Artif Intell Law (2017) 25:293–304
DOI 10.1007/s10506-017-9207-8

On the legal responsibility of autonomous machines

Bartosz Brożek[1,2] · Marek Jakubiec[1]

Published online: 31 August 2017

Abstract The paper concerns the problem of the legal responsibility of autonomous machines. In our opinion it boils down to the question of whether such machines can be seen as real agents through the prism of folk-psychology. We argue that autonomous machines cannot be granted the status of legal agents. Although this is quite possible from purely technical point of view, since the law is a conventional tool of regulating social interactions and as such can accommodate various legislative constructs, including legal responsibility of autonomous artificial agents, we believe that it would remain a mere 'law in books', never materializing as 'law in action'. It is not impossible to imagine that the evolution of our conceptual apparatus will reach a stage, when autonomous robots become full-blooded moral and legal agents. However, today at least, we seem to be far from this point.

Keywords Autonomous agents · Responsibility · Folk-psychology · Law in action

1 Introduction

Our goal in this paper is to consider the question whether autonomous machines can be legal agents, i.e. whether they can be legally responsible for their actions. By an autonomous machine we understand "a system situated within and a part of an environment that senses that environment and acts on it, over time, in pursuit of its own agenda and so as to effect what it senses in the future" (Franklin and Graesser

✉ Bartosz Brożek
bartosz.brozek@uj.edu.pl

[1] Department of Philosophy of Law and Legal Ethics, Jagiellonian University, Bracka 12, Kraków, Poland

[2] Copernicus Center for Interdisciplinary Studies, Kraków, Poland

1997). According to this definition, an autonomous machine is reactive (it responds in a timely fashion to changes in the environment), self-controlling (i.e. it exercises control over its own actions and is not directly controlled by any other agent), goal-oriented (it does not simply act in response to the environment), and temporally continuous (it is a continuously running process) (Franklin and Graesser 1997). Of course, autonomous machines may have additional properties such as the ability to communicate with other agents or change their behaviour in line with previous experience; they may even be equipped with "believable personality", i.e. act in such a way which gives an impression of the possession of individual "character traits" (Franklin and Graesser 1997). The definition provided above is therefore to serve as a kind of conceptual threshold: the conclusions we reach shall be applicable to any autonomous machine, i.e. a reactive, self-controlling, goal-oriented and temporally continuous artificial agent, whether it has other (stronger) properties or not.

We begin, in Sect. 2, by considering two extreme views in relation to this problem, which we deem 'restrictivism' and 'permissivism'. The former denies the possibility of holding autonomous machines legally responsible on purely metaphysical grounds; the latter imposes no restrictions on the possible legal constructions, and hence does not prohibit the introduction of the responsibility of artificial autonomous agents. We argue that both these stances are mistaken. Next, in Sect. 3, we try to show that in order to properly consider our question one needs to analyse the relationship between three different conceptualisations of human behaviour: folk-psychological, scientific (provided by psychology, neurobiology, etc.), and legal. We claim that the law—in order to be an efficient tool for regulating social interactions—must be largely based on the conceptual scheme of folk-psychology. Finally, in Sect. 4, we argue for two claims: that the problem of legal responsibility of autonomous machines is completely different from the question of the responsibility of legal persons such as corporations, foundations, or states; and that the conceptual apparatus of folk-psychology makes it impossible to treat machines as legal agents.

2 Between two extremes

There are two extreme views regarding the legal responsibility of autonomous artificial agents. The first one may be deemed restrictive and boils down to the claim that machines—no matter how "intelligent" or "autonomous"—can never become legal persons and be legally responsible for their actions. The usual argument adduced to substantiate this claim highlights certain properties of the human being—such as intentionality, free will, autonomy or consciousness—which seem to constitute the prerequisites of legal (and moral) responsibility (e.g. Fischer and Ravizza 2000), but which are apparently lacking in any machine. On the other end of the spectrum is permissivism, the claim that the law is a flexible tool of social engineering, which may be used to make anyone—or anything—a legal person (i.e. a legal agent and/or patient). From this point of view, there is no real problem in holding an autonomous artificial agent responsible for their actions; the same holds for any other object, be it a stone, a river or a comet. History of law shows clearly that it is possible: legal systems often excluded some human beings from the pool of legal agents or patients (slaves, children, etc.) or included in the pool the entities which lacked the characteristic features of "full-blooded" actors (e.g., Whanganui river in New Zeland) (Hutchison 2014).

Of course, neither of the two extremes is easy to accept. The restrictive view is firmly based on very strong metaphysical claims, and there are reasons to deem them doubtful. A particularly clear and persuasive rejection of restrictivism can be found in Hage 2017. Hage observes that we usually think of legal responsibility as requiring two things: intentionality and free will. This is the bedrock of our everyday experience of legal (and moral) action. However—argues Hage—this is a "realist mistake". We have a tendency to treat intention to act and the will that leads to the performance of an act as "real things", "out there" in the world, independent of the human mind. Meanwhile, Hage claims that there are no grounds for embracing the realist standpoint. Rather, intention and free will should be understood as things we attribute to one another. The complex network of social interactions is based on such attributions: irrespective of the real nature of human action, we grasp it by deploying a conceptual apparatus which utilises the concepts of intention, free will, autonomy, etc. This shows, Hage argues, that the restrictive conception of the responsibility of artificial agents rests on a grave mistake. If intention and free will are not "real" phenomena, but are mere outcomes of our attributions, they cannot constitute necessary conditions of legal (or moral) responsibility. It follows that there is nothing barring the ascription of responsibility to machines (Hage 2017).

To put it in a different way: the fatal flaw of restrictivism is the fact that it is based on very strong and apparently mistaken ontological assumptions. Attributivism is much more plausible: it not only embraces weaker ontological commitments, but also seems to be more coherent with the actual social practice. People have the default tendency to understand the actions of others in terms of intentions and free will, and to do so they apply no special tests determining whether an agent X has "real" intentions or acts out of free will (see O'Connor 2010, Nahmias et al. 2005). Rather, these things are assumed and such assumptions are upheld as long as there is no evidence to the contrary (e.g., that someone has concealed their intentions, was coerced, etc.). But does it mean that by rejecting the restrictive view we are forced to embrace the permissive one? Is the decision to make artificial agents legally responsible for their actions purely conventional? Is the concept of legal responsibility so flexible that anything may be regarded as a legal agent?

A natural answer to this question is that while there are no conceptual limits to ascribing legal agency, there must be special reasons to do so in relation to anyone or anything which is not a human being. The concept of legal responsibility is flexible, but it cannot be used indiscriminately. This is the line of reasoning employed by Hage when he argues that the attribution of legal responsibility to artificial agents would be justified only if its consequences (such as influencing future behaviour of the agent to whom responsibility is attributed or influencing future behaviour of other agents) were desirable (Hage 2017).

In a similar spirit (Bryson et al. 2017) argue that since legal personhood is a fiction, "the inherent characteristics of a thing are not determinative of whether the [legal] system treats it as a legal person (Bryson et al. 2017)". It follows that—conceptually—artificial agents can be treated as legal persons (agents and/or patients). However, granting machines a legal status may lead to abuse. The danger is not merely imagined, given our experience with legal persons such as corporations or international organisations. "Trying to hold an electronic person to account, claimants would experience all the problems that have arisen in the past with novel legal persons. There almost inevitably would arise asymmetries in particular legal systems, situations like that of the investor under investment treaties who can hold a respondent party to account but under the same treaties is not itself accountable (Bryson et al. 2017)".

Let us repeat that the constraints on ascribing legal agency to machines as suggested by Hage and Bryson et al. are not of fundamental nature. Neither Hage nor Bryson et al. seem to believe that there are conceptual barriers to extend the pool of legal agents so as to include autonomous artificial agents. Rather, they claim that it may turn out to be a bad decision in terms of its consequences, e.g. when it will not lead to the socially desired outcomes or even to legal abuse. In other words, they temper permissivism (the idea that legal personhood may be granted to anything, including machines) only by employing utilitarian criteria, which are external to the problem of whether machines can in principle be legally responsible for their actions.We believe that the situation is more complex.

3 Three conceptual schemes

The distinction between law in books and law in action was introduced by Pound (1910), and popularised in the writings of American legal realists. Painting with a broad brush, law in books is the collection of legal acts, court rulings and other law-related documents, while law in action is the set of norms of conduct actually enforced in the court of law. Horwitz (1992) has called this distinction the "realists's battle cry", and although it falls short of serving as the realists' definition of law, it nevertheless underpins the importance of the practical dimension of legal phenomena. No sophisticated theoretical construction embodied in a legal act can count as law as long as it is not observed and enforced in legal practice. This claim is also not completely alien to legal positivists: even Hans Kelsen, who considered the legal system as an ideal belonging to the sphere of pure "ought", understood that there is no binding law if the legal system is not—by and large—efficacious (Kelsen 1945).

Applied to our problem, the distinction between law in books and law in action yields the following question: it seems possible—according to the position advocated by permissivists—to grant legal status to any object, including intelligent machines. The legislator is free to introduce any legal construction that may serve their goals. However, there is a danger that the construction in question would remain "in books" only, having no real significance for the legal practice. Arguably, there are several conditions for a successful implementation of "law in books" in legal practise: some of them pertain to the realisation of the goals of the regulation (there exists a need to realise the given goal, the legal means of realising it are properly selected), others to the form of the regulation (e.g. Summers 2006). In particular, any law must be understandable to its addressees. It is difficult to imagine "law in action" based on highly complicated provisions, which are extremely difficult to comprehend (a case in point is tax law, which often is difficult to follow and requires the assistance of specialised tax advisors).

In the context of the question pertaining to the ascription of responsibility to autonomous artificial agents, it is necessary to consider the relationship between three different conceptual schemes, which are all related to human behaviour: folk-psychological, scientific, and legal. Let us begin by considering the first two. Folk psychology is usually understood as the ability of mind-reading, i.e. of ascribing mental states to other people (Stich 1983). A more detailed characterisation—albeit not an incontestable one—has it that folk psychology is a set of the fundamental capacities which enable us to *describe* our behaviour and the behaviour of others, to *explain* the behaviour of others, to *predict* and *anticipate* their behaviour, and to produce *generalisations* pertaining to human behaviour (Stich and Ravenscroft 1992). Those abilities manifest themselves in what may be called *the phenomenological level* of folk psychology as "a rich conceptual repertoire which [normal human adults] deploy to explain, predict and describe the actions of one another and, perhaps, members of closely related species also. (…) The conceptual repertoire constituting folk psychology includes, predominantly, the concepts of belief and desire and their kin—intention, hope, fear, and the rest—the so-called propositional attitudes" (Davies and Stone 1995). It is the conceptual apparatus that we deploy every day to understand and predict the behaviour of other people. Interestingly, folk psychology is to a large extent culture-dependent: people belonging to various cultures conceptualise human action in essentially different ways. For example, according to a much discussed study by Michael Morris and Kaiping Peng, Asians tend to explain behaviour by citing situational factors, whereas Westerners explain it by focusing on personal causes such as beliefs and desires (Morris and Peng 1994). The study in question concerned, *inter alia*, the explanation of a mass murder. In *The World Journal* the Chinese reporters focused on the situational causes influencing the behaviour of the mass murderer, and mentioned that the "gunman had been recently fired," "the post office supervisor was his enemy," and that he "followed the example of a recent mass slaying in Texas". On the other hand, in *The New York Times* the American reporters focused on the dispositions of the mass murderer, noticing that he "repeatedly threatened violence," "had a short fuse," "was a martial arts enthusiast," and was "mentally unstable".

The scientific explanation of human behaviour, and the conceptual scheme it utilises, is quite different from folk psychology. Let us have a look at an example pertaining to moral judgment and action. One of the more famous recent attempts to explain the mechanisms behind the human ability to reason morally is Jonathan Haidt's social intuitionist model. In his essay "The Emotional Dog And Its Rational Tail" (2001), Haidt argues that one should reject the rationalistic paradigm in moral psychology, i.e. an approach "that moral knowledge and moral judgement are reached primarily by a process of reasoning and reflection" (Haidt 2001). Instead, he claims that moral judgement always appears in consciousness in an automatic, effortless way, as a result of the workings of moral intuition. Moral reasoning, on the other hand, is a process tied up with effort and takes place usually after the decision has been reached, supporting it only *ex post* (Haidt 2001). It should also be added that on Haidt's view moral intuition is formed in the course of social interactions, and its development is tied up with a number of processes. In this context, a special role is played by the unconscious mechanism of cultural transmission. Numerous experiments and observations suggest that only a relatively small part of our cultural knowledge is learned consciously—children acquire it mainly through the imitation of the behaviour of older children and adults (e.g. Meltzoff and Decety 2003; Want and Harris 2002). Haidt points out that the recent findings regarding the neural foundations of intuition underscore the importance of practice and repetition for the proper training of cultural intuitions. He claims further that the process of social learning depends on the activity of the basal ganglia's circuits, which are also instrumental to motor learning: because of that many social skills are rapid and automatic, just like well-learned motor sequences are. Social skills and judgement processes, learned in a gradual and implicit way, are experienced in the consciousness as arising from nowhere. "The implication of these findings for moral psychology is that moral intuitions are developed and shaped as children behave, imitate, and otherwise take part in the practices and custom complexes of their culture. (…) Even though people in all cultures have more or less the same bodies, they have different embodiments, and therefore end up with different minds" (Haidt 2001).

Let us now come back to the already cited example of the explanation of a mass murder in Asian and Western cultures. While in both cases the crime was judged morally outrageous, the Asian folk-psychological way of seeing things was through a number of situational factors influencing the actions of the mass-murderer; the reporters of *The New York Times*, on the other hand, while describing a similar tragedy, concentrated on the subjective aspects of the situation, such as mental instability of the perpetrator or his enthusiasm for martial arts. In other words, the moral condemnation of the incident is accompanied in both cultures by different explanations of the murderer's behaviour. While Asians tend to see human action as largely a product of circumstances, the Americans take a more individualistic approach to explaining it. From the point of view of Haidt's theory, the condemnation of mass murder can be explained in the following way. First, the act brings about an intuitive moral judgement. This intuition is not inborn—it is trained in the process of enculturation. Second, the intuition may be followed by an *ex post*, rationalised justification. Crucially, Haidt's theory does not determine which aspects of the perpetrator's behaviour should be taken into account in moral judgement: it puts emphasis neither on external circumstances (as the Asian folk psychology does), nor on individual agency (as is the case in the American culture). At the same time, Haidt's model can easily account for such cultural differences. Let us recall that moral judgement is not something one is born with; rather, a number of emotional mechanisms and cognitive skills are shaped by a particular culture to generate both moral intuition and the public criteria for moral justification. In other words, what the Asians and the Americans have in common are some fundamental emotional and cognitive capacities, which are "filled in" with the moral content characteristic of a given culture.

The mechanism described by Haidt captures a part of what may be called *the architectural* level of folk psychology. It explains how people—through their upbringing in a given culture—acquire the ability to apply a conceptual apparatus for explaining, predicting and describing the moral actions of themselves and other people, i.e. how the phenomenological level of folk psychology emerges. From this perspective, the scientific explanation of the human moral behaviour—as proposed by Haidt—is "deeper" than the folk-psychological: the former provides an account of the latter. However, the "deep" architectural level is not directly accessible to a person who makes a moral decision or attempts to explain or describe the behaviour of other people. In everyday situations, when we try to understand the actions of others or held them responsible for what they do, we have no other option but rely on the conceptual repertoire of folk psychology, which includes such concepts as intention, goal, belief and free will. Thus, even though our judgments and actions may in most instances be the outcomes of unconscious processes, we conceptualise them as free and intentional.

Finally, let us consider the legal conceptual scheme. It is easy to notice that it is largely based on the folk-psychological concepts. For example, in relation to crimes we speak of *mens rea* (guilty mind), intention, committing a crime knowingly, maliciously or willingly, with specific intent, etc. (Greene and Cohen 2004). It is difficult—if not impossible—to imagine social life without these concepts. Jerry Fodor goes as far as saying that the elimination of the conceptual repertoire of folk psychology would be "the greatest intellectual catastrophe in the history of our species" (Fodor 1987, xii). In order to get a better understanding of what the rejection of the folk-psychological concepts for explaining action would look like, let us consider the following thought experiment. In a country X, the parliamentary election is won by SPEM, the Solemn Party of Eliminative Materialism. The main goal of SPEM—deeply rooted in the philosophical doctrines of Patricia and Paul Churchland (Churchland 1981)—is to get rid of all the cultural constructions which assume the existence of immaterial reality. In other words, to use a term coined by Daniel Dennett in his *Philosophical Lexicon*, the program of SPEM is to change the country X into *Churchland*, i.e. "a theocracy whose official religion is eliminative materialism" (Dennett 2008). Having finally won the majority in the parliament, SPEM replaces all the legal provisions which utilise folk-psychological concepts pertaining to "immaterial phenomena", such as intention or will, with the description of the relevant physical (materialistic) processes. Thus, for example, the rule which says that "A person commits a criminal offence if he or she acts with intention, recklessness, or negligence" is replaced with "A person commits a criminal offence if he or she has an elevated blood pressure in prefrontal cortex, motor cortex, basal ganglia and cerebellum", while the rule "A person who acts in self-defence is not criminally liable" becomes "A person whose oxytocin level is elevated is not criminally liable". It is not difficult to observe that such radical changes to the conceptual foundations of the legal system would result in complete chaos. The new law would be totally incomprehensible, even if the science behind it was flawless and identified the real mechanisms responsible for human action. There is no escape from the folk-psychological conceptual scheme: any legal system must be based on it, otherwise it would become an incomprehensible, futile exercise in theory-construction. A law taking advantage of the conceptual apparatus of eliminative materialism would remain law in books, never becoming law in action.

4 The legal responsibility of autonomous artificial agents

So far we have argued that (1) neither extreme restrictivism nor extreme permissivism are a viable choice when it comes to deciding whether autonomous machines may be legally responsible—the construction of legal responsibility must lie in-between those two extreme options; and that (2) the concept of legal responsibility is deeply rooted in the folk-psychological understanding of human action—the legislator cannot disregard the fact that people conceptualise their own behaviour as well as the behaviour of others in terms of intentions, goals, beliefs, etc. In other words, we have tried to show that although legal responsibility is not founded on solid metaphysical grounds (the real existence of free will, intentions, etc.), there are conceptual limitations which prevent the legislator from departing too far from the folk-psychological understanding of what responsibility consists in. The addressees of the law would simply fail to understand a view of legal responsibility which is essentially different from how the folk-psychological conceptual scheme structures human action.

Let us now come back to the question whether—and if so, under what conditions—autonomous artificial agents may be held legally responsible for their actions. Arguably, the fact that "genuine" legal responsibility is a theoretical reconceptualisation of the relevant aspects of folk psychology, may have no bearing on the question of the legal status of autonomous machines. After all, legal systems make room for civil and criminal liability of legal persons, such as corporations, foundations, states or municipalities (Wells 2001; Leigh 1982). It would be difficult to argue that a corporation or a state has (or is ascribed) "free will" or "intention" or may act "wilfully" or "knowingly". Nevertheless, we find it possible to hold them legally responsible for certain actions. Similarly, it may not be necessary to ascribe intentions or free will to an autonomous machine in order to deem it responsible for what it does.

We believe that there is an important problem with this argument: the analogy between legal persons and autonomous machines is far from perfect. The reasons for holding corporations and similar legal entities responsible for certain events are quite straightforward. First, in some cases, it may be difficult to identify an individual, whose actions caused the damage (e.g., that a patient at a hospital got infected with some disease), while it is obvious that it happened in relation to the operations of a legal person (e.g., a hospital). Second, legal persons usually have much deeper pockets than individuals, and hence it is reasonable to ask the restitution from them rather than sue an individual. However, let us observe that the actions of a legal person are always traceable back to the actions of an individual person or a group of persons, even if it is not possible to clearly identify them. In other words, legal responsibility of corporations and similar entities is connected to acts performed by their representatives or employees. From this perspective, holding legal persons legally responsible does not go against the folk-psychological understanding of agency; rather, it is a kind of "prosthesis" the legal system utilises to ensure a more efficient and swift restitution of justice. Moreover, on closer inspection things are more complicated than they appear on the surface. A case in point is criminal responsibility of legal persons. In some jurisdictions (e.g., in Germany), a legal person cannot be subjected to criminal proceedings; in other legal systems (e.g., in France or in Poland), criminal responsibility of legal entities is possible, but only under the condition that an individual physical person, who committed the act "on behalf" of the legal person, is identified. This clearly shows that the idea of the legal responsibility of legal persons is not a natural extension of the folk-psychological view of human agency: even if we allow legal persons to be charged with criminal offences, much is done to link it to the actions of individual human beings (Khanna 1996).

Given those considerations, it is safe to say that legal responsibility of an autonomous machine would be quite different from the legal responsibility of a legal person. Although the pragmatic justification would be quite similar (e.g., to ensure a swift and efficient administration of justice), the actions of an autonomous artificial agent (as we understand the term in this paper) would not be directly traceable back to the actions of a human agent. Thus, we believe that the fact that there exists legal responsibility of legal persons has no relevance what so ever for the question of the accountability of autonomous machines. Behind the legal person there always is some human being. An autonomous artificial agent, on the other hand, would need to be regarded as a real actor in the web of social interaction, not a mere legal facade, which facilitates the pursuit of justice or offers more efficient means of redress.

Therefore, we believe that the problem pertaining to the legal responsibility of autonomous machines boils down to the question of whether such machines can be seen as genuine agents through the prism of folk psychology. Let us consider the following thought experiment. We meet a new person, Mr. Y. He is an intelligent and likeable man, easy to get along with. He is also ready to help, when the need arises, excels in conversation and enjoys spending time with his new friends. He acts with dignity in difficult situations, and whenever he is mistreated by his hotheaded employer, we feel sympathy for him. However, one day we learn that he is a sophisticated android, equipped with state-of-the-art autonomous algorithms, enabling him to navigate the troubled seas of social interactions. The algorithms are so perfect that they calculate the adequate reaction to any social encounter, are able to recognise complex patterns of facial expression in any interlocutor, estimate the intensity of the emotions expressed by others, and mimic the behaviour of someone who is in pain, feels sorrow, is happy or troubled. The algorithms are so cleverly designed that Mr. Y, who in some tasks - such as playing chess or doing complex calculations - is orders of magnitude better than we are, never displays those abilities when interacting with us. The question is, whether we change our attitude towards Mr. Y? We believe that most of us would treat him differently. Before discovering his real nature, i.e. the fact that Mr. Y is a purely deterministic device, we would have treated him as any other human being, attributing him beliefs, intentions, emotions, and free will. After realising that Mr. Y is an android, we would probably change our attitude. After all, Mr. Y does not really feel physical or emotional pain, so why be sympathetic when his boss treats him badly? When helping us, he had sacrificed nothing—he just acted as his algorithms dictated—so why should we be grateful? He is not a moral or a legal agent, but a sophisticated machine, which resembles a car or an iPhone much more than it resembles a human being. His actions are dictated neither by sentiments towards us nor by reason, and so he is as blameworthy for his deeds as a falling stone, which accidentally injures a passerby.

Of course, this claim may be contested. For instance, one may observe that people often have a tendency to ascribe intentions to inanimate objects. Ever since the famous experiments of Heider and Simmel (1944), it has been well known that we are able to spontaneously conceptualise the observed movement of even simple geometrical figures such as circles and rectangles as an intentional action. This tendency may be well rooted in our evolutionary past: it has been speculated that from birth humans have the ability to perceive things in two diametrically different ways, as governed either by causality or intentionality (Bloom 2004). However, such a "default" conceptualisation of some phenomena as intentional surely has its limits. On the one hand, it is unlikely that one would be willing to ascribe agency or intentionality to circles and rectangles beyond the confines of an relatively isolated event. To put it differently, it is unlikely that one would ascribe intentions, goals and beliefs to inanimate objects in any systematic way. On the other hand, research in developmental psychology suggests that even if small children are eager to see robots as animate and intelligent agents, the older they become the less animistic are their intuitions regarding machines (Okita and Schwartz 2006). This can be interpreted as evidence that the gradual shaping of folk-psychology in the process of inculturation leads to the exclusion of robots from the pool of intentional agents. These observations seem to reinforce our intuition that once the discovery is made that Mr Y is an android, he would no longer be treated as a moral agent by most people.

If we are right, then the same applies to all kinds of autonomous machines, which resemble human beings even less than Mr Y. From this perspective, it is hard to imagine the communal understanding and acceptance of granting autonomous machines the status of legal agents. Of course, this conclusion in an essential way depends on our current folk-psychological conceptual scheme. After all, in our past we have treated animals and even inanimate objects such as rivers or volcanoes as actors in the web of social interactions. It is not impossible to imagine that the evolution of our conceptual apparatus will reach a stage, when autonomous robots become full-blooded moral and legal agents. However, today at least, we seem to be far from this point.

5 Conclusion

In this paper, we have argued that autonomous machines cannot be granted the status of legal agents. Although this is quite possible from purely technical point of view, since the law is a conventional tool of regulating social interactions and as such can accommodate various legislative constructs, including legal responsibility of autonomous artificial agents, we believe that it would remain a mere “law in books”, never materialising as “law in action”. The reason is that the law, and in particular such a fundamental institution as legal responsibility, must be comprehensible for the people who are subject to legal rights and obligations. Meanwhile, as we have argued, our perception of human action and responsibility—i.e. our folk psychology—is not suited to see autonomous machines as the authors of their actions. This may change with the evolution of our conceptualisations of agency, but such changes require much time.

It should be added that our position does not favour any concrete view of moral or legal agency. In particular, it does not lead to accepting a kind of utilitarianism, which posits, inter alia, that autonomous machines cannot be moral (or legal) agents, since they cannot suffer from any punishment—they feel neither pain nor pleasure (Sparrow 2007). We do believe that the ability to experience emotions is a part of the folk-psychological understanding of personhood and agency. However, the ability in question is not the sole foundation for ascribing moral or legal responsibility: it is important, insomuch as it constitutes an aspect of the folk-psychological conceptualisation of agency. In other words, when referred to by the followers of Bentham, the capacity to feel pain and pleasure becomes a normative criterion for ascribing moral (or legal) responsibility; similarly, when Kantians speak of an autonomous application of reason (i.e. not influenced by factors external to reason itself, e.g. emotions) to moral and legal questions, they are formulating a normative criterion of moral (legal) agency: an agent incapable of such autonomous reasoning cannot be subject to moral (legal) rights and duties. Our claim is not normative, it is rather descriptive and conceptual: the ascription of moral and legal responsibility is always mediated through the folk-psychological understanding of agency.

Acknowledgements This paper was written within the research project ‘Law & Neuroscience. A New Paradigm in the Philosophy of Law’ sponsored by the Foundation for Polish Science.

References

Bloom P (2004) Descartes’ baby. Basic Books, New York

Bryson JJ, Grant TD, Diamantis ME (2017) Of, for, and by the people: the legal lacuna of synthetic persons. Artif Intell Law. doi:10.1007/s10506-017-9214-9

Churchland P (1981) Eliminative materialism and the propositional attitudes. J Philos 78:67–90

Davies M, Stone T (eds) (1995) Folk psychology: the theory of mind debate. Blackwell, Oxford

Dennett D (2008) The philosophical lexicon. http://www.philosophicallexicon.com. Accessed 12 Apr 2017

Fischer JM, Ravizza M (2000) Responsibility and control: a theory of moral responsibility. Cambridge University Press, Cambridge

Fodor J (1987) Psychosemantics. MIT Press, Cambridge

Franklin S, Graesser A (1997) Is it an agent, or just a program?: a taxonomy for autonomous agents. In: Müller JP, Wooldridge MJ, Jennings NR (eds) Intelligent agents III agent theories, architectures, and languages. ATAL 1996. Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence), vol 1193. Springer, Berlin

Greene J, Cohen J (2004) For the law, neuroscience changes nothing and everything. Philos Trans R Soc Lond B Biol Sci 359(1451):1775–1785

Hage J (2017) Theoretical foundations for the responsibility of autonomous agents. Artif Intell Law. doi:10.1007/s10506-017-9208-7

Haidt J (2001) The emotional dog and its rational tail: a social intuitionist approach to moral judgement. Psychol Rev 108:814–834

Heider F, Simmel MA (1944) An experimental study of apparent behavior. Am J Psychol 57:243–249

Horwitz M (1992) The transformation of american law, 1870–1960: the crisis of legal orthodoxy. Oxford University Press, Oxford

Hutchison A (2014) The Whanganui river as a legal person. Altern Law J 39(3):179–182

Kelsen H (1945) General theory of law and state. Harvard University Press, Cambridge

Khanna VS (1996) Corporate criminal liability: what purpose does it serve? Harv Law Rev 109:1477–1534

Leigh LH (1982) The criminal liability of Corporations and other groups: a comparative view. Mich Law Rev 80(7):1508–1528

Meltzoff AN, Decety J (2003) What imitation tells us about social cognition: a rapprochement between developmental psychology and cognitive neuroscience. Philos Trans R Soc Lond B Biol Sci 358(1431):491–500

Morris M, Peng K (1994) Culture and cause: American and chinese attributions for social and physical events. J Person Soc Psychol 67(6):949–971

Nahmias E, Morris S, Nadelhoffer T, Turner J (2005) Surveying freedom: folk intuitions about free will and moral responsibility. Philos Psychol 18(5):561–584

O'Connor T (2010) Free will. The stanford encyclopedia of philosophy. https://plato.stanford.edu/entries/freewill. Accessed 12 Apr 2017

Okita SY, Schwartz DL (2006) Young children's understanding of animacy and entertainment robots. Int J Humanoid Robot (IJHR) World Sci 3(3):393–412

Pound R (1910) Law in books and law in action. Am Law Rev 44:12–36

Sparrow R (2007) Killer Robots. J Appl Philos 24(1):62–77

Stich S (1983) From folk psychology to cognitive science. MIT Press, Cambridge

Stich S, Ravenscroft I (1992) What is folk psychology? Cognition 50:447–468

Summers RS (2006) Form and function in a legal system: a general study. Cambridge University Press, Cambridge

Want SC, Harris PL (2002) How do children ape? Applying concepts from the study of non-human primates to the developmental study of 'imitation' in children. Dev Sci 5:1–14

Wells C (2001) Corporations and criminal responsibility. Oxford University Press, Oxford