SHOULD AUTONOMOUS AGENTS BE LIABLE FOR WHAT THEY DO?

Jaap Hage*
Universities of Maastricht and Hasselt
jaap.hage@maastrichtuniversity.nl

1. Introduction

After decades of developments in information technology and artificial intelligence, 'autonomous' systems have come to play important roles in our society. In the Netherlands decisions about study grants are in first instance taken by a computer program. Computer programs keep track of the stocks of shops and order new supplies if necessary. These orders are accepted by other computer programs, which prepare the physical delivery of the ordered goods. Websites offer goods for sale, and sometimes make the price dependent on information they gather about potential customers.[1] These events and acts all take place in a virtual world, but also physical things have become autonomous agents. For quite some time already aircraft autopilots play an important role in air planes.[2] Weapons have been developed that take decisions about life and death without direct human interference, and the development of self-driving cars has drawn much attention.[3] The 'internet of things' is still on the brink of emerging, but it seems a safe prediction, that some of these 'things' will turn out to be autonomous systems.[4]

These new developments raise many questions. How will human beings respond to autonomous systems if they have to interact with them?[5] Is it morally right to have machines take actions in matters of life and death?[6] And – from a legal perspective – is it possible and desirable to hold autonomous agents responsible and liable for what they did?

These last questions represent the core of the present article. Before going into more detail on the arguments to come, we must first mention a preliminary question: does it make sense to write about autonomous agents and their acts if we are not dealing with human beings? Can we say that software applications and automated physical systems really act, or are the real agents always human beings, who may use software and hardware tools to realize their

* The author thanks Antonia Waltermann for valuable comments on an earlier version, and the participants in the workshop on Liability Law of the Ius Commune Conference on November 24th 2016 in Maastricht for illuminating questions at after presentation of this article.

1 See, for instance, https://www.cnet.com/news/now-showing-random-dvd-prices-on-amazon/ (last visited on 22-11-2016).

2 See https://en.wikipedia.org/wiki/Autopilot (last visited on 22-11-2016).

3 E.g. https://ikhlaqsidhu.files.wordpress.com/2013/06/self_driving_cars.pdf (last visited on 22-11-2016).

4 See the contribution of Marco Loos to the present volume.

5 See, for instance, the overview of developments concerning sex robots on http://www.mirror.co.uk/all-about/sex-robots (last visited on 22-11-2016) and the development of 'social robots' (https://www.researchgate.net/publication/308928172_A_Human-Robot_Competition_Towards_Evaluating_Robots%27_Reasoning_Abilities_for_HRI; last visited on 22-11-2016) .

6 See http://www.cc.gatech.edu/ai/robot-lab/online-publications/formalizationv35.pdf (last visited on 22-11-2016).

intentions? If the latter is the case, the issues of responsibility and liability do not arise for these other systems. As will turn out later, the answer to the question whether it makes sense to write about autonomous non-human agents is a substantial part of the answer to the question whether it is possible to hold these agents responsible. The answer to this preliminary question can therefore only be given in the context of the main argument of this article.

Although the questions with which this article deals are highly relevant for the way in which law deals with autonomous agents, they are not typical legal questions. They cannot be answered by means of law itself, and traditional legal methods such as invoking legislation, case law and doctrinal literature, are therefore not useful. The argument of this article is mainly philosophical, and the method that is predominantly used is analysis.

In the following sections, the question will be addressed whether autonomous agents can be held responsible for their acts. In this connection autonomous systems are taken to be non-human systems which do things which would be considered acts if performed by humans.

The argument, which leads to the conclusion that autonomous agents can be held responsible for their acts, will be easier to follow if its main lines are presented first. The argument is based on an analogy between human beings and autonomous agents and its main element is that if humans can be held responsible, so can, in principle, autonomous agents. This argument can only be convincing if the relevant similarities between human beings and autonomous agents are more important than the relevant differences. An important part of the argument is therefore aimed at showing precisely this. The main point here is that the argument does not claim that autonomous agents are actually like human beings, but rather that human beings are actually like autonomous agents. This analogy can only lead to the conclusion that autonomous agents can be held responsible if it is assumed that human beings can be held responsible, even if they – as the argument assumes – are like autonomous agents. This will be argued indeed, and leads to the transition from the question whether human beings and autonomous agents can be held responsible and liable to the question whether it is desirable to do so. The answer to this last question is guardedly affirmative: it depends on the circumstances, but yes, sometimes it is desirable to hold human beings and autonomous agents responsible and liable for what they did. Therefore it sometimes makes sense to do so.

2. The mental and the physical aspects of acts

The law normally holds human beings responsible for their acts. This claim is somewhat ambiguous, because of the many meanings the word "responsible" can have.[7] In this article we are interested in two senses of responsibility: responsibility in the sense of having performed some act, and responsibility as having to bear the consequences of the act, including legal liability. We will call the former act-responsibility, and the latter liability. In the present section we focus on act-responsibility.

[7] Hart famously distinguished four main kinds of responsibility (see HLA Hart, *Punishment and Responsibility*, 2nd edition, Oxford University Press 2008, p. 210-237). More elaborate distinctions are also possible.

Where act-responsibility is concerned, it is almost a tautology that human beings are typically responsible for their acts.[8] However, if humans are typically responsible while the responsibility of autonomous agents still needs to be argued, there must be one or more differences between humans and autonomous agents which seem at first sight relevant. Two differences which might make the difference are that humans act intentionally and on the basis of a free will, while autonomous agents have no intentions, nor a will, let alone a free will.

Let us assume for the sake of argument that, as a matter of fact, autonomous agents lack intentions and a will. However, they have in common with human beings that their 'behavior' has a physical aspect and that this physical aspect partakes in the causal[9] processes that govern all of physical reality. Even if human behavior is intentional, and based on a free will, what actually happens fits in the same causal chain of events in which also the 'acts' of autonomous agents fit. The question is therefore whether intention and free will make a difference; whether they play a role in the causal chains that lead to the physical aspects of acts. If they do not play such a role; they are in that sense redundant, and it seems dubious to base a difference in the attribution of responsibility on such redundant phenomena.

There are many reasons to assume that intention and free do *not* play a role in the causal chains that lead to the physical aspects of acts. The present physical research paradigm, which assumes that physical events are caused by other physical events, works quite well and this paradigm leaves no room for intervening mental phenomena like intentions or the will.[10] It is completely unclear how physical events might be influenced by mental events and there is no evidence that such influence exists.

Admittedly it *seems* obvious that mental events such as decisions influence human behavior, but when we look at, for example, the brain processes which cause muscle contractions, the obviousness disappears. The obviousness of mental events causing physical ones only exists as long as we only look at human acts, and not at their physical manifestations in the shape of muscular contractions. As a matter of fact, even where acts are performed intentionally, there is some evidence that the intention to act only comes into existence after the physical process initiating the act has already started.[11] This makes it unlikely that the intention caused this physical process.[12]

---

[8] Almost, because we typically say that somebody is responsible in this sense for bringing about particular results, and not that somebody is responsible for, say, whistling. However, when it comes to acts that consist in bringing about a particular result, having performed the act and being responsible for the act that consists in bringing about the result (e.g. closing the window, committing a murder) are the same thing.

[9] For reasons that will become clear in section 3.3, the word "causal" may be misplaced here. "Based on natural laws" would be a more adequate, but also more awkward formulation.

[10] This research program is described as naturalism in the illuminating catalogue of arguments against free will that was drawn up by Nahmias (Eddy Nahmias, 'Is Free Will an Illusion?', in Walter Sinnott-Armstrong (ed.), *Moral Psychology. Volume 4: Free Will and Moral Responsibility*, Cambridge (Mass.): MIT Press 2014, p. 1-15.

[11] See for instance Benjamin Libet, 'Do We Have Free Will' in Walter Sinnott-Armstrong and Lynn Nadel (eds.), *Conscious Will and Responsibility*, Oxford University Press 2011, p. 1-10. This evidence has been disputed for methodical reasons (e.g. Adina L. Roskies, 'Why Libet's Studies Don't Pose a Threat to Free Will', in Sinnott-Armstrong *o.c.*, p. 11-22). However, this methodical disputation does not provide evidence

This means that even if humans act on the basis of[13] intentions and free will while autonomous agents do not, this difference does not make a difference for what happens physically. It is therefore not at all obvious that this alleged difference between humans and autonomous agents should make a difference for holding agents, be they humans of autonomous agents, responsible for their acts. Whether such a difference should be made depends on the reasons why we do hold humans responsible, even if the mental aspects of their behavior do not influence the physical aspects thereof.

3. The attribution of agency and responsibility

3.1 Experience of agency

Typical human agency is intentional, meaning that the acting person experiences himself[14] as acting, and experiences the act as being brought about by him, based on his will to act. This *experience* of one's own acts as being caused by one's intention to act should not be confused with the *perception* of independently existing facts such as seeing that some person acts, or that a will causes acts.[15] The sentence Jaap saw that Esther hit Veerle is only true if Esther, Veerle and Jaap existed, if Esther hit Veerle and if Jaap saw this happening. The sentence Jaap had an experience of Esther hitting Veerle, on the contrary, can also be true if Esther, Veerle and the hitting event were all figments of Jaap's imagination. Therefore it is not without a doubt possible to conclude from the fact that Jaap had such an experience that the event really took place, or even that Esther and Veerle are real existing persons.

However, experiences are the building blocks of empirical knowledge; not infallible building blocks but nevertheless the starting point of theory construction. This also holds for the experience of acting, the experience of freely taking decisions on what to do, and the experience of causing events to happen. Experiences of agency and their abundance explain why humans beings for many centuries, if not millennia, conceptualize many events in terms of acts, agents, intentions and (free) will, and it also explains why legal discourse uses these same concepts on such a large scale and in such central contexts as the attribution of criminal and civil liability. The questions whether this conceptualization is valid in the sense that what

---

for the claim to the contrary that intention and free will do play a role in causal processes, and is therefore not very strong in the argument in favor of a role for intention and free will..

[12] It does not make it unlikely that the intention caused the act (even though it did not cause the muscular contraction that is part of the act), but that has to do with the fact that the notions of intention, causation and act all belong to the practice of agency that will be discussed in section 3.3. Within this practice it may make sense to say that the intention to perform some act caused this act (Searle's notion of prior intention, cf. John R, Searle, *Intentionality*, Cambridge: Cambridge University Press 1983, p. 91-98), but this does not imply that the practice of interpreting acts as being caused by prior intentions makes sense in the light of modern science.

[13] This formulation "on the basis of" is left deliberately vague, because, if it is not causation, it is not clear what the relation is between on the one hand intention and free will, and on the other hand acts .

[14] Here I follow the convention that if the gender of a person is not relevant for the text, an author should use her or his own gender to refer to humans. In my case this means that I use male pronouns. I can only encourage female authors to use female pronouns.

[15] This point was already made, be it in quite different wordings, by Searle in his *Intentionality*, p. 124.

is experienced also exists in reality cannot be answered solely on the basis of our experience of our own acts, however. From the fact that we experience ourselves as free agents who decide what we will do and who intentionally do what we decided we cannot derive the certainty that we do have a free will, and that this will determines what we will do. We must also look at the (physical) sciences to get a picture of what exists independent of our experience and to confront our experiences with the results of these sciences.

At their present stage of development, the physical sciences seem to leave no room for intentions and manifestations of will which *as such* cause acts.[16] The question therefore becomes how to fit these findings of the sciences with the way humans experience themselves. A plausible answer is that we should distinguish between two views of agency, the 'realist' one and the 'attributivist' one.

3.2 The realist and the attributivist view of agency

According to the realist view, the intention to act and the will that leads to the performance of an act are 'real' things, which exist independently of human experience. Our experience of them is on this view very much like perception, like an awareness of something that exists independent of the experience. Think for instance of the experience we have when we perceive a house. The house exists independent of out perceiving it, and the statement that John sees the house expresses are relationship (seeing) which exists between john and the house as independent entities. If "intention" and "free will" are interpreted as referring to 'real' intentions, and 'real' free will, intentions and will should exist independent of our experience of them. If intention and will are seen according to this realist theory, the view that human acts are caused by the intention to act and as the result of a free will, are according to modern science false. There is no reason to assume that intention and will influence the muscular movements that represent the physical aspect of our acts.

According to the attributivist view, intention and free will are attributed, or ascribed, to human agents. A person who ascribes intention and free will to himself can base this attribution on how he experiences his own acts. Attribution of intention and free will to others is based on an analogy to one's own experience. This attributivist view is supported by the fact that we do not only experience intentions in our own acts but that we also recognize them in the acts of others. Of course we cannot see these intentions, because - as the saying goes - we cannot look into somebody else's head.[17] However, going by the context in which a physical act was performed, we can recognize the intention in the act, very much like we recognize the will in juridical acts. This recognition does not concern some independently

---

[16] The clause "as such" was added to account for the possibility that mental phenomena are in a sense identical to physical phenomena – configurations of the brain are the most likely candidate – and in their quality of physical phenomena cause the muscular movements which count as acts. This view is called the 'identity theory of mind-body'. An overview of several identity theories can be found in David M. Rosenthal, 'Identity Theories', in Samuel Guttenplan (ed.), *A Companion to the Philosophy of Mind*, Oxford: Blackwell 1994, p. 348-355.

[17] Of course we can look into somebody else's head. What we cannot do is to look into somebody else's mind. However, we cannot look into our own mind either, since minds are not things that are amenable to being looked into. Our minds consist of *experiences*, and only our own experiences constitute our minds. We cannot experience somebody else's experiences, although we can have similar experiences.

existing entity, but is essentially the attribution of an intention to acts, based on facts that we can 'really' see, or perceive in some other way. If somebody does something which is understandable and which does not seem to be caused by an external force, we take it that the agent acted intentionally. And if there are no reasons to assume that the act was caused by a factor that should not have caused it, we also assume that the act was the result of a free will.[18] Since it is up to us which factor we consider as illegal causes, it is up to us to determine which acts we count as based on a free will, as voluntary. Free will is on the attributivist view a matter of attribution, very much like intention.

3.3 Expansion of the attributivist view

From a purely physical (realist) point of view, there is little ground to distinguish between acts and other events; there are only events and in some of them human bodies are involved. This is different from the attributivist point of view. From this perspective, not all events in which human bodies are involved count as acts, while some events (or even absence of events) can count as acts even though no human body was involved. This is the case when omissions count as acts.[19] So, on the attributivist view, acts are the result of attribution too, and - since agency presupposes acts - the same holds for agency.

Moreover, it also holds for causation. As any lawyer who has studied the legal notion of causation knows, it is not a discovery of an independently existing fact when some act or other event is found to be the cause of particular damage, but it is a matter of attributing the status of cause to the act or event.[20] In social and physical sciences it is not very different. The reasons for attributing the status of a cause to something may differ from one field to another, but causation is always a matter of attribution rather than discovery.[21]

This, finally, leads us to act-responsibility for an act that consists in causing some result. If agency and causation are a matter of attribution, then responsibility must be a matter of attribution too. On the realist view, responsibility does not exist. There is no responsibility to be discovered in the 'outside world' analogously to the way we can discover a pond in the forest or a birthmark on somebody's skin. We cannot discover that somebody was responsible for some act, although we can discover facts that are grounds for attributing responsibility to somebody. For instance we can discover that Jane pushed a button and use this discovery as a ground to hold Jane responsible for the fact that the light went on and the fact that a prowler

---

[18] Notice that this "should" makes freedom of the will a normative phenomenon. More on this normative notion of a free will in Jaap C. Hage, 'Rechtstheoretische analyse van de partijautonomie in het overeenkomstenrecht', in Ilse Samoy (ed.), *Evolutie van de basisbeginselen van het contractenrecht*, Antwerpen: Intersentia 2010, p. 241-262, in particular sections 26 and 27.

[19] See for the status of omissions as acts the illuminating account in Johannes Keiler, 'Commission versus Omission', in Johannes Keiler and David Roef (eds.), *Comparative Concepts of Criminal Law*, Cambridge: Intersentia 2015, p. 47-77.

[20] See, for instance, HLA Hart and T Honoré, *Causation in the Law*, 2nd edition (1st edition 1959), Oxford: Oxford University Press 1985.

[21] Notice that this claim about causation in the physical sciences is limited to the notion of causation, according to which one fact or event *necessitates* some other fact or event. The claim that causation in the sense of necessitation is a matter of attribution does not extend to the regular connections that are discovered to hold between physical facts and events. In contrast to causes, these laws *may* exist in a mind-independent reality.

was chased away.[22] Responsibility is best accounted for on the attributivist view, and is then the result of attribution, rather than a 'real' phenomenon.

3.4 The reality and relativity of what is attributed

We should be careful not to assume that attribution is really nothing. What is the result of attribution is not 'real' in the sense of existing independent of human experience, but that holds for practically everything in law. Legal rights, legal duties, competences, judges, and criminal suspects, they are all the result of the application of a legal rule and as such the result of attribution.[23] We should not make the mistake to assume that what is the result of attribution does not really exist. What is the result of attribution is by definition not 'real' in the sense of mind-independent, but it does exist as the result of attribution. There are facts which are the result of attribution, and the sentences describing them are true in the same way as sentences describing 'real' facts. The sentence that Willem-Alexander was in 2015 the King of the Netherlands is as true as the sentence that in that same year the North Sea bordered on the English east coast.

This parallel to what exists in law is useful to illustrate that attribution is not always what one single person attributes; it can also be what 'we' attribute, or what the law, in the shape of rules, attributes. A person can hold himself responsible for some accident, while others do not agree that he is responsible. The law can have its own view as to whether this person is responsible for the accident. Therefore responsibility is in a sense always relative to one or more persons who, or to a set of rules which attribute(s) responsibility.

3.5 Attribution to autonomous agents

What is 'real' does not depend on attribution and is mind-independent. What is the result of attribution, on the contrary, depends on human minds. This mind-dependency may be direct, as when somebody considers what somebody else 'does' as an act. It may also be indirect, as when the members of a tribe attribute the failure of rain to the anger of the gods which makes the anger of the gods the cause of the rain failure even if some members of the tribe do not believe this. It is also indirect when the law attributes the status of owner of a house to the daughter of a deceased person who used to be the owner. (The daughter is taken to have inherited the house.)

Because attribution is mind-dependent, agency and responsibility may theoretically be attributed to anything, and on any ground. It is possible to consider events as the acts of animals or of gods, or as the acts of organizations, and we may hold animals and organizations responsible and liable, even criminally liable, for their 'acts'. If we can attribute agency to organizations and hold them responsible and liable for their acts, we can do the same for autonomous agents. From the perspective of what can be done, there are no difficulties for the attribution of agency, responsibility and liability to autonomous agents.

---

[22] Cf Donald Davidson, 'Actions, Reasons and Causes', in Journal of Philosophy 60 (1963). Also in his *Essays on Actions & Events*, Oxford: Clarendon Press 1980, p. 3-20.

[23] Cf. N. MacCormick and O. Weinberger, *An Institutional Theory of Law* Dordrecht: Reidel 1986.

The question is not whether such attributions *can* be made, but whether it is *desirable* to do so.

4. The desirability of attribution

We attribute agency and responsibility to humans, and the reasons why we do so may illuminate our thinking about the attribution of agency and responsibility to autonomous agents.

4.1 Intuitive and reflected attribution

The attribution of agency and act-responsibility to humans beings is in first instance not done with a particular purpose in mind. It is more likely the outflow of how we experience ourselves as involved in agency, an experience which is projected on other humans. Whether we will attribute agency and act-responsibility to autonomous agents depends in first instance on our propensity to see these systems as similar to human agents. In this connection it should not remain unnoticed that we often in first instance seem to attribute agency to non-human actors, as when we say that the dog is asking to be walked, or that the computer formatted our text wrongly. To say that such attributions are merely metaphorical may reflect that we do not think it desirable to make such attributions, but ignores our intuitive analogizing between human and non-human 'behavior'.

However, we do reflect on whether it is desirable to attribute agency to animals and organizations, and - in extreme cases such as severe mental illness - even to human beings. Moreover, if the outcome of this reflection is negative – it is not a good idea to attribute agency – the result is often that we stop attributing agency. What originally seemed to be an agent – the dog that 'asks' to be walked, or the computer that 'asks' for our password – turns on closer inspection only to be similar to an agent.

4.2 When attribution of intention is desirable

When is it desirable to attribute agency to some system, be it human, otherwise animal, or non-animal? To answer a closely related question – the question when it is desirable to attribute intentional states such as beliefs and desires to a system - Dennett introduced the notion of the "intentional stance".[24] Dennett contrasted the intentional stance to the astrological, the physical and the design stance.

On the astrological stance, we explain the past behavior of a system and predict its future behavior on the basis of the date and time of the origin of this system and of the 'laws' of astrology. For example, we would predict that somebody will soon find a lover from the date he was born. This stance is never desirable, because the predictions will fail too often and therefore also the explanations become untrustworthy.

The physical stance is desirable if we succeed in predicting the future behavior of a system and explaining its past on the basis of the systems physical characteristics and our knowledge of the laws of nature. This physical stance works well for much of the non-animal part of the

[24] Daniel Dennett, 'Intentional Systems' in his *Brainstorms*, Brighton: The Harvester Press 1981, p. 3-22, and 'True Believers' in his *The Intentional Stance*, Cambridge (Mass.): The MIT Press 1987, p. 13-35.

world, and many scientists expect it to work just as well in the long run for the animal part. As a matter of fact, large parts of medicine are presently based on the adoption of the physical stance towards human bodies.[25]

The behavior of some non-animal systems can even better be predicted and explained from the intentions of their makers than from their purely physical characteristics. Take for instance a television set. If we push a button on the remote control, the set turns on. That it will do so is in our experience so natural that we do not even need the user manual of the set to switch the TV on in this way. We can predict the behavior of the set, because we expect it to be designed to turn on if the button is pressed. Theoretically it should also be possible to predict the reaction of the set to the button push from the set's physical characteristics, but that is so difficult that most humans are not able to do so. The design stance - assuming that the set was designed to operate in a particular way – is in this case more fruitful than the physical stance. Therefore it is desirable to take the design stance towards television sets.[26]

It should be noted that the physical and the design stance do not exclude each other. We can fruitfully both approach a television set as a purely physical system and as a system designed to fulfil certain functions. It is a matter of what works best whether the physical or the design stance should be adopted, and it is for instance not so that we should take the design stance to things that were designed. If a television set is defect, we'd better approach it from the physical stance. Moreover, it is useful to adopt the design stance to the results of evolutionary processes, even though evolution is not a matter of design. For instance, it makes sense to ask what the purpose is of the long neck of a giraffe in order to explain then neck's length and to predict whether the length will still increase in the centuries to come.[27]

The intentional stance, finally, should be adopted towards a system when the behavior of the system is best explained and predicted if we attribute beliefs and desires to the system. This works well for human beings: we can predict the presence of students in the lecture hall from their desire to learn (about the exam) and their belief that the lecture will reveal what will be asked on the exam. It also works quite well for higher animals: we can explain the search behavior of a chimpanzee from its belief that there is food hidden at some place, and its desire to have the food.[28] It even works for non-animal systems: we can predict the move of a chess-playing computer program from its desire to check-mate and its belief that a certain move will check-mate the opponent.

It may be objected to the last example that the behavior of the program can also be explained from the design stance. The observation that the design stance may also lead to explanations of the program's past behavior and correct predictions of its future behavior may be correct,

---

[25] Notice the move from "humans" to "human bodies". Not everybody likes this move in medicine.

[26] That the set is a television set, rather than a heap of electronics, is also part of approaching the 'heap' from the design stance.

[27] However, some see the fact that the living parts of nature are fruitfully approached from the design stance as a reason to assume that these parts of nature must have been designed. This is a central theme of Daniel Dennett, *Darwin's Dangerous Idea*, London: Penguin Books 1995. See also Richard Dawkins, *The Blind Watchmaker*, London: Longman 1986.

[28] Cf. Frans B.M. de Waal, *Are We Smart Enough to Know How Smart Animals Are?*, London: Granta Books 2016.

but the adequacy of the design stance does not preclude an even greater adequacy of the intentional stance. In fact, the two may be combined in a fruitful fashion: assume that the system was designed to use a strategy (try to check-mate) and to employ its knowledge of the board position, the possible moves, and the value of the pieces to execute its strategy. The adoption of the design stance may in this fashion lead to the adoption of the intentional stance and to the attribution to the chess-playing program of desires (the strategy) and beliefs (knowledge of the board position, the rules of chess, and an evaluation function for board positions).

4.3 When attribution of agency is desirable

The step from chess-playing programs to 'intelligent' physical systems such as cruise missiles and self-driving cars is only minor. Although it is possible to approach them from a purely physical stance, and also from a design stance, their behavior is most fruitfully explained from the assumptions that these systems reach for some goals (destroy the target; arrive at a particular destination), and that they employ their knowledge (about their geographical position, about the location of defense-missiles, or about the availability of roads) to achieve their goals.

Moreover, it is also a minor step from the intentional stance to the agency stance, if these two can be separated at all. The cruise missile adapts its course to avoid defense-missiles, or to correct for heavy winds; the self-driving car breaks to avoid a collision, or takes a detour because it received info that some road is blocked because of road works. Adapting course and breaking are best explained and predicted on the assumption that these systems *act* on the basis of their strategies and their knowledge. To this extent it is desirable to attribute agency to some autonomous agents.

In addressing the question whether and when it makes sense to adopt the intentional stance towards systems, Dennett focused in particular on the facilitation of predictions about the system's future behavior. It is desirable to adopt the intentional stance towards a system if this makes it easier (than any other stance) to predict the system's behavior. Analogously one might argue that it is desirable to adopt the agency stance towards systems if this makes it easier (than any other stance) to predict the systems behavior. However, it is not unavoidable to make predictive power the one and only criterion for the desirability of the adoption of the agency stance attributing agency, or – which boils down to the same thing – for the attribution of agency. Other criteria, next to or even instead of, predictive power may also be used. For example, one may have religious reasons to attribute agency only to human beings. Or one may believe that in the end science will reveal that there are substantial differences between humans and other systems, which justify the view that only humans should be seen as agents. However, the somewhat artificial nature of these examples suggests that predictive power may be the best criterion for which stance to adopt, although there is no ground to postulate that it is the only possible criterion.

5. The attribution of liability

The attribution of agency can be contemplated from the perspective of explanatory and predictive power. However, agency can also be contemplated from the perspective of

purpose: why should we call something an act and attribute this act to an agent? The answer to this latter question will often be that we attach consequences to agency. If an agent is act-responsible for some event, this is a ground to call this even an "act" and to hold the agent liable for what he did. Moreover, this relation between the two kinds of responsibility works in both directions: we attribute agency because we want somebody to be liable and we hold somebody liable because he performed the act in question and - for some cases of liability - because he thereby caused the damage.

This raises the question why we would want to make an agent liable. The question is very abstract, and to make the attempt to answer it more concrete, we will focus on two questions, assuming that the answers will be related:

1. Why should we want to make one person (a human being) liable for the damage suffered by some other person?
2. Why should we want to make an autonomous agent liable for the damage suffered by some person?[29]

The questions do not distinguish between contractual and tort liability, because *prima facie* there seems to be little reason to make the answer to the general liability question dependent on whether there is a contract. Of course that does not exclude that the answer to the question whether there is liability in a concrete case may depend on whether the basis for liability is tort or contract, or still something else.

The answer to the first question can go into two directions. The one direction is that somebody should be liable for somebody else's damage because he deserves to be liable, most likely because he caused the damage. We will consider this direction in more detail in section 6. The other direction is that liability serves some purpose, such as prevention of future damage, the achievement of distributive justice, or popular satisfaction. Popular satisfaction as ground for liability is strongly connected to the idea that the liable person deserves to be liable, because the idea that the person who caused damage must compensate it is quite popular and the effectuation of this idea may therefore lead to greater popular satisfaction. We consider the purposive direction in section 7.

6. Deserved liability

One representative of the view that an agent is liable for particular damage because he deserves to be liable is Ripstein.[30] According to Ripstein, tort law is based on two principles. The one principle is that people accept reciprocal limits on their freedom. For our present purposes, this principle is not very relevant. The second principle is that people bear the costs their unlawful conduct imposes on others. This principle is reformulated as the principle that if one person wrongs another, the latter's loss becomes the former's to deal with. It is this last formulation which suggests the relevance of desert for liability. Apparently the fact that the behavior that caused damage was a wrong is a reason to make the wrongdoer pay for the

[29] These questions exclude criminal liability, although *prima facie* there is little reason to treat criminal liability differently from civil liability.

[30] Arthur Ripstein, 'Philosophy of Tort Law', in Jules Coleman and Scott Shapiro (eds.), *The Oxford Handbook of Jurisprudence and Philosophy of Law*, Oxford University Press 2002, p. 656-686.

damage. This only makes sense if a wrongdoer somehow deserves to become liable, because if that is not the case, it is still an open question why somebody who violates a norm of conduct should become liable for the resulting damage. That question might be answered by pointing out desirable consequences of moving liability from the victim to the wrongdoer, but then the justification is not based on the wrongness of the act anymore, but on the purpose of moving liability. Seeing in wrongness a reason to attribute liability makes only sense on the assumption that a wrongdoer somehow deserves to become liable.

6.1 Justification within a practice and justification of a practice

When it comes to the explanation how tort law operates, or – perhaps better - how tort law has operated in the past[31], Ripstein is likely to be right. Somehow people assume that it is not always allowed to damage somebody else's interests, and that if one does so nevertheless, this is a reason why the tort-feasor has to compensate the damage that resulted from his illegal behavior. This is part of the practice of tort law, a practice which also assumes that it is possible to identify acts, agents, and damage, and to identify some act as the cause of particular damage. Working from this practice it is possible to attribute particular damage to a particular act and a particular agent, and to attribute responsibility for the damage to the agent. That this practice presupposes that the agent deserves to be liable can be seen from the demand, now less commonly made than it used to be, that the agent could be blamed for his act. Given this practice, the judgement that some person is liable for the damage of some other person can be justified if the conditions for liability specified by the practice are satisfied. This is how tort law operates, and how tort law justifies liability.

However, justification of liability on the basis of tort law only makes sense if the practice of tort law itself makes sense. To see this dependence of justification from within a practice on the justification of the practice itself, it is useful to consider a practice which many people do not consider to be justified, the practice of rain dancing. This practice uses dancing as a way to make it rain. Suppose that a rain dance, if it is to work, must satisfy certain demands, such as a sacrifice of a goat preceding the dance. Is this sacrifice justified? Yes it is from the perspective of the practice, because if the sacrifice is not made the rain god may be insulted and does not reward the ensuing dance with rain. However, if we step outside the practice, as the readers of this text will most likely do automatically, the sacrifice does not make sense, because there are no rain gods that can be insulted by not sacrificing a goat, and a rain dance will not work anyway. Therefore the sacrifice of the goat, if it makes sense at all, only serves to placate the ignorant observers who demand this sacrifice to be made.

Perhaps many readers will not like the parallel, but a similar argument may be made about the practice of tort law which attributes liability to agents because they deserve to be liable because of their unlawful act. Only if this practice as a whole makes sense, including its presuppositions about acts, agents, damage, and causation, it makes sense to justify the attribution (and enforcement!) of liability from within this practice.

---

[31] As is well-known, there is a development in tort law from fault liability (based on desert) to strict liability. This development is in line with the argument of this article, although the view that liability is the result of purposive attribution was not necessarily the main cause of the development as it has actually taken place.

6.2 The hermeneutic fallacy

It is tempting to justify the practice of tort law by trying to understand it 'from within'.[32] The justification then consists of identifying the elements of the practice and justifying them on the basis of the principles underlying the practice. This is what Ripstein did when he tried to justify tort law by identifying two principles which could in his opinion explain the more detailed rules of tort law. We find a similar approach in the work of Weinrib, when he tried to understand – but understanding is meant as justification here – private law from its internal purposes, rather than from external purposes which are served by it.[33] Moreover, the same approach to the attribution of liability can be found in the work of Dworkin, who also claims that the practice of law must be understood from within.[34]

Understanding is a psychological category, which has no other standards than the feeling that one understands and there is nothing wrong with attempts to understand a practice from within if such attempts lead to the feeling of understanding. However, given the psychological nature of understanding it cannot justify a practice that is understood. It is a fallacy to assume otherwise, and we may well call this common fallacy the hermeneutic fallacy: a social practice is justified if we understand why it does what it does. This hermeneutic fallacy is a special case of the fallacy to derive how something ought to be done from the way it is actually done. From the fact that some social practice, e.g. tort law as it used to function, exists, it cannot be concluded that it should continue to function that way, or that the practice should continue to exist. This does not mean, of course, that the practice is wrong, but only that the reasons for its existence cannot be found within the practice itself.

Rain dancing cannot be justified by understanding it, and neither can liability law. So, even if Ripstein is correct in his finding that tort law is amongst others based on the principle that people bear the costs their conduct imposes on others, this does not show that it is right to attribute liability for damage to some person for the reason that this person caused the damage and therefore deserves to be liable.

6.3 Capacity and desert

The argument of the previous two subsections only rebuts one type of argument why liability is based on desert, the argument from understanding the practice as it actually is; it does not attack that conclusion itself. It is now time to address the conclusion: should people be held liable for damage for the reason that they deserve to be liable?

---

[32] Lucy writes in this connection about the "methodological injunction": the demand that "… any adequate theoretical account (or explanation or understanding) of any social action, practice or institution must, in first instance, capture the way in which that action, practice or institution is understood by those whose patterns of behavior and thought constitute that action, practice or institution" (William Lucy, *Philosophy of Private Law*, Oxford: Oxford University Press 2007, p. 272. As long as this methodological injunction is seen as a technical norm, prescribing what is necessary if one wants to understand a social practice, it may be well defendable. However, it becomes problematic as soon as this understanding of a social practice starts to function as a justification.

[33] See Ernest J. Weinrib, *The Idea of Private Law*, Oxford University Press 1995, chapter 1.

[34] Ronald Dworkin, *Justice for Hedgehogs*, Cambridge: Harvard University Press 2011, p. 227-231. On page 229 Dworkin almost explicitly presents interpretation as a means of justification.

The answer to this question depends on whether people sometimes deserve to be liable. If nobody would ever deserve to be liable, it is rather useless to sustain the practice of holding people liable because of desert: nobody would ever be liable. As a matter of fact, we do assume that people sometimes deserve to be liable and we do sometimes attribute liability to persons because they deserve it. However, the question is not whether this practice exists, but whether it is justified, whether it makes sense. That depends on the reasons why we assume that people sometimes deserve to be held liable. Which reasons are actually recognized is a matter of empirical investigation, but here we will investigate one such a reason which may be the most important one. Somebody deserves to be liable because through what he did he caused damage to somebody else and he should not have done what he did because of the damage it would cause.[35] This reason makes sense on the presumption that people can avoid wrongful acts for the reason that they recognize that wrongfulness. In the vocabulary that is nowadays fashionable one might say that desert depends on reason-responsiveness.[36] Are people reason-responsive in the sense required for the assumption of desert? It depends on how we understand reason-responsiveness, but it will be argued that in the most relevant sense people are not reason-responsive.

However, first we need to consider a brief note on the attribution of capacity. The question what a person could, or could not, do depends in an important sense on the capacities we attribute to this person. Capacity and therefore also reason-responsiveness are a matter of attribution, just like act, agency, causation, responsibility and liability. The argument for this claim will be presented in section 6.6, but if we assume that the claim is correct, the observation - or rather attribution - that somebody is reason-responsive does not help us much further in our search for the foundation of liability. The broad practice of agency attribution, which does not only include the attribution of agency, but also the recognition of acts and causation, cannot be justified by the nth element of this very practice. In order to avoid this complication we will replace the question whether a person could have avoided his damage-causing behavior by the question whether it would have been possible that he avoided this behavior. The questions may seem the same, but the crucial difference is that the former question is typically answered through the broad practice of agency attribution which includes the attribution of capacities, while the latter question has no direct ties to agency and does therefore not depend on this practice.

6.4 What is a capacity?

Let us assume for the sake of argument that an agent has the capacity to do something if it is possible that he does it. But what does that mean? Possibilities are the most interesting in case they were not realized, because if something is the case it is obvious that it must have been possible. It is notoriously difficult, however, to establish the existence of possibilities in case

---

[35] Not coincidentally, this reason is very close to the second principle underlying tort law which was identified by Ripstein.

[36] See for instance Stephen J. Morse, "Criminal responsibility and the disappearing person", *Cardozo Law Review* 28 (2007), p. 2545-2575, and Michael McKenna and D. Justin Coates, "Compatibilism", in *The Stanford Encyclopedia of Philosophy* (Summer 2015 Edition), Edward N. Zlatan (ed.), URL = <http://plato.stanford.edu/archives/sum2015/entries/compatibilism/>

they were not realized. To deal with this problem, a thinking device was constructed: possible worlds theory.[37] The basic idea underlying possible worlds theory is that something is necessary when it is the case whatever else may be the case. For instance, whatever the other facts may be, in any case every colored object has a surface and whatever the other facts may be, the number 5 is bigger than the number 3. Therefore, necessarily every colored object has a surface and necessarily 5 is bigger than 3. A different way of expressing that something is the case whatever else may be the case is to say that it is the case *in all possible worlds*. In all possible worlds every colored object has a surface and in all possible worlds the number 5 is bigger than the number 3.

Something is possible if it is the case *in at least one possible world*. The actual world consists of all the facts as they happen to be, while a different possible world contains a set of all facts as they might have been under different circumstances. In the real world John has brown hair, but under different circumstances, in some other possible world, John is red-headed. Because there is some alternative possible world in which John is red-headed, it is possible that John is red-headed. In fact, he is not, but he might have been. Something is possible if it is the case in some possible world. That may be the actual world, but that is not necessary.

This idea can also be applied to acts and agents. In the actual world, Jane did not visit her mother, but in some other possible world she did. Therefore, actually Jane did not visit her mother, but it would have been possible that she did so. In this sense, Jane had the capacity to visit her mother. This captures the notion of a capacity quite well. That would mean that an agent has the capacity to do something (including omission) if there is some possible world in which he does it. Notice that this notion of capacity avoids the use of attribution, at least that seems at first sight to be the case.

6.5 Possible worlds and constraints

We now have a definition of what it means that a person has a certain capacity, but it may seem that this definition has replaced one problem, the nature of capacity, with another problem, the nature of a possible world. What makes a set of facts a possible world? Here the notion of a constraint plays a role.[38] Not all facts can go together. To give an obvious example: the fact that it is raining (here and now) cannot go together with the fact that it is not raining. Incompatible facts cannot be part of one and the same possible world. That is a constraint on possible worlds. A logical constraint in this case, because it is a matter of logic that these facts cannot go together. Next to logical constraints, there can also be physical constraints. The laws of physics can be interpreted as constraints on worlds that are physically possible. It is, for instance, physically possible that a metal bar is red, but it is physically

---

[37] The idea of possible worlds theory can be traced back at least to the German philosopher Leibniz (1646-1716), who in his *Theodicies* defined necessity as that what is the case in all possible worlds. A technical account of possible worlds, under the for logicians usual name of model-theoretic semantics or model theory, can be found in several chapters of Brian F. Chellas, *Modal logic; an introduction*, Cambridge: Cambridge University Press 1980.

[38] More about constraints and their relation to possible worlds in Jaap Hage, 'The (onto)logical structure of law', in Michal Araszkiewicz and Krzystof Pleszka (eds.), *Logic in the Theory and Practice of Law Making*', Cham: Springer 2015, p. 3-48.

impossible that a metal bar is heated but does not expand. There is no physically possible world, no world that satisfies all the physical constraints, in which a metal bar is heated but does not expand. And neither is there a physically possible world in which something travels faster than light in vacuum.

We are now in a position to define possible worlds more precisely. A possible world is a world that satisfies a set of constraints. A logically possible world satisfies the laws of logic; a physically possible world satisfies the laws of physics. A particular world counts as possible if it satisfies a set of constraints. Only relative to such a set does it make sense to ask whether something is possible or necessary. Necessity or possibility *tout court*, without being made relative, does not make sense.

Both logically and physically it is possible that John is red-headed, but is it still possible if we take into consideration that John just finished dying his hair brown? That is apparently not the case, and it is worthwhile to consider more closely why that is not the case. Both with logical and with physical necessity (and possibility) the necessity is the result of constraints that consist of laws, the laws of logic and of physics respectively. A law expresses a necessary general connection between *types* of facts, for instance the type of fact that something is a metal bar that is being heated and the type of fact that this something expands. When we speak of possible worlds, such laws are the most obvious constraints to take into account. However, it is not necessary to take only laws into account as constraints. There is no fundamental reason why particular facts should not be considered as constraints too. One such a fact might be that John just finished dying his hair brown. Given that fact, it is necessarily the case that John's hair is brown, and impossible that his hair is red. And given the fact that the train Marco was on departed five minutes ago, it is impossible that he was seen at the railway station one minute ago. In particular in connection with capacities it is important not to take only laws into account as constraints on possible worlds, but also facts. If it is claimed that Jane could not visit her mother, this claim will probably not only be based on the laws of nature (purely physical necessity), but also on facts concerning Jane's personal history.

6.6 The relativity of capacity

An agent has the capacity to do something if there is a possible world in which the agent does it. Now we know that this specification of capacity is still too vague: we also need to specify relative to which set of constraints the capacity exists. The crucial question is: which set of constraints should be taken into account in determining whether a particular agent had the capacity to perform some act, or to refrain from performing it. Here we will not attempt to answer this question in abstract, but focus merely on the characteristics of individual agents.

It is clear that in determining the capacities of a particular agent, we should take some personal characteristics of this agent into account. Going only by the laws of logic and of physics, which are the same for everybody, every agent would have the same capacities. That would be an unattractive finding, and to avoid it, we must take personal characteristics into account in determining which capacities some agent has. [39] But which personal characteristics

[39] A related account of what an agent can or cannot do, is Tony Honoré, 'Appendix: Can and can't' in his *Responsibility and Fault*, Oxford: Hart Publishing 1999, p. 143-160.

should be taken into account? If the agent cannot drive a car, we should most likely take that into account. So if Jane could not drive a car, she did not have the capacity to visit her mother (let us assume) and most likely she should not be held responsible for not visiting her.[40]

Stepping back from this casuistry, the general issue is the following: if all facts regarding an agent are taken into account, as well as all physical and other possibly relevant laws, there are two possibilities. On a deterministic world view, only one possible course of the world is possible if all previous facts and all laws are taken into account. On a non-deterministic view, it is arbitrary what the course of the world will be, even if all previous facts and all laws are taken into account. However, on both views it does not seem reasonable to attribute responsibility to an agent for what he did. In the deterministic case not because the agent could not do anything else than what he actually did. In the non-deterministic case because it is arbitrary what the agent does, and therefore not dependent on the agent himself.[41] There is no middle way according to which the agent determines what he will do, because everything about the agent that might be relevant for the determination what he will do is *ex hypothesi* included in the set of all constraints. Given these constraints the agent is either determined to do what he will do, or his act will be arbitrary, and therefore not dependent on the agent himself. In neither case there is a ground for attributing responsibility.

The distinction between what an agent did and what he had the capacity to do makes only sense if not all facts are taken into account as constraints on what is possible. For instance, in determining whether Charles had the capacity to avoid the car accident we do take into account that in general Charles is capable to drive a car, but we do not take into account that on this very occasion he was distracted by his quarrelling children on the back seat. Therefore we conclude that Charles could have avoided the accident, and since he did not avoid it, we hold him responsible and liable for causing the accident.

When we take this approach, the question arises which facts should be taken into account, and which facts should not. Capacity becomes a normative issue, the issue which facts *should* be left out of consideration to determine what else the agent could have done next to what he actually did. Perhaps this seems an acceptable approach; after all it is what lawyers are actually doing when they ask whether a defendant could have acted differently than he actually did. We should realize, however, that if we make capacity a normative notion, it becomes a matter of attribution (as already mentioned in section 6.3) and we can no longer adduce the capacity of an agent as a reason for holding the agent responsible. What we actually do is to give one single normative judgment concerning both the capacities and the responsibility of the agent. Either we judge the agent to have the relevant capacities and to be responsible, or we judge him to lack the capacities and not to be responsible. This judgment cannot be founded in the capacities of the agent, because these capacities are themselves part of the judgment. To do otherwise would be a circular argument.

---

[40] That might be different if it was Jane's fault that she cannot drive, but for now we will ignore the possibility of responsibility without capacity and constructions like *culpa in causa*.

[41] Cf the first four essays in Timothy O'Connor (ed.), *Essays on Determinism and Free Will*¸ Oxford: Oxford University Press 1995.

6.7 Conclusion on deserved liability

To summarize the above argument: we are left with two options. The one is to take all constraints into account in determining what an agent had the capacity to do. If we take this approach, the behavior of the agent is either determined or arbitrary, depending on whether one accepts determinism. In neither case the agent deserves to be held responsible and liable for damages. The other option is to count some constraints in and to discount some other constraints in determining the capacities of an agent. This is what lawyers actually do. However, then the social practice of counting or discounting constraints determines capacity: capacity has become a matter of attribution too. This practice cannot be adduced as justification of the practice of agency attribution. If we assume that responsibility should be attributed on the basis of desert, capacities would be the main justification for the practice of agency attribution, and if this justification does not cut ice, the practice would not be justified at all.

7. Purposive attribution of liability

However, desert is not the only possible foundation for a practice of agency attribution, including the attribution of responsibility. The alternative is that agency and responsibility are attributed for some purpose. Attribution based on desert is backward looking and requires that the past could have been different than it actually was. If determinism is correct, the past could not have been any different and it does not make sense to attribute agency and responsibility because the agent deserved it. If determinism is incorrect, the past was a random affair, and also then it does not make sense to attribute agency and responsibility because the agent deserved it.

This dilemma is avoided if attribution takes place in order to serve some purpose, because such a practice would be forward looking. It would make sense if what we do now, attribute agency and responsibility or not, influences the future. In section 7.2 it will be argued that purposive attribution of agency makes sense for humans and in section 7.3 the argument will be extended to autonomous agents. However, first we will in section 7.1 rebut an objection against purposive attribution, an objection based on the misguided idea that according to determinism it does not matter what we do or decide. Finally, we will in section 7.4 devote a few words to the way in which responsibility for autonomous agents can be implemented.

7.1 Determinism and fatalism

Any act that is performed in order to bring about future effects presupposes that what happens now influences the future. This means that events should not follow each other in an arbitrary fashion. The question whether this also presupposes determinism, the view that the future can only develop in one single way, is not easy to answer, but if determinism is true this does not make purposive action senseless. The view that this is different since if everything is determined it does not make any difference what we do now, rests on a confusion between determinism and fatalism. Because this confusion is quite common, it is useful to devote some attention to it here.

Fatalism is the view that the future will end up in a particular fashion anyway and that what we presently do cannot influence that outcome. This is compatible with the idea that this unavoidable outcome can be reached along different paths. To use a somewhat pessimistic example: whether you exercise a lot or not, in the course of time your strength will decay anyway.

A determinist, on the contrary, can adhere to the view that what we presently do influences the future. The future is determined by everything that is presently the case, but this everything includes what we presently do. Therefore our decisions and acts do influence the future. A different question is whether we are free to act and whether more than one act is possible. If determinism is true, that is not the case: we are only capable to do what we actually do. However, that does not mean that it does not matter what we do or that deliberations on what to do are useless. The deliberations can influence our acts, even though we are not able to deliberate differently than we actually do.

If determinism is true, there is only one possible future, and this seems to confirm fatalism. However, this appearance is deceptive, because fatalism does not only involve that a particular outcome will certainly occur; it also involves that what we presently do has no influence on this outcome. This latter element does not follow from determinism and it is this element that would make purposive action senseless. We see that purposive action is incompatible with fatalism, but compatible with determinism because determinism does not imply this second element of fatalism.

7.2 Three grounds for attributing responsibility to human agents

The practice of attributing responsibility, both act- and liability-responsibility, may be justified if the attribution of responsibility brings desirable consequences. This raises two questions, one evaluative and the other empirical. The evaluative question is which consequences are desirable. In this article the evaluative question will be avoided, by postulating three purposes which are deemed to be desirable[42]:

- influencing future behavior of the agent to whom responsibility is attributed,
- influencing future behavior of other agents, and
- satisfaction of the desire of people who believe that agents should bear the consequences, of what they did.

The empirical question is whether the attribution of responsibility to human agents contributes to the achievement of these purposes. The answer to this question can only be found through empirical research, and the following attempt to an answer is therefore speculative.

It seems likely that if agents are held responsible for their acts, and if they believe that this is the case, this may influence their behavior. It seems also likely that other agents will be influenced by the belief that agents in general are responsible for their acts. Finally, it is likely that presently many persons will feel satisfaction if persons are held responsible for their doings because they (falsely) believe that agents deserve to be held responsible for what they

[42] The author adheres to some form of utilitarianism, and believes that the three mentioned purposes are merely instrumental to the maximization of happiness. Cf. Jaap Hage, 'Recht en utilisme. Een pleidooi voor utilisme als richtlijn voor de wetgever', in *Law and Method*, November 2015, DOI: 10.5553/REM/.000011.

do. Perhaps this belief will disappear in the (far) future, but as long as it is still prevalent, it can underlie the practice of holding people responsible.

7.3 Should autonomous agents be held responsible?

We spent quite some time considering the argument that humans should be held responsible because they deserve it. Our conclusion was that this argument cannot support its conclusion, and that we need other grounds for attributing responsibility to humans. In the case of autonomous agents it is most likely not necessary to spend time on arguing that they do not deserve to be held responsible, since the reasons why humans do not deserve responsibility hold *a fortiori* for autonomous agents: their 'behavior' is either determined or arbitrary, and in neither case it can be said that autonomous agents deserve to be held responsible for their doings.

Because it is so obvious that autonomous agents do not deserve to be held responsible, the third purposive reason for holding them responsible may not apply to them either. Since most people do not believe that autonomous agents are 'real' agents, and therefore also not that they deserve to be responsible, there is little gained by attributing them responsibility.[43]

Because human agents presently do not identify themselves with autonomous agents, the argument that autonomous agents should be held responsible because this influences the future behavior of other agents is weaker in the case of autonomous agents than it is in case of human agents. Weaker, but not completely without force. One reason that the argument still has some force is that humans might believe that if even autonomous agents are held responsible for their 'doings', then certainly humans agents will be held responsible and that this belief influences the future behavior of human agents. A second reason is that autonomous agents might 'reason' that if other autonomous agents are held responsible, they will also be held responsible, and that this belief influence the future behavior of these autonomous agents. However, this second reason presupposes that being held responsible can influence the future behavior of autonomous agents at all. This is a crucial presupposition and we will turn to it now.

Will the future behavior of autonomous agents be influenced by the fact, if it were to be one, that they are held responsible for their doings? Asking the question suffices for making it clear that it cannot be given a general answer: it depends on the nature of the autonomous agent. Let us first consider autonomous agents which are 'mere' computer programs such as programs involved in e-trade or in taking of (simple) administrative decisions. Such programs can be less or more sophisticated, and their reaction to responsibility depends on their sophistication.

If a program is relatively simple it does not take responsibility into account, and then attributing responsibility to it will have no effects on the program's behavior. Then a practice of attributing responsibility makes little sense, at least not for these specific programs.

---

[43] However, Joanna Bryson pointed out to me that (some) people do think that autonomous agents who earned money, e.g. by stock trading, deserve to be taxed. Popular satisfaction would then be increased by making these systems liable to taxation.

A program can also be more sophisticated in the sense that its programmer has taken into account that the autonomous agent running this program will be held responsible. For instance, the program may be made extra careful not to harm other agents, be they human or non-human. If it is allowed to be a little disrespectful to human dignity, this may be compared to eugenics in which the human genome is manipulated to make humans more obedient to rules. Perhaps we would not and should not want this with respect to humans, but in case of non-human agents this might be a desirable development. If attributing responsibility to autonomous agents would contribute to this development, it would be a reason for having this practice.[44]

The variant in which a programmer takes into account that its product will be held responsible makes the effects of responsibility dependent on the reaction of a different system – most likely a human being – to the responsibility of the autonomous agent. The parallel with human responsibility would be bigger if the autonomous agent itself reacted to being held responsible. An intelligent program may possess knowledge about its potential responsibility and take this knowledge into account in deciding what it will do. This knowledge may be generally available, but may also be the result from being held responsible on a particular occasion. The adaptation of behavior to potential or actual responsibility presupposes that the agent is not focused on a single task such as taking a particular kind of administrative decisions or conducting e-trade, but that it performs tasks like that in the context of wider tasks such as contributing to the well-being of society, or the maximization of its profits. Presently there are, to the author's knowledge, no practically functioning systems which can do this, but for the theoretical question that is not very relevant. If such systems would exist - and it is quite likely that they can already be created - it would make sense to hold them responsible for their doings, both in abstract as well as in concrete cases.

For systems which are also physically autonomous, such as self-driving cars, cruise missiles, and some robots, the situation is not fundamentally different from that of autonomous agents which are merely computer programs. A computer programs runs on a machine which typically has peripherals for input and output. The only thing that makes physically autonomous systems different from mere computer programs is the nature of the peripherals. These different peripherals can make that the nature and impact of what the physically autonomous system does radically differ from what the mere computer program does, but for the fundamental issue whether it makes sense to hold them responsible this does not seem to make any difference.

So there seem to be no fundamental reasons why autonomous systems should not be held liable for what they do. However, this fundamental argument does not show that it is presently desirable to hold autonomous systems liable. Making the programmers or the users of autonomous systems liable for what the systems do may be more efficacious from a purposive

---

[44] Elbert de Jong raised the question whether programmers who can make programs adapt to liability cannot also make them avoid damage causing behavior. That would directly lead to the desired consequences (less damage) and therefore be preferable to the indirect route of making autonomous systems liable. Liability would still make sense, however, for those cases where a system has a conflict of duties. Then the system can comply with one of the duties and compensate the damage that results from not complying with the other duty.

perspective. Whether this is the case is an empirical question, and the answer may vary from system to system and may change in the course of time. However, the argument of this paper should have made clear that although the practical desirability may still be an open question, the difference between humans and autonomous systems as such does not justify a different treatment as far as liability is concerned.

7.4 How to implement the responsibility of autonomous agents?

Actually it is not part of the question this article aims to answer, but for practical purposes it is important to know how the responsibility of autonomous agents may be given shape. The answer to this question can benefit from our experiences with autonomous agents which already exist for quite some time and which do function in society on a large scale: organizations. Let us take a company with limited liability as an example, and let us call this company for the sake of exposition C.

C acts in society by means of its peripherals, which are typically called "agents" or "representatives". Some of these agents are humans, but others are also autonomous agents, such as the internet-based program that sells the products of C to online customers (including some autonomous agents). The events in which these peripherals are sometimes attributed to C as its acts, and C is therefore considered to be an agent. Some of the acts performed by C are generally praised, and the rumor goes that C has even been nominated for a Nobel prize. However, some other acts of C have been classified as causes of damage, and C is held liable to pay damages. This means that C has the obligation to compensate the damage and that it defaults on its obligation if it refrains from doing so. In case such a default would take place, the assets of C are collateral for the obligation and are amenable to attachment. In case C would commit crimes the company may be fined, and if the crimes are serious enough, its existence may be terminated.

As this example illustrates, law has already an elaborate set of means to deal with the responsibility of some autonomous agents for their doings. With some adaptations these means can also be made suitable to deal with the responsibility of computer-based and physically autonomous systems. The 'only' thing that is needed is to treat modern autonomous agents as legal persons just as we already do with the more old-fashioned autonomous agents.

8. Conclusion

The main step that needs to be taken to answer the question whether autonomous agent should be responsible for their doings is to stop thinking from an anthropocentric perspective. Intelligence, beliefs, desires, acts and agency, even when attributed to human beings, are no 'real' things which exist in a mind-independent realty, but they are the result of attribution, just as damage and causation. The question is not whether attribution is true or truthful, but whether it is desirable or useful. We have seen that the attribution of agency and responsibility are not useful for the reason that the agents deserve to be held responsible for what they did. Acts are either determined or random, and in neither case does it make sense to attribute responsibility for them to agents. However, it can make sense to attribute responsibility in order to influence the future for the best. This holds for human agents, and

also for autonomous agents, for both to the extent that their behavior can be influenced through the attribution of responsibility. The legal implementation of this responsibility takes place by treating agents, whether they be humans, organizations, or other autonomous agents as legal persons.

There is only one difficulty: it is very hard for humans to abandon their anthropocentric world view. It is still common to distinguish between humans and animals, and most likely it will stay common for quite some time that humans will treat the differences between them and non-living systems[45] are qualitative rather than quantitative. And yet, stopping to do so is the first step towards an adequate treatment of modern autonomous agents.

[45] The distinction between living and non-living systems may turn out to be as arbitrary as the distinction between agents and non-agents.