"Autonomy" and the autonomy of autonomous technical systems

Niels Gottschalk-Mazouz
Institute of Philosophy, University of Bayreuth & University of Stuttgart

AAP (Sydney), 6.7.2010

Since the 1990s, engineers speak with an eye on robots, software agents, etc. of "autonomous technical systems." What does "autonomous" mean here, in machines, and to what extent is our (human) autonomy affected, now that technology is getting "autonomous"? These questions will be answered in five steps, (I) a presentation of the usual (prescientific) understanding of autonomy, (II) a discussion of the concept of personal autonomy in philosophy, (III) the analysis of the notions of autonomy in the discussion of "autonomous" technical systems, (IV) a reflection on how our (i.e. human) autonomy is affected by those autonomous technical systems, and (V) some remarks on how such systems should be established accordingly.

I. (The prescientific understanding of autonomy, very short)

Historically, the concept of autonomy comes from the political sphere (Pohlmann 1971). The city-states in ancient Greece formulated a claim for Self-determination, against threats from within (tyranny) and from the outside (Domination). This autonomy-as-selfdetermination was always one of limited self-determination – limited by rules of superior political alliances. This understanding has, in my opinion, survived until today, for example in the concept of an "autonomous region" in Politics (that allow for limited self-determination within a nation state) but also in Philosophy, where personal autonomy means self-determination vis-à-vis the conditions one finds oneself in and that one cannot change. The same seems to be the case in the discussion of freedom and Determinism, where "autonomy" names the kind of freedom that is all we can have - under alleged determinism (as in Walter 1998).

Just like in politics, what is meant is a self-determination that is limited (by external and internal Determinants), but those very limitations make self-determination possible in the first place.

Please note that already in the ancient texts, like Plato's REPUBLIC, the analogy between the political and the mental has been exploited.

II. (autonomy in modern philosophy)

The most influential definition of autonomy in modern philosophy goes back to Immanuel Kant. Autonomy is now focussing the individual human being, the person. Autonomy, with Kant, is understood as having a certain status ad a certain form.

(Status:) Autonomy is understood transcendentally (and not empirically), i.e. it is the condition of possibility ("Bedingung der Möglichkeit", Kants terms) of responsible action, and that is of morally justified action. Conditions of possibility, as opposed to conditions on actuality, which are the conditions that apply to the concrete action in its actual form. Conditions of possibility are those, that govern responsible action as such, independent of any actual, contingent empirical circumstances. Transcendental autonomy is what we have to presuppose if we want to take somebody to be a responsible actor. It is absolute (not depending on any empirical conditions), unlimited (not diminished by empirical circumstances) and not gradual, and it is a normative (and not a descriptive) concept. It expresses, at least that is how I think one should read Kant, it expresses a certain form of recognition of one's self or of others as capable of responsible action, as having (with that) a dignity etc. Any responsible actor is as such also one that acts only in a morally justified manner, i.e. in a way that his maxim (the rule that governs his action) is permitted by the categorical imperative.

(Form:) In Kant, autonomy is tied to the concept of law. Moral quality, for Kant, is inseparable from the form of a rule and, ultimately, of a law, as can be seen from the categorical imperative's general version and from many other paragraphs in the "Groundwork of Metaphysics of morals", including those that connect morality and autonomy in the way just outlined. So, in the very first sentence of the section entitled

"The autonomy of the will as the supreme principle of morality "of the Groundwork of the Metaphysics of Morals (cf. GMS BA 87), we can read: "Autonomy of the will is the nature of the will, by which he is a law to himself - independent of all characteristics of the objects of this will."

Critics have pointed out immediately that this act-maxim-law apparatus is quite complex and not very adequate to our moral practice. We do not try to think of a rule that might cover our action (a maxim) and then evaluate this maxim only in terms of law-likeness (whatever its contents might be); we do not estimate – one might say - the rule- or law-likeness as such ("rule-worship"). Moreover, the whole concept of rules and laws in morality might follow a model of god-imposed laws plausible only the judeo-christian tradition where there <u>is</u> a god doing so in the first place (Schneewind 1998).

One could add that is a rigid technicist notion of ought and is, of regulation - in the technical sense - of one's own behaviour.

I do not want to discuss whether this criticism (or: all of this criticism) is adequate as a criticism of Kant. Allthemore because <u>other</u> modern concepts of autonomy have been developed that do <u>not</u> focus on law-like form (see already Frankfurt 1971). The new headlines are now: coherence, responsiveness to resoning, responsiveness to reasons (see Buss 2002 for a summary). It is not easy to see what all these conceptions of autonomy (personal autonomy; autonomous action; …) have in common: But I think that essentially it is a two-fold activity of normative self-determination that is expressed by the concept of autonomy: On the one hand, an <u>identification and schematization of the own activity</u> ("you know what you do"; in Kant: Assigning a maxim). On the other hand, a <u>distancing reflection</u> of the identified activity that is evaluative (broadly construed, i.e. including morals; in Kant: Checking the maxim against the cat. imp.). Only insofar as I <u>know</u> what I am doing, and whether I <u>want</u> to do it, only insofar as I have <u>epistemic</u> and <u>normative authority</u> (as we might say) over my actions – only insofar as I have this authority I am an autonomous actor (and person).

On the very least an autonomous actor does not merely react to circumstances.

That persons do have the capability to determine their doings in these two respects (interpretation; evaluation), that is what I see as the common (minimal) ground that is expressed by the ascription of autonomy. I think. Kant pointed out that to determinate ourselves in these two respects is never only a matter of the circumstances, at least not if we do *want* to conceive ourselves as being autonomous, as being a responsible actor. We interpret and evaluate vis-à-vis the circumstances we find ourselves to be in. But there is a logical gap between these circumstances (Kant: the empirical, the inclinations etc.) and the results of our interpretation and evaluation. This logical gap is what requires and constitutes, one might say, the autonomous person.

Please note that this leaves open how the (interpretative/evaluative) insights can alter behaviour. It is even compatible with a view that these insights cannot alter it at all (Stoa, Spinoza, … Determinism). Even though Kant and many others of course have been pointing out that these insights somehow must be able to make a difference in our behaviour. This, I think, is in line with the metaphysics that is implied in our ordinary language talk about actions and responsibility. But, above all, it is in line with the "engineering perspective" of autonomy in machines that are *designed* such that their interpretation and evaluation does make a difference in their behaviour. So I think that we should take this here for granted and move on to the discussion of what engineers (and engineering scientists) mean when they speak of autonomous systems.

(Some more words on Kant:) Kant seems to have had the view that free action as such is not accessible by empirical investigation because this would be a metaphysical confusion. It would lead to a clash of presuppositions. The empirical is conceived under the regulative idea of being causally closed, of being determinated by antecedent conditions: only this makes empirical science possible. Empirical inquiry thus presupposes that the object of study is determinated. And insofar as we perform empirical studies of human beings, we will only find empirical determinations. For there is nothing other that we could possibly find.

Causality is conceived by an epistemological subject, who attributes law-likeness to any observed repeating sequence of events. Which is why we can not take the interactions of this subject with any objects as causal if we want to catch the epistemological dimension of it. Only insofar as a subject is an object (as any other objects) it is standing in causal relations (and only in these). But this means missing exactly the subjective features! The same argument goes for the subject as an actor.

Our metaphysics of action, on the other hand, just presupposes that actions are not causally determinated. It does not, and cannot, say how this is possible in empirical terms. We and up with a dualism of stances towards the world: A theoretical stance that takes everything, including us, as determinated (Kant: "theoretical reason"), and a practical stance that takes us as being free actors.

So freedom of will and freedom to act, in the Kantian sense, are not empirically accessible.

III .

In the debate over autonomous technical systems, i.e. robots and the like, it has been suggested – in vague analogy to Kantian autonomy – that autonomy may mean the ability of these systems to impose rules on themselves (see Smithers 1997 for such a proposal). But this speech remains entirely metaphorical, if it is not *technically* further specified. Moreover, a closer look at the various uses of the word autonomy in autonomous technical systems – and the definitions that were proposed – reveals that autonomy is not always demanding rules. Rather it points out a series of mostly self-imposed "engineering contraints" (see Maes 1991) that are much more down-to-earth than "imposing rules on themselves" would suggest. I will now present eight notions of autonomy, starting with the most basic, that I found in the writings on autonomous technical systems by (mostly) engineering scientists. I will always try to find a more telling word for the exact property that is pointed out with the word "autonomy":

1st An independence from the electrical outlet, or generally, the ability, to get along without external supply of energy or material. In this sense of "autonomous", autonomous systems are those that, I would say, are autark, self-sufficient. (E.g., energy self-sufficient).

2nd The ability to move freely (ie not only along given tracks). In this sense, autonomous means being mobile.

3rd The ability to complete tasks without user/operator intervention, without remote control. In this sense, autonomous means automatic.

Sometimes these items have been combined, such as in this definition here, which combines the latter two (by Todd 1986, p. 233): "autonomous robots are usually taken to mean free-ranging mobile robots which are not teleoperated but plan and execute their own actions."

Please note the “their own actions”, which is not explaining in technical terms what is the case so does not help much in a definition (though we will get back to this point later).

In the robotics scene, the third item (that of no remote control) is further differentiated and remote controlled, semi-autonomous and (fully) autonomous systems are distinguished. Semi-autonomous systems can be commanded to perform certain more complex actions, such that control in exercised on a higher level. You can give commands like “follow this line” rather than low-level commands like “turn right”, “forward” or directly controlling motor action.

OK, so far we had “autonomous” as self-sufficient, mobile, and automatic. Here are the next three: Autonomy shall mean that:

4th The future behaviour of the system depends only on its internal states (Autonomous in this sense means independent from the environment, not influenced by the environment. (this is according to McFarland / Bosser 1993, p. 145-147).

But one also finds the following, quite opposite definition: Autonomy means then that:

5th The same task can be completed in different situations. So the system is adaptive to the environment, and

6th The same task can be completed in an ever improving way, so the system is learning.

The next (and last) two definitions are the most complicated, because they make reference to the perspective of an observer. Autonomy will then not be, as it was in the 6 definitions already given, an objective predicate that is observer-independent, but a subjective predicate that is relative to the observer (this is an idea that was present at the beginning of the discussion about autonomous technical systems, see Covrigaru / Lindsay 1991, and that has been picked up more recently by the leading robotics groups in Germany and Switzerland. So

7th Autonomy means that the system is not only adaptive and automatic, but operates in a way that an external observer can identify types of behaviour that were not explicitly specified while constructing the system (Knoll / Christaller 2000). Autonomous in this sense means being innovative.

And here is another similar definition:

8th "The more the controlling agent A is aware of the internal condition of B and The better he knows the laws by which the state of B can be influenced, the better A is able to control B. (...) B's autonomy with respect to A is - qualitatively spoken - inversely proportional to the knowledge that A has about B." (Pfeifer 2003; my translation).

So this is interesting: A System is thus autonomous to the degree that its internal states and their laws of change are not known, so being autonomous means being opaque, unpredictable.

OK. I will now try to organize these eight notions of autonomy: "(energy) self-sufficient," "mobile", "automatic", "environment-independent", "adaptive," "learning," "innovative" and "unpredictable". I will treat them as pointing out various respects of autonomy, and come up with one covering definition in more general terms.

So what are general features of these aspects?

1. We said that the latter two are relative, to an observer. This observation can be generalized, because on closer inspection the other respects are also specifying relations:

- "Energetically self-sufficient" and "environment-independent" indicate such a negative relation to the environment
- "automatically" indicates a negative relation to the user,
- "adaptive" a positive relation to the environment and a negative to the user,

- "learning" (just as “innovative”) finally a positive relation to the environment and a negative to the designer or maybe also the user.

These relations can be seen as different respects of autonomy.

2. Even if it is not stated explicitly, the autonomy can only exist under certain “enabeling” conditions that hold in the very same respect (“dimension”, if you want). Take energetic autonomy, for example: It exists only under certain environmental conditions that include energetic conditions, such as adequate sunlight (if solar cells are used) or adequate charge of batteries. So the system is energetically autonomous only in the sense that it does not require certain forms of energy input (but requires others). The same goes for the other respects of autonomy.

In general, I would say, the ascription of technical autonomy therefore has the following form:

***x* is autonomous in (positive or negative) respects *a* under (usually not explicitly stated) boundary conditions *b*.**

Such an ascription sets the system in one or more respects in relation, it specifies namely (i) exchange relations with the system-environment, (ii) interaction relations with the user and manufacturer (control or programming) - and possibly even third parties (monitoring); (iii) epistemic relations concerning the knowledge of user / designer / third parties.

Please note that these respects are not completely independent from each other, because epistemic relations influence our interactions and these interactions influence the exchange relations of the system.

So that's how one can generally summarize the results:

autonomy specifies relations of exchange, interaction and knowledge of *x* in one or more respects *a* under the assumption of boundary conditions *b*.

IV.

On this basis, now, our autonomy and that of technical systems can be compared and related to each other:

First, there is an asymmetry in the sense that our autonomy is the expression of a morally relevant reflective intentionality of people, whereas technical autonomy is a relation to people (or to the environment). (It can be useful, however, to attribute intentions to technical systems, I shall come back to this point later).

Second, there are complementary relationships of control, because we control the system on the basis of what the system cannot provide by itself, and what the system does provide by itself cannot be controlled by us anymore. This is not a symmetric relationship, however, because these control relationships are set up by us, in the dimensions of exchange, interaction, knowledge.

Third, the way that it feels like to interact with autonomous technical systems can be very different. There are basically three ways to conceive oneself vis-à-vis an autonomous technical system:

a. The system can provide support that goes unrecognised, then I experience myself as an enhanced subject (and when the system fails, as incapable).

b. If the support is to be recognised, it has to be represented somehow. I can then treat the system as a powerful instrument. But this works only insofar as I know about the determinations of the system, about how to control it. Only then I will know whether it was me or the system that failed when my action has not been not successful.

c. Sometimes, though, we experience autonomous systems also as subjects. We, as users, treat them sometimes as if they were having intentions. This is encouraged by the design of these systems, by their exterior (that may have a human shape) or the way they interact with us (speech etc.). Why is it useful for us to do so (even if we know very well that we did ascribe these intentions and that it is not the machine that is attributing these intentions to itself)? I think intentional vocabulary is used when one wants to be able to behave and interact if determinations of behaviour (be it one's own or the other's, by the way) are not known. So I think there is an intrinsic connection to the last of the above aspects of autonomy (that was: opaque/unpredictable).

If we treat autonomous systems in the latter way, as if they were having intentions, it is (at least most of the time) clear to us that these are not real intentions, because we ascribe them from the outside. The system has no epistemic or normative authority over its intentions because we do not take it to be able to interpret and evaluate its actions in the sense that autonomy in the Kantian sense would require. That is why fulfilment of these purposes, fulfilment of the preferences of a machine, are not taken to be relevant as such: There is no point in respecting the autonomy of a machine if we know already that it is a machine. Whereas respecting the autonomy of human beings (in the Kantian sense) means exactly to respect the ability to interpret and evaluate one's own actions.

As I said, we do treat autonomous systems sometimes as if they were having intentions. It is, on the other hand, for those machines sometimes useful to be able to process intentional behaviour or vocabulary of humans, of users. Such an as-if-understanding might facilitate the human-machine-interaction. Autonomous technical systems may even use an internal architecture that is called a belief-desire-intention architecture. That is when their determinations are bundled on lower levels of description in such a way that an abstract control layer can be implemented that allows the engineer or programmer to ascribe as-if-intentions to the machine. This might help the machine in reacting adequately to intentional behaviour or vocabulary of a user, then.

So, on the one hand, part of what constitutes our autonomy can be ascribed to machines as well. That is, having and developing intentions. On the other hand, part of wthat constitutes the autonomy of autonomous technical systems can also be ascribed to humans. They can also be seen as to be "(energy) self-sufficient," "mobile", "automatic", "environment-independent", "adaptive," "learning," "innovative" and “unpredictable”. Sometimes we might even use the word “autonomous” for any of it. We use “autonomous” then not in the transcendental, but in the empirical sense, i.e. as a description. We can even measure the degree of autonomy that a human has, in the empirical sense. Of course, when it comes to humans, we would think of some further respects of autonomy (in the empirical sense), stronger than the eight presented above: Like, e.g., as is frequently found in the literature, being responsive to reasons in an adequate way etc. It is in Bioethics that one finds such richer concepts of empirical human autonomy (cf. Miller 1995), and maybe one day our machines will be complex enough to discuss their autonomy in the terms of bioethics as well. But for the time being this is just pointless, for they are so dumb…

Transcentental autonomy however, that is autonomy in the Kantian sense, autonomy that is morally important, is of a different kind. It is no empirical predicate at all. It rather points out a presupposition of our everyday practices (and, of course, of parts of our jurisdiction, like criminal law, or human rights). Logically it is independent of any empirical fact. This means that machines can as good as we are (or even better) in empirical respects, perhaps in every empirical respect, and we still do not have to ascribe autonomy to them because of their performance. Maybe one day we will want to do so, but there are no empirical facts that will ever require this. It is ourselves who decide whether we want to do this, because ascribing autonomy in the normative, in the Kantian sense, is an act of recognition of ourselves and others as moral persons.

But, even though there is no logical connection between the empirical and the transcendental, there is a pragmatic connection. For the ascription of both types of autonomy serves practical purposes that overlap. We do ascribe autonomy (in both senses) to structure our actions and interactions. These actions and interactions may fail, however. And if they constantly do so, we may well ask ourselves whether we want to continue ascribing transcendental autonomy. So it may become implausible to ascribe transcendental autonomy. And that is when the actions that we, humans and ascribers of autonomy, carry out under this ascription – when these actions constantly fail. While there is no objective standard of failure here, I think that we simply would not uphold ascribing autonomy forever if the system we are confronted with behaves e.g. “just too simple”. Of course, it is only from a perceived level of structural and behavioral complexity (and in humans: development) on that we find it plausible to ascribe autonomy in the first place. So the plausibility of ascription of normative autonomy depends on empirical autonomy, on demonstrated empirical autonomy as one might say, i.e. on performance, in the eight given respects (and possibly, in humans, more than these eight). Though this cannot be a strict relationship.

All one can say at the moment, I think, is that empirical performance seems to be a prerequisite of a lasting ascription of normative autonomy. Machines are just not good enough to make us wanting to uphold such an ascription for longer time - even if we tried. Maybe one day machines will be better, see Chalmers' discussion of the singularity, and then we will find out whether we feel some urge (and, then, will resist this urge ot not) to ascribe also normative autonomy to machines. I think that the crucial point will finally be whether these machines couple to the "space of reasons" (Searle), whether one can see them as the authors of their own speech acts. But technology still has a very long way to go until this might be the case.

V.

So the last remarks may have seemed quite science-fiction like to you. But the issues of robots as possible moral subjects, and of the moral quality of their actions, have not only been discussed, as for some time, in literature and film, but also in philosophy (see eg IRIE 2006). Most of the time they are imagines as robots or central computing systems (running wild). Autonomous technical systems however are not the same as robots. All technical systems that have sensors and actuators, that can sense and effect something in the real or in the virtual world, are candidates for autonomous technical systems. I think that the real challenge is the development of a whole "robotic infrastructure". So we should not think (only) of industrial robots or service robots (with legs and arms etc.). But of modern cars, of mobile phones and personal computers, of computerised access systems (door locks, cameras) and the like. A robotic infrastructure consists of sensors (e.g. cameras), processors and memory (e.g. in your PC or Google's servers), and actuators (locks). Not all of it needs to be built into the same device (as in a robot), because devices contibuting these elements are networked (and will be even more in the future): They work in combination with other technological artefacts, and they perform the better the more they are networked (form a system or a cloud).

Some important questions have already been discussed with robots in view, but others should be discussed as well I think with this infrastructure in view. What has been discussed (see Christaller et al. 2001) with respect to autonomous robots is essentially that we should design them as transparent as possible and if necessary give additional information on their capabilities and doings. Only then we will not be negatively surprised by their autonomy. (please note that this will let them appear less autonomous in the eighth sense of autonomy given above…).

So, let's have a look at the proposals of "transparent design" discussed in the literature.

- Some authors really proposed to signpost these robots as "attention: autonomous", something that does not tell much given the plurality of possible notions of autonomy, so one should be more precise I think, like signposting: "attention: free-ranging". Maybe a form-follows-function approach might also help a user in recognising capabilities of a machine.

What else?

- The systems should, one can read in the literatur, be able to make as clear as possible not only what they <u>can</u> but – above all – what they <u>are</u> about to do (as they <u>can</u> do many things). For example, if carrying out orders they may constantly (or on demand) repeat these orders or more generally their intentions.

- They should make clear that they learn from user interactions (to make clear to the user that he has some responsibility for future system behaviour).

- They should log their performance and their learning processes such that they can be subject to revision by the user.

- They should ask for permission by affected users if this permission is in doubt. *Or, more generally spoken, they should draw on the interpretative and evaluative ressources of the users as those that count and should constantly do so (if you want, recalibrate themselves).*

Some of these proposals seem a little helpless, even naïve to me. My main concern is that successful technology is such that it does <u>not</u> require the user to answer all the time to the question "are you shure?", to review logs etc. (maybe the PC is the big exception here…). It requires to "just function" in a supportive way. Constantly giving explicit information and asking for approval may very well confuse the user, for shure it requires him to concentrate on the machine and keeps him from doing what he originally wanted to do. Moreover, if these machines learn or gather data (and this is signposted or so) and then you should be careful because they do, what does this mean for you? Shall you decide in every instance whether the machine shall gather and/or forward the data, this will be very hard to do, and this not so much because there will be many such instances (which will be), but because it is not at all clear what content this data has, in general: what it means that something is learned or stored or forwarded, what it is relevant for. It will not even be all propositional or audio/video, but it will be training data, lower level stuff that is non-propositional or not well structured. This is allthemore the case as it will be data that still has to be processed (in the robot or in the network) to be of value. If there is some storage or upload function built into the machine that we are faced with, it is hard to tell what we are consenting to to be uploaded when we are asked for consent. So it will be, if at all, only possible to agree or disagree to the fact that (and maybe: when) there is an upload but not to what is uploaded (because you just do not know what it is, i.e. what it means that is uploaded).

So, the robotic infastructure will pose challenges on the informational level, and this allthemore as it seems so be a networked infrastructure. These challenges are already experienced with the internet and with pervasive computing, i.e. the networking of distributed objects like mobile phones, and personal computers and servers etc. They are discussed under the heading of privacy, of informational self-determination. But from the viewpoint developed here this is only part of the problem. The other part is the potential of this smart infrastructure is to be a robotic infrastructure, i.e. to be able to actually do something (more narrowly spoken, to enable us or to inhibit us in doing something). So it is not only information that is at stake. It is that our means of action are affected directly. It is (direct) control. Again, consent does not solve the issue. Suppose that you aggree that your garage door of office building door is controlled automatically. If it is a networked system, then you do not know (exactly) what this will mean. For it depends on information and its processing that is not yet available or done. There are a couple of science fiction stories around about “smart houses” that look in (or help others to lock in) their inhabitants. Thisis what I mean by control as opposed to mere surveillance (which is normally seen as the lurking danger of all this networking). It is one thing if others can find out at will whenever you have passed through a door (and maybe take some ation against you later). But it is another thing if others can keep you at will from passing through this door in the first place.

So, in a negative utopia, others can not only find out all that I do, but others can make me do or keep me from doing as they like - without having to take the detour of sanctioning me using surveillance knowledge, or confusing my beliefs or other purely informational mechanisms, but by depriving me directly of means of action to their like. So the danger is not only the surveillance society but – as I would put it – the control society. And this now affects our, human, empirical autonomy in just the eight ways spelled out above. Because if a machine is autonomous in one of these respects, it is no longer us who control it, but the machine. Or whoever controls the machine. This has been spelled out in terms of movement, but it is true for the other respects as well.

Since we now see the political dimension of it, the dimension of power relations, we understand that something is missing in the way that the labelling requirements have been discussed. Robots should not only tell what they are about to do. I would also add that for intentions are ascribed and programmed, for these machines are carrying out orders or follow human purposes, so they act on behalf of us and we act through them, and that is why it should also be clear whose orders they carry out. It is not enough to know what the system is doing, one also will want to know who made the system do what it does (who, if anybody, "controls the machine"). Because as artificial actors, these machines mediate power relations. Just imagine: If such a machine will not let me pass, it should be able to find out why, i.e. not only because of which characteristics of myself, because of which rule, but also (and maybe even more) because which person's or corporate body's command to do so. This is even more challenging if we face a networked object where it is not at all obvious who is in charge of it (or who is contributing how to the effects of this object: Manufacturers, their suppliers, programmers of used software, the consenting user, the government or criminals or curious hackers maybe – who knows?). Only if this accountability will be possible also in a world of robotic infrastructure, we can continue to hold people responsible for their actions, and this is one major point in ascribing autonomy. So normative, transcendental autonomy may well be at stake (in the sense of becoming pragmatically pointless and thus implausible) in the course of developing autonomus technical systems. It is our autonomy that we are worried about then, and that we should be worried about. Our autonomy, today in the empirical and tomorrow maybe also in the transcendental sense.

Thank you for your attention!

Brandom, R. (1994): *Making it explicit. Reasoning, Representing, and Discursive Commitment*. Cambridge, MA

Buss, S. (2002): "Personal Autonomy", *The Stanford Encyclopedia of Philosophy (Winter 2002 Edition)*, Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/win2002/entries/personal-autonomy/>.

Christaller, T.; Decker, M.; Gilsbach, J.-M.; Hirzinger, G.; Lauterbach, K.; Schweighofer, E.; Schweitzer, G.; Sturma, D. (2001): *Robotik. Perspektiven für menschliches Handeln in der zukünftigen Gesellschaft*. Heidelberg

Collins, H. M.; Kusch, M. (1998): *The Shape of Actions. What Humans and Machines Can Do*. Cambridge, MA.

Covrigaru, A. A.; Lindsay, R. K. (1991): "Deterministic autonomous systems". In: *Artificial Intelligence Magazine,* S. 110-117.

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

Gottschalk-Mazouz, N. (2007): „Überwachung und Macht & die Spezifik technisierter Überwachung“, in: Gaycken, S.; Kurz, C. (Hg.): *1984.exe - Gesellschaftliche, politische und juristische Aspekte moderner Überwachungstechnologien*, Bielefeld, S. 161-183.

Gottschalk-Mazouz, N. (2008): „Internet and the flow of knowledge: which ethical challenges will we face?“ in: Hrachovec, H.; Pichler, A. (Hg.): *Philosophy of the Information Society. Proceedings of the 30. International Ludwig Wittgenstein Symposium Kirchberg am Wechsel, Austria 2007. Volume 2*, Frankfurt, Paris, Lancaster, New Brunswik: Ontos, S. 215-232.

IRIE (Hg.) (2006): *International Review of Information Ethics*, Bd. 6 (Ethics in Robotics), URL = < http://www.i-r-i-e.net/issue6.htm>

Kant, I. (1785/1968): *Grundlegung zur Metaphysik der Sitten*, zit nach der Weischedel-

Werkausgabe, Bd. 7, Frankfurt 1968.

Knoll, A.; Christaller, T. (2000): „Selbstrepräsentation, Selbstwahrnehmung und Verhaltenssteuerung von Robotern“. In: Sandkühler, H.-J. (Hg.): *Selbstrepräsentation in Natur und Kultur*, Frankfurt, S. 109-132 (http://www.iais.fraunhofer.de/fileadmin/images/pics/Abteilungen/IL/TC/Publikationen/Christaller2000.1.ps.gz).

Maes, P. (Hg.) (1991): *Designing Autonomous Agents*. Cambridge, MA.

McFarland, D./Bosser, T. (1993): *Intelligent behaviour in animals and robots*. Cambridge, MA.

Miller, B. (1995): "Autonomy", in Reich, W. T. (Hg.): *Encyclopedia of Bioethics*, Revised Edition. New York, S. 215-220

Pfeifer, R. (2003): „Körper, Intelligenz, Autonomie“. In: Christaller, T.; Wehner, J. (Hg.): *Autonome Maschinen. Maschinen werden selbständig - was kommt auf uns zu?* Wiesbaden: Westdeutscher Verlag, S. 137-159 (http://www.ifi.uzh.ch/ailab/people/gomez/PfeiferPublications/Koerper-Intelligenz-Autonomie(GermanPaperFinalOne).pdf).

Pohlmann, R. (1971): "Autonomie". In: Ritter, J. (Hg): *Historisches Wörterbuch der Philosophie*, Bd. 1, S. 702-719.

Rammert, W. (2003): *Technik in Aktion. Verteiltes Handeln in soziotechnischen Konstellationen*. Technical University Technology Studies Working Papers TUTS-WP-2-2003, Berlin.

Smithers, T. (1997): "Autonomy in Robots and Other Agents", In: *Brain and Cognition* 34, S. 88-106.

Todd, D. J. (1986): *Fundamentals of robot technology*. London.

Walter, H. (1998): *Neurophilosophie der Willensfreiheit. Von libertarischen Illusionen zum Konzept natürlicher Autonomie*. Paderborn 1998.