Artif Intell Law (2017) 25:255–271
DOI 10.1007/s10506-017-9208-7

Theoretical foundations for the responsibility of autonomous agents

Jaap Hage[1]

Published online: 31 August 2017

Abstract This article argues that it is possible to hold autonomous agents themselves, and not only their makers, users or owners, responsible for the acts of these agents. In this connection autonomous systems are computer programs that interact with the outside world without human interference. They include such systems as 'intelligent' weapons and self-driving cars. The argument is based on an analogy between human beings and autonomous agents and its main element asserts that if humans can be held responsible, so can, in principle, autonomous agents, as humans are more like autonomous agents than is often assumed (rather than the other way round).

Keywords Attribution · Autonomous agents · Capacity · Liability · Responsibility

1 Introduction

This article has two purposes. The first is to argue that it is possible and perhaps more sensible, to hold autonomous agents, and not merely their developers, owners or users, responsible in the sense of being 'legally liable' for their doings. The second—more important—purpose is to outline an account of autonomous agents as bearers of responsibility. In this connection 'autonomous systems' are taken in the broad sense to refer to computer programs that interact with the outside world without human interference. They range from programs that trade on the stock market, to 'intelligent' thermostats and weapons and self-driving cars.

✉ Jaap Hage
jaap.hage@maastrichtuniversity.nl

[1] University of Maastricht, Maastricht, The Netherlands

The argument presented in the subsequent sections has the following outline. After a brief section in which two kinds of responsibility are distinguished; it is argued, in Sect. 3, that agency, responsibility and liability are not found in a mind-independent reality, but rather are attributed to elements of a social practice that will be called the 'practice of agency'. This practice may be based on the way human beings experience themselves and their fellow humans, but does not necessarily have a firm foundation in the 'real', mind-independent world. This practice might have been different from what it actually is and might attribute agency, responsibility and liability to autonomous systems just as easily as it actually attributes these characteristics to human beings. It is argued that a major reason to treat humans and autonomous systems differently in this respect—that humans act intentionally and on the basis of a free will—has lost much of its credibility in the light of modern science.

Assuming that different treatment of humans and autonomous systems is not obvious, the question is whether different treatment would be desirable. That question is addressed in the Sects. 4 and 5. One reason to treat humans and autonomous systems differently would be that humans, in contrast to autonomous systems, sometimes deserve to be liable. Section 4 contains the core of the argument and is devoted to the refutation of the view that humans and machines should be treated differently. In particular it includes the refutation of compatibilism in it its most popular guise—the capacities approach—which is the main justification for retributive theories concerning the attribution of liability.

If humans should not be treated differently than autonomous systems because humans sometimes deserve to be liable, the question becomes for what reason humans should be held liable and whether this reason also applies to autonomous systems. Section 5 gives short answers to these two questions: Humans should be held liable if this has good consequences, and the same holds for autonomous systems. Whether we should ascribe acts to autonomous systems and hold these systems liable for their acts therefore depends on the desirability of the consequences of doing so, exactly as it does for human beings. The section hints at some considerations that might be relevant in this connection.

This paper is concluded in Sect. 6.

2 The meanings of 'responsibility'

Discussions of responsibility of humans, but also of autonomous agents, are burdened by the ambiguity of the very notion 'responsibility'. Any argument about who is to be held responsible, and for what, must be clear about the meaning of 'responsibility'.

A well-known set of distinctions concerning responsibility derives from the work of Hart (2008, pp. 210–237). Hart distinguished between role-responsibility, causal responsibility, liability-responsibility and capacity-responsibility. As the name indicates, role-responsibility is connected to some social role. For example, the dean of a faculty is role-responsible for the well-being of the faculty. Strictly speaking, role-responsibility is not a special kind of responsibility, but rather a frequent ground for the existence of responsibility in a still unspecified sense: people are responsible because they fulfil a particular role.

Causal responsibility exists when a result is attributed to some event or agent, which is held to be the cause of the result. For example, an earthquake is causally responsible for the fires that resulted from it, and Claude is causally responsible for the car accident because of his lack of attention. Here 'being responsible' does not mean anything other than 'being the cause'.[1] A special case of causal responsibility is that an act is ascribed to an agent, who is then held to be causally responsible for the performance of this act. This variant on causal-responsibility exists in particular for acts that are defined in terms of the consequences they brought about. Examples of such acts are closing the door, defined in terms of the result that the door is closed, or homicide, defined in terms of the result that a person is dead. Only when this result is brought about by the agent, can the act be ascribed to the agent as 'his' act. Later in this article we will use the term 'responsibility' *tout court* for this variant of causal responsibility where the cause is an agent and the result of this cause defines an act.

To be liability-responsible means that one has the bear the consequences. In private law these consequences typically are seen as liability for damages. In criminal law it typically means liability to be punished. Often the single word 'liability' is used for this variant of responsibility. Whether a person is liable can be judged from different points of view, the legal and moral points of view being most prominent. An agent can legally and/or morally be liable for the damage that he caused. However, liability from one particular point of view or another is not a different kind of liability. The point of view merely influences the standards for the existence of liability and the nature of the consequences which the liable person must bear.

Capacity-responsibility, finally, concerns the possession of certain capacities without which an agent cannot be held liable for his doings. For example, a person who is mentally retarded may not be capacity-responsible for placing the burning candle in a place where it may cause a fire. Just like role-responsibility, capacity-responsibility may not really be a separate kind of responsibility, but rather a ground for (the absence of) liability. A person who lacks the relevant capacities may not be liable for what he did, and the absence of liability-responsibility is attributed to a lack of capacity-responsibility.

For our present purposes the discussion whether there are really many different kinds of responsibility is not very relevant. It is however important to distinguish between liability and being responsible for an act in the sense of being its causal origin. We will use the word 'liable' for the first kind of responsibility, and the word 'responsible' for the second kind.

[1] The notion of 'cause' suffers from its own ambiguities, which may manifest themselves as ambiguities of 'causal responsibility'.

3 The attribution of agency and responsibility

3.1 The mental and the physical aspects of acts

It is almost a tautology that human beings are responsible for their acts, because if people are not responsible, the acts would not count as 'their' acts. However, if humans are 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. One difference might be that 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 autonomous agents lack intention and a will. In spite of this lack, they have in common with human beings that their 'behavior' has a physical aspect and that this physical aspect is part of the processes that constitute physical reality. Even if human behavior is intentional and based on a free will, what physically happens fits in the same chain of events in which the 'acts' of autonomous agents fit. The question is therefore whether intention and free will make a difference, and whether they play a role in the chains that constitute the physical aspects of acts. If they do not play such a role, they are in that sense redundant. It seems dubious then to base a difference in the attribution of responsibility on such redundant phenomena.

There are many reasons to assume that intention and free will do *not* play a role in the chains of facts and events that constitute the physical aspects of acts. The physical research paradigm which assumes that physical events are only linked to other physical events in a law-like fashion, works quite well and leaves no room for intervening mental phenomena like intention or will.[2] It is completely unclear how physical events might be influenced by mental events as such, and there is no evidence that such an influence exists.[3] It should be emphasized in this connection that the two words 'as such' in the previous sentence are crucially important. The possibility should not be excluded beforehand—and actually it is quite likely—that intentions and will have counterparts in brain states. These brain states play a role in the processes of the physical world. However, this does not mean that intention and will *as mental phenomena* (as 'qualia'; Tye 2016) influence the physical world.

Therefore, even if humans act on the basis of intentions and free will—whatever that might mean—while autonomous agents do not, this does not make a difference for what happens physically. It is for this reason not at all obvious that the alleged difference between humans and autonomous agents should mean that only human agents are held responsible for their acts and autonomous agents not. Whether such a difference should be considered depends on the reasons why we hold humans responsible, even if the mental aspects of their behavior do not influence the physical aspects thereof.

[2] In the illuminating catalogue of arguments against free will that was drawn up by Nahmias (2014), this research program is dealt with under the name 'naturalism'.

[3] There is an enormous amount of literature on these issues. Useful overviews are Kane (2002, 2005).

3.2 Experience of agency

Typical human agency is intentional: the acting person experiences himself as acting on purpose 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 an act.[4] The sentence 'Jaap saw Esther hit Veerle' is only true if Esther, Veerle and Jaap exist, if Esther really did hit Veerle and if Jaap truly saw this happening. The sentence 'Jaap had an experience of Esther hitting Veerle' can, on the contrary, also be true if Esther, Veerle and the hitting event were all figments of Jaap's imagination. Therefore it is only guardedly 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 existing persons.

However, experiences are the foundation of empirical knowledge; not an infallible foundation 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 can explain why humans beings conceptualize many events in terms of acts, agents, cause and effect, intentions and (free) will, and also 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 question whether this conceptualization is valid in the sense that what is experienced also exists in reality cannot be answered solely on the basis of our experience of our own acts. We must also look at the sciences to get a picture of what exists independently of our experience and to confront our experiences with the results of science. At their present stage of development, the sciences seem to leave no room for intentions and manifestations of will which *as such* cause acts. The question therefore becomes how to combine these findings of science with the way humans experience themselves.

3.3 The realist and the attributivist view of agency

A good starting point for combining the human experience of agency with the findings of the sciences is to distinguish between two views of agency. According to the realist view, the intention to act and the will that leads to the performance of an act are taken to be 'real' things, which exist independently of human experience. Our experience of them is in this view very much like perception; an awareness of something that exists independently of the experience. Think for instance of the experience we have when we perceive a house. The house exists independent of our perception, and the statement that John sees the house expresses a relationship (seeing) which exists between John and the house as independent entities. Analogously, if 'intention' and 'free will' are interpreted as referring to *real* intentions, and *real* free will, intentions and will are taken to exist independently of our experience of them. It is not clear whether anybody explicitly adheres to this

[4] This point was already made, be it in quite different phrasing, by Searle (1983, p. 124).

realist view of agency. That is not surprising, because modern science does not provide us with reasons to assume that intention and will as mental states (qualia)[5] 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 attributes 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 can then be based on an analogy with 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.[6] This recognition does not concern some independently existing entity, but is essentially the attribution of an intention to act, 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 for granted that the agent acted intentionally. Further, if there are no reasons to assume that the act was caused by a factor that should not have caused it, and if the act had no perceivable 'illegal' cause, we thus assume that the act was the result of free will. Since it is up to us what we consider illegal causes, it is also up to us to determine which acts we count as based on a free will, or as being voluntary. Free will is in the attributivist view a matter of attribution, very much like intention.

3.4 Expansion of the attributivist view

From the attributivist point of view 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 for instance the case when omissions count as acts. A father is deemed to have acted if he intentionally refrained from feeding his baby child in time. So, in the attributivist view, acts are the result of attribution too, and—since agency presupposes acts—the same holds for agency.

Moreover, this 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 rather is a matter of attributing the status of cause to the act or event (see e.g. Hart and Honoré 1985). In social and physical sciences it seems to be 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.[7]

[5] Again, this leaves the possibility open that intention and will are states of mind that are realized by brain states and that these brain states cause muscular movements. If that is the case, intention and free will are *as mental states* are, redundant for the causation of the muscular movements.

[6] Other support for this view can be found in Wegner (2002).

[7] 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. Cf. Psillos (2009).

This, finally, leads us to responsibility for an act that consists of causing some result. If agency and causation are a matter of attribution, then responsibility must be a matter of attribution too. If one considers responsibility as a mind-independent entity in the 'outside world', this 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. Responsibility is best accounted for with the attributivist view, and is then the result of attribution, rather than a 'real' phenomenon.

3.5 Attribution to autonomous agents

By definition, what is 'real' does not depend on attribution and is mind-independent. What is the result of attribution, on the contrary, depends on the human mind. This mind-dependency may be direct, as when somebody considers what another 'does' as an act. It may also be indirect, as when 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 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 to whom the house used to belong. (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 grounds. 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, gods and organizations responsible and liable for these 'acts'. This however is only from a historical perspective done by analogy to the attribution of agency to human beings. Ontologically speaking, there is no difference between attribution of agency to humans and to other agents.

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. The question is not whether such attributions *can* be made, but whether it is *desirable* to do so. We attribute agency and responsibility to humans, and the reasons we do so may illuminate our thinking about the attribution of agency and responsibility to autonomous agents.

3.6 The desirability of attribution

The attribution of agency and act-responsibility to human beings is in the first instance not done with a particular purpose in mind. It is more likely the result of how we experience ourselves as involved in agency, an experience which is projected onto other humans. Whether we attribute agency and responsibility to autonomous agents depends in the 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 the 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 ignores our intuitive analogizing between human and non-human ‘behaviors’. 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—is revealed on closer inspection only to be similar to an agent.

When is it desirable to attribute agency to some system, be it human, animal or otherwise? To answer a closely related question—the question of when it is desirable to attribute intentional states such as beliefs and desires to a system—Dennett (1981, 1987) introduced the notion of the “intentional stance”. This intentional stance should be adopted when the behavior of the system is best explained and/or 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 expectation 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 somewhere, and its desire to have the food (De Waal 2016). 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 beat the opponent.

The step from chess-playing programs to ‘intelligent’ physical systems such as cruise missiles and self-driving cars is only minor. The behavior of ‘intelligent’ physical systems is most fruitfully explained by the assumption that these systems reach for some determined goal (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. The cruise missile adapts its course to avoid defense-missiles, or to correct for heavy winds; the self-driving car brakes to avoid a collision, or takes a detour because it received info that a road is blocked because of road-work. Adapting course and braking 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 in fact desirable to attribute agency and responsibility to some types of autonomous agents.

4 Retributive attribution of liability

The attribution of agency can be contemplated from the perspective of explanatory and predictive power, but also 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 responsible for some event, this is grounds to call this event an ‘act’ and to hold the agent liable for what he did. Moreover, this relation between responsibility and liability 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. One answer is that somebody should be liable for damage caused to another because he deserves to be liable. This is the retributive view on liability, and we will consider it in some detail in the present section. The other answer is that liability serves some purpose, such as prevention of future damage, the achievement of distributive justice, or popular satisfaction. This is the consequentialist view, which we will consider in Sect. 5.

4.1 Justification within a practice and justification of a practice

It is not always permitted to damage somebody else's interests, and if one nevertheless does so, this is often seen as a reason why the tort-feasor who caused the damage, must compensate it. This is part of the practice of tort law. This same practice also includes the attribution of intentions and culpability to agents, and the recognition of causal relations between acts and their consequences. Working from this practice it is possible to attribute particular damage to a particular act and a particular agent, and to also attribute liability 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, the justification of liability on the basis of tort law only makes sense if the practice of tort law itself makes sense.[8] It is tempting to justify this practice by trying to understand it 'from within'. The justification then consists of identifying the elements of the practice and justifying them on the basis of the principles underlying the practice. This approach to the attribution of liability can amongst others be found in the work of Dworkin (2011, pp. 227–231), who claims that the practice of law must be understood from within. Understanding is a psychological category, which has no other standard 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 the 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 (the 'naturalistic' fallacy). The fact alone that some social practice such as tort law exists, does not lead naturally to the conclusion that the practice should continue to exist.

[8] Cf. Rawls' (1955) distinction between justification within a practice and justification of a practice (criminal law in his case).

The above argument only rebuts one type of argument of why liability is based on desert: the argument from understanding what the practice actually is; it does not attack that practice itself. It is now appropriate to address the general conclusion: should people be held liable for damage for the reason that they deserve to be liable? That depends on the reasons why we assume that people sometimes deserve to be held liable. Which reasons are actually recognized is a matter for empirical investigation, but here we will investigate one such a reason, perhaps the most important one. A person may deserve to be liable because through his actions, he caused damage to another and should not have done what he did because of the damage it would cause. This makes sense on the presumption that people can avoid wrongful acts for the reason that they recognize their wrongfulness. In the vocabulary that is currently fashionable, one might say that desert depends on reason-responsiveness, the capacity to act on reasons (Morse 2007; McKenna and Coates 2015). Are people reason-responsive in the sense required for the assumption of desert?

4.2 Capacities and possibilities

The question concerning what a person can or cannot 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 as acts, agency, causation, culpability, responsibility and liability. The argument for this claim will be presented in Sect. 4.4, but if we assume that the claim is correct, the observation that somebody is reason-responsive does not help us much further in our search for the foundation of liability. The social practice of tort law cannot be justified by the Nth element of this very practice. In order to avoid this complication we will replace the question of whether a person could have avoided (had the capacity to avoid) his damage-causing behavior by the question of 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 on the basis of the practice of tort law or agency, which includes the attribution of capacities, while the latter question has no direct ties to agency and therefore does not depend on this practice.

Let us therefore 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 most interesting in the cases where they are 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 cases where they were not realized. To deal with this problem, a thinking device was constructed: possible worlds theory.[9] The basic idea underlying possible worlds theory is that something is necessary when it is the case whatever else may also be the case, that is: in all possible worlds. Something is possible if it is the case in at least one possible world.

[9] 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 which 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 Chellas (1980).

The actual world consists of all the facts that happen to obtain, while a different possible world contains a set of all facts that might have obtained under different circumstances. In the actual 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.

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, even though Jane did not visit her mother, 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 an 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 so it seems at first sight.

4.3 Possible worlds and constraints

We now have a definition of what it means that a person has a certain capacity. 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 (Hage 2015a). 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. It is a logical constraint as 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 both heated and red, but it is physically 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 a 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. Only relative to such a set, it makes 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? It is apparently not the case, and it is worthwhile to consider more closely why not. Both with logical and 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.[10] 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.

4.4 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 we 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 the abstract, but focus merely on the characteristics of individual agents.

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 an agent has.[11] But which personal characteristics 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.

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. In a deterministic world view, only one possible course of the world is possible if all previous facts and all laws are taken into account. In 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, in 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 do nothing else other what he actually did, and in the non-deterministic case not because it is arbitrary what the agent did and therefore not dependent on the agent himself. 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 of what he will do is *ex hypothesi* included in the set of all constraints. Given these constraints the agent either has no real choice what he will do, or his act will be arbitrary. In neither of these cases does the act depend on the agent himself and in neither of them is there a ground for attributing responsibility.

[10] In a deontic setting, this idea is explored in dyadic deontic logic. See Von Wright (1971).

[11] A related account of what an agent can, or cannot, do, is found in Honoré (1999).

The distinction between what an agent did and what he had the capacity to do only makes sense if not all facts are taken into account as constraints on what is possible. For instance, in determining whether Claude had the capacity to avoid the car accident, we take into account that in general Claude is capable to drive a car, but we do not take into account that on this occasion he was distracted by his quarrelling children in the back seat. Therefore we conclude that Claude 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, an issue of which facts *should* be left out of consideration in determining 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 Sect. 4.2) 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/attributive 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 on the capacities of the agent, because these capacities are themselves part of the judgment.

To summarize the above argument: we are left with two options. 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 does the agent deserve to be held responsible and liable for damages. The other option is to count some constraints, and to discount others, 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, because it would be part of that very practice.

5 Purposive attribution of liability

Desert is not the only possible foundation for the attribution of agency and liability. The alternative is that agency and liability are attributed for some purpose. Attribution based on desert is backward looking and presupposes 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 therefore will not make sense to attribute agency and responsibility because the agent deserved it. If determinism is incorrect, the past was a random affair, and it further makes no 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 attribution of agency and liability influenced the future.

The practice of attributing responsibility and liability may be justified if this attribution brings desirable consequences. This raises two questions, one evaluative and the other empirical. The evaluative question asks which consequences are desirable. In this article the evaluative question will be avoided, by postulating three purposes which are deemed to be desirable:[12]

- influencing future behavior of the agent to whom responsibility and liability has been 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 answer to the question of whether the attribution of responsibility to human agents contributes to the achievement of these purposes can only be found through empirical research, and the following attempt at an answer is speculative. It seems likely that if agents are held responsible and liable for their acts, and if they know 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 and liable 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 do. Perhaps this belief will disappear in the (far) future, but as long as it is still prevalent, it will underscore the practice of holding people 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 liable, the third purposive reason for holding them liable may not apply to them either. Since most people do not believe that autonomous agents are 'real' agents, and therefore that they do not deserve to be liable, there is little gained by attributing them liability.[13]

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 in the case of human agents. Weaker, but not completely without force. One reason that the argument still has some force is that humans

[12] The author adheres to some form of utilitarianism (Hage 2015b), and believes that the three mentioned purposes are instrumental to the maximization of happiness.

[13] Joanna Bryson has pointed out that (some) people do believe that autonomous agents who earn money, for instance, by trading in stock, deserve to be taxed. Popular satisfaction would then be increased by making these systems liable to taxation.

might believe that if even autonomous agents are held responsible for their 'doings', then certainly human 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' influences 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 there 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 more or less sophisticated, and their reaction to responsibility depends on their sophistication. If a program is relatively simple, it does not take responsibility into account, and attributing responsibility to it will have no effect on the program's behavior. Then a practice of attributing responsibility makes little sense, at least not for these specific programs. 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, 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.

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 of being having been 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 the 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 program runs on a machine which typically has peripherals for input and output. The only thing that makes physically autonomous systems different from a mere computer program is the nature of the peripherals. These different peripherals can make the nature and impact of what the physically autonomous system does radically differ from what a mere computer program does. However, for the fundamental issue of whether it makes sense to hold them responsible this does not seem to make any difference.

6 Conclusion

There seem to be no fundamental reasons why autonomous systems should not be held liable for what they do. However, this 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 perspective. Whether this is the case is an empirical question, and the answer may vary from system to system and may change over 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 responsibility and liability are concerned.

Acknowledgements The author thanks Antonia Waltermann and the anonymous reviewers for AI and Law for valuable comments on an earlier version and Rebecca Kumi for copy-editing.

References

Chellas BF (1980) Modal logic; an introduction. Cambridge University Press, Cambridge

de Waal FBM (2016) Are we smart enough to know how smart animals are? Norton & Company, New York

Dennett D (1981) Intentional systems. In: Brainstorms. The Harvester Press, Brighton, pp 3–22

Dennett D (1987) True believers. In: The intentional stance. The MIT Press, Cambridge, pp 13–35

Dworkin R (2011) Justice for hedgehogs. Harvard University Press, Cambridge

Hage J (2015a) The (onto)logical structure of law. In: Araszkiewicz M, Pleszka K (eds) Logic in the theory and practice of law making. Springer, Cham, pp 3–48

Hage J (2015b) Recht en Utilisme Een Pleidooi voor Utilisme als Richtlijn voor de Wetgever. Law Method. doi:10.5553/REM/.000011

Hart HLA (2008) Punishment and responsibility, 2nd edn. University Press, Oxford

Hart HLA, Honoré T (1985) Causation in the law, 2nd edn. Oxford University Press, Oxford

Honoré T (1999) Appendix: can and can't'. In: Responsibility and fault. Hart Publishing, Oxford, pp 143–160

Kane R (ed) (2002) The Oxford handbook of free will. Oxford University Press, Oxford

Kane R (2005) A contemporary introduction to free will. Oxford University Press, Oxford

McKenna M, Coates DJ (2015) Compatibilism. In: Zlatan EN (ed) The Stanford encyclopedia of philosophy. http://plato.stanford.edu/archives/sum2015/entries/compatibilism/. Accessed 14 Aug 2017

Morse SJ (2007) Criminal responsibility and the disappearing person. Cardozo Law Rev 28:2545–2575

Nahmias E (2014) Is free will an illusion. In: Sinnott-Armstrong W (ed) Moral psychology volume 4: free will and moral responsibility. MIT Press, Cambridge, pp 1–15

Psillos S (2009) Regularity theories. In: Beebee H, Hitchcock C, Menzies P (eds) The Oxford handbook of causation. Oxford University Press, Oxford, pp 131–157

Rawls J (1955) Two Concepts of Rules. Philos Rev 64:3–32

Searle JR (1983) Intentionality: an essay in the philosophy of mind. Cambridge University Press, Cambridge

Tye M (2016) Qualia. In: Zalta EN (ed) The Stanford encyclopedia of philosophy (Winter 2016 Edition). https://plato.stanford.edu/archives/win2016/entries/qualia/

von Wright GH (1971) A new system of deontic logic. In: Hilpinen Risto (ed) Deontic logic: introductory and systematic readings. Reidel, Dordrecht, pp 105–120

Wegner DM (2002) The illusion of conscious will. The MIT Press, Cambridge