not necessarily come of God any more. In many modern theories, the idea of society has replaced the notion of God (e.g. Emile Durkheim [2]). It is society we partially depend on now from which we strive to gain some autonomy by leading an individual way of life. Society is to be understood as the community of language users. In the last section, we will show that, paradoxical as it may be, it is the *dependence* of an individual on a community of language users and its interaction with it that provides the pre-condition of autonomy. Before leading up on these thoughts about the relation between the autonomous person and society, we must indicate for the reader what we mean by the word 'person'.

2.2 Autonomy and Personhood: The role of higher-order "volition"

In philosophical tradition, autonomy is closely related to *personhood*. So a proper definition of personhood as it is employed by one of the leading authors in the field could help in gaining insight into the notion of autonomy. Daniel C. Dennett [3] gives advice concerning which entities should be called persons and which not (≈other minds problem). Although Dennett's investigations go back into the seventies, it is legitimate to revive his conceptual tools, since they deserve being brought in again to, inversely, add value the notion of non-person (developed elsewhere, *cf.* Schmidt [4]) that will be important for the next part of this paper.

Dennett gives six mutually interdependent criteria.

"The *first* and most obvious theme is that persons are rational beings […]. The *second* theme is that persons are beings to which states of consciousness are attributed, or to which psychological or mental or intentional predicates are ascribed. […]. The *third* theme is that […] our treating of him or her is somehow and to some extent constitutive of being a person. […]. The *fourth* theme is that the object to which this personal stance is taken must be capable of reciprocating in some way. The *fifth* theme is that persons must be capable of verbal communication. The *sixth* theme is that persons are distinguishable from other entities in being *conscious* in some special way [...]. Sometimes this is identified as self-consciousness. Three philosophers who claim - in very different ways - that a special sort of consciousness is a precondition of being a moral agent are Anscombe, [...], Sartre [...] and Frankfurt [...]." […] (Dennett 1978: 270 n.)

In respect to Dennett's work, we claim that the conditions of personhood having to do with the fourth and fifth themes specifically merit our attention as the Self (non-physical person) is inscribed in society and as such does not exist in absence of other persons (the Other): To claim some autonomy, a being or artefact is able to recognize other people as persons and to grasp their intentions properly. It must expose a real dialogical character. This is what Dennett picks up when he deals with the Gricean notion of communication (*cf*. end of the paper; Dennett[3]).

In any case, autonomy, somewhat paradoxically, requires social interdependence.

Dennett discusses his six features of personhood in a Wittgensteinian or pragmatic way. Rather than working from speculative ontology, he asks which *stance we actually take* towards these beings. He explores how we "explain and predict their behaviour by *ascribing* beliefs and desires to them" (Dennett [3]: 7) So Dennett deals with *our* presuppositions about the systems. According to Dennett, we take an 'intentional stance' towards a being whenever we ascribe *intentionality* to it. We usually take this 'intentional stance' if we recognize that a being behaves as if it had desires and a will of its own, i.e. if it exposes visible signs of *intention*. If our predictions about the system's behaviour are fulfilled and have turned out successful, then we legitimately ascribed an intentional structure to it.

Then Dennett goes a step further than only analyzing intentionality or our intentional stance to define *personhood and the respective stance*. He maintains that "being an intentional system is not sufficient condition for being a person, but is surely a necessary condition" (Dennett [3]: 271). What would be a sufficient condition of personhood for Dennett? Furthermore, what are the most striking characteristics missing from a machines, robots or... non-persons?

As artefacts, they do not have *second - or even higher-order volitions*. Having higher order volitions is identical to what Harry Frankfurt (cited in Dennett [3]: 283 ff.) maintains to be the criterion for being a person and thus being autonomous. To put it very briefly, according to Harry Frankfurt, second- (and higher-) order volitions are reflexive, i.e. they bring in a level of self-consciousness where we reflect upon our own first order desires. A being that has only first order desires like being hungry or eager to take drugs, to use Frankfurt's example, is only a so-called '*wanton*'; in living on this level, he or she is *not a person* and is not bearing the burden of *personal responsibility*. Personhood only enters the scene as soon as someone actively acts upon his or her basic desires, responsibly choosing which one of them he or she accepts and which ones he or she refuses to have lived out (Dennett [3]: 285, quoting Frankfurt [5]: 14 ff.). The second or higher-order level is what Frankfurt calls a person's *will* or *volition*, being his or her capability of choosing between different strives and desires, and by doing so forming his or her own character while consciously exerting control over his or her own basic level of desires.

To sum this up, we require two criteria for our strong notion of autonomy to be fulfilled. Firstly, an autonomous agent must, by definition, abide by self-imposed moral laws that prove its independence from the original intentions of its creator. This is autonomy defined in terms of a *relationship between creator and artefact*. Secondly, there is *autonomy in terms of self-reflexivity*. In this additional second meaning of the term, a being or creature is autonomous if it is not a slave and passive victim of its own immediate impulses, but is a *real person* able to 'consciously' take up a certain attitude towards its own immediate 'feelings'. In short, we do not accept mere robotic "wantons" as autonomous agents. Being autonomous requires the capability of distancing oneself from immediate impulses by means of a self-imposed second-order system of wishes, i.e. a kind of moral law. Only if an artefact or human being is able to expose this kind of self-reflexivity and thus to be autonomous in this second sense would we say it is a real being, or a *person*. Otherwise the notion of non-person would be the 'default value' deemed suitable for the kind of being in question here.

3 'Autonomy' in the strong sense of the term

The next step to be taken is, in a Dennettian way, to ask ourselves whether we should ascribe second- or higher-order volitions to certain artefacts. It turns out that, even if one accepts the idea that they might have some rudimentary *basic desires* as e.g. Dietrich Dörner does (Dörner [6]: 557-574), this does not mean that we are automatically inclined to ascribe second-order volitions to it.

In table 1 picturing the hidden layers of personhood, the constraints of everyday communication are not taken into account.

Table 1: The hidden layers of personhood

Level of autonomy: person, autonomous agent	**Moral choice, self-imposed laws:** intentionality as long-term rationality, following moral rules that exceed immediate gratification of desires	**Higher order *volitions*:** self-reflexivity, moral self-consciousness; only in human beings, not in artefacts
Level of non-autonomy: "wanton", non-autonomous being	**Mere instrumental rationality:** how to satisfy basic needs regardless of moral rules	**Basic desires:** e.g. hunger and thirst; possible in artefacts

According to Frankfurt, there is a level of "basic desires" like hunger and thirst. But apart from that, human beings usually have the capability to refrain from immediately satisfying these basic desires. To a certain degree, we have the freedom of choice to consider which of our desires we want to satisfy and which ones we do not. That means, we consciously evaluate our own basic desires, and decide about their fulfilment according to moral reasons.

Compared to the basic level of desires, this latter capability of moral choice is of higher order. It requires self-consciousness and enables an agent to develop self-imposed moral laws and to abide by them, even if threatened by unpleasant consequences. These self-imposed moral laws a person chooses for himself or herself to follow are called "volitions". The volitions form the level of moral self-consciousness that designates personhood and autonomy. An example of an act of moral choice is to stay hungry if another person is in need of food.

The higher order volitions that characterize personhood and autonomy are situated in the realm of the social we will deal with in the last section of the text.

For example, Dietrich Dörner's virtual agents, who live on a virtual island, represent their respective 'physiological' states of affairs by picturing them on an internal scale of measure, and then behave accordingly (e.g. they 'read' from their own 'bodily and environmental parameters' that they are hungry or angry and so forth. The whole of their self-presentation at a given moment might be called self-consciousness in a functionalist sense. (Dörner [6]: 301 ff.) There is, of course, only information processing going on, but these processes are represented for the "artefactual being" and trigger a given behaviour. So it is claimed that these artefacts entertain a first-person perspective and thus a minimalist form of self-consciousness (for robotics cf. also Cruse [7]).

It is hard enough to accept the idea that such a minimalist form of self-consciousness and 'desire' is possible for some artefacts. But even if we do accept it, we are not automatically willing to ascribe second-order volition to them, or something like a self-imposed moral law. Therefore we should not call these artefacts persons, or autonomous agents. On the contrary, we should name these artefacts Dennettian wantons as they only react to certain 'physical' or functional needs without entertaining a second-order system of morals.

This is why we are not willing to ascribe autonomy to them in the strong and genuine sense of the term – even if they obviously get along in the absence of their creators, even if they learn to adapt to the environment and even if they show something "comparable" to self-consciousness. Autonomy, in the strong sense, requires a being's capability for developing a self-imposed moral system made out of second-order volitions. As we understand it, beings without these capacities must not be claimed autonomous without betraying the whole history and the essence of the term. This moral system must be developed independently of the intentions of the being's creator.

Only upon seriously considering the significance of second-order volition will we arrive at what Grice claims to be the basis for communication: What is needed is the ability to recognize other people as persons and to treat them accordingly. A person must have an understanding of the fact that other people are persons as well, and have their own beliefs, intentions and desires. A person must be able to ascribe intentionality to others and recognize the fact that they are persons too. In short, to be counted as a person, a being must have an idea of what goes on in other persons' minds and how the world may be perceived by them. Artificial beings must have a theory of mind to do all the things that robotologists – however outlandishly, discretely or secretly– expect them to do. This would be the necessary condition for artefacts to be social beings instead of "(artificial) autists", as well as to establish communication to serve as the basis for authentic man-machine communication (cf. Dennett [3]: 270; ibid. 280).

4 Conclusion: On what humans are not

One may argue that all types of machines do not have the wherewithal to become *totally* autonomous. We did so above, but this would mean that we accept a very peculiar definition of personhood, that in which people would be *totally* autonomous thinkers. This is not the presupposition that *we* mean to imply, even if today it is in vogue in the sciences of the artificial. Thinking on one's own is one thing; thinking *in exclusion of other*s is another. What we meant to do in arguing against full-blasted machine autonomy was simply to introduce the reader to the fact that other inspirational options in the "mechanisation of Man" need to be explored.

As we have seen above, looking at how ordinary people and scientists – other ordinary people – use language can be revealing. In fact the frequent use of the word "autonomy" in the literature concerning the artificial world is what raised our eyebrows. The downfall of using autonomy-rich terminology in these A.I. related fields is that many use it to mean "human-like" and this is totally erroneous as humans are, at best, only semi-autonomous. The importance of the use of language by humans, as so simply and vividly put forth by M. Tournier in his world-taking philosophical story success about Robinson Crusoe (entitled *Friday*), shows that we are forever indebted to fellow language users to correction of perception of the world; our fellow representatives of personhood make the proper elaboration of meaning possible. Alone and living in total autonomy on his island, Robinson started to suffer from the *absence* or *non-presence* of the other. He would become mentally, morally and spiritually low. Why? Was he not enjoying total freedom? As members of human community, we look to others in order to check (through discussion) the reliability of our senses. The *concept* of other humans, moral and intellectual, provides this possibility as it *is* the structure that helps oneself know one's position in society and makes one's own point of view on things possible.

The private domain of human life is *dialogical*. In hypothetical communication, we test our thoughts, sayings etc. against that structural slot provided by the concept of other –"What will she think if I...?", "what will he say?"– before committing oneself to decisions and actions of public consequence –"if others in my community do *X*, then I must do *X*" or "if the others think *Y* about this matter, mightn't I think *non-Y* or *Z*?". In society, imposing one's own individual principles is always performed in consideration of group-imposed guidelines or others' individual self-imposed laws. The average Dennettian wanton does not master these considerations enough to form the necessary self-will to create these such socially-induced decision settings.

In this paper, although we separate non-autonomy from genuine autonomy and show the logical difficulties artefactual entities have with respect to this difference, it would seem of paramount importance at this point of time to point out the inappropriateness of pushing the artificial towards total autonomy, that is if the goal is to copy man (as in the case of building humanoid robots), simply because autonomous is what humans are *not*; however in the future, if a totally autonomous robot were to be useful to humans, it most likely would have a utilitarian function *because* man cannot be fully autonomous.

The semi-autonomous character of a (human) person is constantly checked, validated and depended upon in normal human life (this article was written for those susceptible to read it in mind).

5 References

[1] Schneewindt, J.B., *The Invention of Autonomy: A History of Modern Moral Philosophy*, Cambridge UK, 1998.

[2] Durkheim, É., *The Elementary Forms of the Religious Life,* London, Allen and Unwin, 1912.

[3] Dennett, D.C., *Brainstorms: Philosophical Essays on Mind and Psychology,* Hassox. Chapter 13: “The Abilities of Machines”, pp 256-266, Ch. 14: “Conditions of Personhood” pp 267-285, 1978.

[4] Schmidt C.T.A. (2004), "A Relational Stance in the Philosophy of Artificial Intelligence", *European Conference on Computing and Philosophy*3-5 June, University of Pavia, Italy: Kluwer Academic Publishers.

[5] Frankfurt, H., “Freedom of the Will and the Concept of a Person”, *Journal of Philosophy*, LXVIII January 14, 1971.

[6] Dörner, D., *Bauplan für eine Seele,* Reinbek (2001).

[7] Cruse, H., “Feeling our Body – The Basis of Cognition?” in *Evolution and Cognition* 5, Vol. 2, pp. 162-173, 1999.

[8] Dennett, D.C., *The Intentional Stance*, Cambridge Mass., 1987.

[9] Dreyfus, H., *What Computers Still Can’t Do: A Critique of Artificial Reason*, Cambridge Mass: The MIT Press, 1972.

[10] Grice, H.P., "Logic and Conversation", in P. Cole and J. Morgan, eds., *Syntax and Semantics*, vol. 3, Academic Press, pp 41-58, 1975.

[11] Kant, I., *Critique of Practical Reason*, transl. Lewis White Beck, Indianapolis: Bobbs-Merrill, 1959.

[12] Kant, I., *Foundations of the Metaphysics of Morals*, transl. Lewis White Beck, Indianapolis: Bobbs-Merrill 1959.

[13] Locke, J., *An Essay Concerning Human Understanding*, Peter H. Nidditch (ed.), Oxford, Clarendon, 1979 (repr.).

[14] Putnam, H., "Brains in a Vat" and "A Problem with Reference". *Reason, Truth and History,* Cambridge MA: Cambridge University Press, 1981.

[15] SCHMIDT C.T. (2004), "Humanoids, from Interfaces to Intelligence. Really? A Philosophical Statement on Retrograding or Scientists Caught Back-peddling", *The 2004 American Association of Artificial Intelligence Fall Symposium Series* on 'The Intersection of Cognitive Science and Robotics: From Interfaces to Intelligence', October 22-24, Washington D.C.

USA, *Technical Report N° FSS-04*, Menlo Park CA: The AAAI Press pp. 55-60.

[16] Schmidt C.T.A. (2002), "Socially Interactive Robots. Why Our Current Beliefs about Them *Still* Work", *Proceedings of the IEEE International Workshop on Robot and Human Interactive Communication* RO-MAN 2002, Sept. 25-27 2002, Berlin Germany.

[17] Smith B. (draft), "Ontology and Information Science", *Stanford Encyclopaedia of Philosophy*.

[18] Strawson, P.F., “Persons”, in Minnesota Studies in Philosophy of Science vol. II. Feigl, Scriven, and Grover Maxwell (ed.), 1970.

[19] Strawson, P.F., *Individuals: An Essay in Descriptive Metaphysics*, London, Methuen (1959).

[20] Tournier, M., *Vendredi ou les limbes du Pacifique,* Gallimard, 1967.

[21] Voltaire, F.-M.A., *Lettres philosophiques* (1734), Raymond Naves (ed.), Paris, Garnier 1969.

[22] Wittgenstein, L., *Philosophical Investigations*, Oxford, Blackwell, 1953.