ICRES 2019: International Conference on Robot
Ethics and Standards, London, UK, 29-30 July 2019.
https://doi.org/10.13180/icres.2019.29-30.07.020

A PRAGMATIC APPROACH TO THE PROBLEMOF AGENCY FOR ROBOT ETHICS

ENDRE E. KADAR

Department of Psychology, University of Portsmouth, King Henry Bldg.
Portsmouth, PO1 2DY, Hampshire
E-mail: endre.kadar@port.ac.uk; kadar_e@yahoo.co.uk

Recent increase in robot autonomy resulted in increase in safety and ethical concerns. Arguably, designing robots with high level of autonomy for safe and ethically acceptable interaction with humans implies a profound understanding of human behaviour control including social interaction. This problem is often framed as creating an artificial agent that can safely interact with humans and its behaviour is ethically acceptable. In robotics, there seems to be a consensus on the definition of autonomous agents despite the fact that there are many variants in use. The present paper provides a critical review of this "standard view" and argues for a more inclusive definition of agency and moral agency, which is labelled as a pragmatic approach because this is more in line with the more inclusive everyday notion of agent and artificial agents need to be safe and ethical designs that are for everyday use, for fluent interaction with humans.

1. Introduction

Autonomous robots whose interaction with humans is safe and ethically acceptable are artificial agents who are behaving safely and in a morally responsible way. Research highlighted that designing these agents engineers need to understand human perception and action in behaviour control including those skills that are needed in interaction (Kadar et al., 2017; Kadar, 2019). In other words, designing artificial agents for fluent interaction with humans requires profound understanding of human social interaction skills. Biological agents possess complex action repertoire and high level of autonomy within their natural environment. These agents often engage in social interactions with other biological agents within and across species in a variety of ways and they do it in a surprisingly easy way. In contrast, designing artificial agents with complex action repertoire and high level of autonomy within a limited and often simple environment is still a challenging task despite huge progress in robotics. Also, these artificial agents are usually designed to interact with only humans and that requires implementation of basic skills of interaction with humans because the performance of an artificial agent should be similar to human-like behaviour otherwise the interaction would be difficult. If an artificial agent with high level of autonomy is also designed for interacting with humans as a moral agent, the design would further be biased towards human-like behaviour. Thus, it is not surprizing that robot engineers use humans as the prototype to conceptualise autonomous moral agent and adapt human characteristics as requirements for artificial agents. Accordingly, the term 'agency' is commonly used to denote the ability of humans to perform intentional actions. Autonomous agents can have their goal as the target of their intention and they persist in performing the required behavior control to achieve that goal. Moral agency requires additional criteria to satisfy behavior standard and these requirements are believed to be associated with consciousness (Himma, 2009). Whilst

autonomous agents could be modeled from various biological species, moral agents are mostly designed based on human moral standards. First we will be tracing the origin of this standard view. Then a critical assessment of the standard view will be provided. Finally, we will argue for the need for a broader view that is more in line with the everyday use of the term agency. Based on this review the present paper proposes a more pragmatic approach to the way artificial moral agency should be defined.

2. The standard view of agency and moral agency

The discussion of the importance of agency and its role in everyday life has a long history in philosophy and it can be traced back to Aristotle but the modern use of the term and related concepts such as free will, mind, the relationship between the mental and physical are mostly associated with Descartes, Hume, Kant and other more recent thinkers. In contemporary analytic philosophy, it is most commonly associated with the influential work of Anscombe (1957) and Davidson (1963). Although Anscombe's and Davidson's views differ significantly in many respects, they seem to agree that action is to be explained in terms of the intentionality of intentional action. To put differently, agency is associated with goal directed actions, that can cause chains of events that cannot be explained without an agency, or the explanation of certain chains of events cannot be reduced to causes that do not include an agency. Not surprisingly this is a controversial issue in many ways, but perhaps the most difficult aspect of this view is the fact that intentionality is usually considered as a mental state. Specifically, the problem is widely discussed in the discipline of *philosophy of action,* which is actually largely focusing on the notion of intentional action. The notion of moral agency seems to be less controversial in philosophy. For instance, *Routledge Encyclopedia of Encyclopedia* defines moral agents whose behaviour are expected to meet the demands of morality. According to *Stanford Encyclopedia of Philosophy,* a moral agent is an agent who is open to responsibility ascriptions, while *Internet Encyclopedia of Philosophy* states that the main characteristic of moral agents is that they can be held accountable for their actions (i.e., praised or blamed, punished or rewarded). In sum, these definitions imply that a moral agent's behavior is governed by moral standards but they emphasize different aspects of meeting this moral standard. Stating that an agency has moral duties or moral obligations is to say that the agent's behavior should be guided by and can be evaluated under those standards. Moral agents are subject to moral standards. In other words, they are *accountable* (or morally responsible) for their behavior under those standards. There are multiple reasons why moral agents are assumed to be conscious beings and as such, there is an associated old tradition of excluding non-humans from the class of moral agents. Perhaps, the simplest factor that can reveal the need for consciousness is the non-biological, the cultural nature of moral standards that implies learning, which requires complex mental skills.

These definitions are widely discussed and often debated in philosophy but for the purpose of the present discussion our main concern is that they seem to be difficult to use in robotics for defining artificial agents and artificial moral agents. The main difficulty lies in the mental (intentionality and consciousness) and cultural requirements that are considered unique to humans. The next section will reveal that these stringent requirements of mental and cultural processes typical of humans are not necessary and they can be ignored in discussing behavioral patterns of interacting biological agents.

3. A critical assessment of the standard view of agency and moral agency

Descartes followed the old western tradition of separating humans from other animals by using a few unjustified assumptions. He believed that humans have free will, possess mind with the ability of rational thinking. In contrast, he argued that animals are mindless machine like creatures without mental life and even without feelings. These are the two pillars of the standard view of agency and moral agency, which can be critically reevaluated in the light of scientific findings. Thus, the standard approach to agency and moral agency can be critically assessed without entering into speculative philosophical reasoning. There are two fundamental assumptions of the Cartesian approach but these are not supported by scientific evidence.

First, humans are not very good at mental planning and setting out intentions before initiating intentional actions. Also, humans are not only far from perfect in applying moral standards in their everyday behavior but quite often they are confused about important aspects of the moral standard of their culture. For instance, Carucci (2016) presented a few examples in workplaces to demonstrate why normal ethical people do not meet their standard by make unethical choices. According to *Stanford Encyclopedia of Philosophy* "there is considerable psychological and anthropological evidence that a small number of core moral values are espoused universally, such as: benevolence (avoiding harm to others and offering aid when the costs are not high); fairness (reciprocating help and sharing goods); loyalty (especially to family and community); respect for authority (of one's parents and community leaders, when it is exercised responsibly); personal purity in body and mind (notably as it reflects moral character); and freedom (especially from oppressive control by others). see Haidt (2012); Haidt & Joseph (2004). ….. Nevertheless, these values are often interpreted and applied differently not only across cultures but also across time within roughly the same cultures. " One does not need to dig deep into the scientific literature to find examples for disagreements. It is widely accepted that telling lies is wrong. In particular, telling lies regularly is judged as extremely unethical even if occasional lies can be tolerated assuming they serve a positive purpose. The current president in the United States is a clearly a perpetual liar and his followers are his "true believers" who are surprising tolerant and accept this kind of behaviour because they believe that this behaviour can also serve a positive purpose for some members of the society.

Second, scientific research has revealed that mental life is not unique to humans. Darwin (1872) has already anticipated that there must be a certain level of continuity in the evolution of mental life in animals including the evolution of intentionality and various types of consciousness. During the 20th century it became a common place that animals show goal directed behavior, which is the core of intentional action. Even primitive creatures such bacteria show behavior that seem to imply intention of moving into areas with preferred conditions and moving away from non-preferred conditions (Papi, 1992). But intentional mental state may not be necessary for goal directed behaviour. Similarly, recent research findings clearly suggests that there are animals who can be considered as conscious and even self-aware despite that fact that their consciousness is different than human consciousness. Surprisingly, there seems to be a consensus among neuroscientists, that the octopus is a very sophisticated animal with mental life and complex skills even though its nervous system is quite different than other invertebrate CNS (The Cambridge Declaration, 2012).

Based on these rather surprising new findings, perhaps it is no surprise that research suggests that moral behavior can also be observed in some animals. In a recent book, Bekoff and Pierce (2010) provided numerous examples for ethical behavior in various species. They suggest that some of the key components of social interaction, the building blocks of ethical behavior are older than humans in evolution. Human morality can be rooted in psychological tendencies and capacities such as empathy, reciprocity, a desire for co-operation and harmony that appeared earlier in evolution than primates and humans. Bekoff and Pierce (2010) provide examples for empathy and proto-ethical behaviour in dolphins, whales, bats, and rodents. Examples include even cross species empathy. For instance, a herd of 11 elephants rescued antelope from an enclosure in South Africa. The matriarch of the herd unfastened all of the metal latches holding the gates closed allowing the antelope to escape.

These examples for cross species empathy and simple moral behaviour patterns, however, should not be surprising because humans have experienced cross-species moral interactions with various species. Bekoff and Pierce (2010) mentioned dolphins helped humans, and there are large number of examples for human-dog interaction that includes empathy, caring and helping as every dog owner could confirm. Neuroscientists could also support some of these findings by showing that distantly related mammals such as whales and dolphins have the same structures in their brains that are thought to be responsible for empathy in humans (de Waal & Preston, 2017). Needless to say, however, that cross-species interaction can also be violent. This is typical of predator – pray relations in animals, where one species is feeding on the other, but there are also examples for territorial animals that are protecting their habitat by violent attack against any intruders.

There are also examples in more primitive animals for posing danger for other species including humans. Certain types of bacteria infection could be deadly not because the bacteria are intentionally killing humans, but because of their nature. As they multiplied inside the host animal they cause damage internally, which could become fatal. There are species carrying infectious viral or bacterial disease that could be fatal for humans but they are affected by these viral or bacterial infections. This is similar to the case in epidemics amongst humans when the carrier of a disease is not even aware of the infection but the disease can be transmitted to other humans unintentionally. These are examples for safety rather than ethical behaviour but as soon as the disease carrier (host) becomes aware of the disease and not protecting other humans by avoiding contact with them an ethical dimension of the issue emerges as well. These examples with human animal interactions for good or ill could provide valuable lessons for dealing with or managing human robot interactions while maintaining safety and keeping ethical concerns suppressed.

4. Lessons from biology for social management of robot – human interaction

Our examples for cross species interactions could provide valuable insights into how human robot interaction should be viewed and assessed. Currently, there is excessive research on imaginary situations of robots becoming so smart that they could pose danger for humans. These examples of danger are based on science fiction scenarios of artificial agents that could possibly take over the control over humans or pose danger for us including ethical concerns. Apart from the problem of autonomous robots behaving out of control by mistake or some damage to the robot, these futuristic scenarios will be mostly real danger for the future but they are not imminent threat and consequently they are not as important as other examples that

are obvious analogues to current human – animal cross species interaction. These cross species interactions can provide valuable insights on how we can cope with various robot-human interactions safely while also avoid or minimize ethical concerns.

Robots like biological agents can become dangerous intentionally and/or unintentionally. In Nature, humans regularly interact with various biological agents who can have complex skills and pose dangers (e.g., snakes, tigers, etc.) but they could also be beneficial (e.g., dogs, horses, etc.). The human contact with these biological agents is partly regulated by laws in different countries in different ways (e.g., pets are often defined by laws, dangerous animals are kept out of human contacts, wild animals kept in cages in zoos, etc.). In these partly regulated human-animal interactions, the importance of mental capacity/ability (intentionality or consciousness) of the animal is not considered as fundamental. Biological agents with various level of complexity can either pose danger or can be beneficial for humans. The dangers and benefits are often intrinsic to the biological/evolutionary design of the animal and humans cannot change it, but humans can have limited control on those design characteristics by taming a specific individual animal. For instance, tigers are wild animals and pose dangers, whilst snakes are not always dangerous but they are posing ethical concerns for humans because humans are not familiar with snakes and usually cannot distinguish between poisonous and not poisonous snakes by the look of the snake.

The laws that regulate human contacts with animals (conditions for pets, animal rights, etc.) are pragmatic. Similarly, everyday human behaviour and attitude towards animals are also driven by pragmatic considerations. Biological entities of all kinds are called biological agents regardless of their mental capacity if any. They can be a source of various types of impact, such as a possible cause of negative impact on humans like carrier of fatal disease (like bats can carry ebola virus in central Africa) but they can also be a resource (dogs, horses, camels, elephants, etc.) for humans. There is a wide range of examples how animals and humans are related in everyday interactions. One extreme and natural form is not sharing habitat (living in separate niche and avoiding close encounters). Keeping wild animals in zoos in enclosed space is an artificial replication of the natural separation of these animals from direct human contact. At the other end of this spectrum of animal – human contact is sharing habitat with animals. Pets are in direct and regular contact with their owners. An interesting new trend can also be observed in various parts of our planets. Some colonies of wild animals move into human habitats in search of food because they lose their habitat or some other reasons (e.g., wild pigeons are populating cities around the globe, wild foxes can be seen regularly in British towns at night, monkeys in India are living in cities, etc.). Some of the species that choose to live closer to humans are often brave enough for begging for food without any fear from humans. These changes in animal attitude towards humans are not without danger because these animals can be hosts of various infectious diseases. The risk, however, does not seem to be high based on empirical experience. There are only a sporadic cases of infection contracted from contact with wild animals.

In robotics, some of the definitions are already avoiding intentionality requirements on autonomous robots, but we propose here that the common requirement of action should also be relaxed. Pragmatically, the key issue is whether an agent could be a cause of safety or ethical concerns regardless of its "mental capacity" such as intentionality or consciousness. In sum, the essential aspect for everyday use of the term is wherever there is potential to cause changes for good or ill for humans or in Nature, this source could be named as an agent, a

source of change. This is a pragmatic definition because it can eliminate the difficult search for the original source of change, which could sometimes be nearly an impossible task.

5. Proposed definitions for agency and moral agency

By looking at natural agents such as various animals and their possible interaction with humans, we can arrive at a pragmatic position of accepting the everyday notion agency and moving away from philosophical definition of autonomous agency that implies action with intentionality. In everyday language use, we talk about actions of chemical agents, biological agents, collective agency and other forms of agency. This new approach is focusing on the ability to be a cause, a source of a chain of events that could be either beneficial or could have a negative impact for humans. In the 1993 edition of the Webster's New Encyclopedic Dictionary action is defined the following way:

> 1: a proceeding in a court of justice by which one demands or enforces one's right or the redress or punishment of a wrong 2: the working of one thing on another so as to produce a change <the action of acids on metals> 3: the doing of something usually in stages or with the possibility of continuation <the action of singing> 4 a: a thing done: deed b pl : behavior, conduct c: readiness to engage in daring activity: initiative <a person of action > 5: combat in war 6: the unfolding of the events of a drama or work of fiction: plot 7: an operating mechanism <the action of a firearm> 8: an area or state of vigorous activity <where the action is> (1993: 11)

As we already anticipated, none of these definitions refers explicitly to the concept of intention, which several authors use traditionally to define a true 'action' (e.g. Anscombe,1957, Davidson,1963 , and Searle (1983)). Moreover, two of the listed definitions (2 and 7) are even compatible with the idea of a non-biological agency exerting action (i.e., 'action of acids on metals' or 'action of a firearm').

Regardless of the differences in the meaning, the term 'action' refers to the process of initiating some kind of change, which is usually called a cause rather than action. In that sense, we are going back to the bare essentials of changes and the everyday intuitive idea of whatever the cause of changes are, they are called agents. These could be biological creatures but they could be some chemicals, or other nonliving systems. Aristotle has already been mentioned in this paper and it is time to rephrase our problem of defining agency using Aristotelian thoughts on causation as sources of changes. Artistotle's Science (Natural Philosophy) is sometime labeled as organic based on the fact that motion/change/ was its central concept and each motion is associated with a mover (Bk. 8 of the *Physics* argues for the additional thesis that for each motion, whether natural or contrary to nature, there needs to exist a mover.) To put differently, within this system agency is central because the notion of mover could be used to define agency, which includes the first mover (God), living creatures and could also be artificial mechanisms (e.g., tools, machines, etc.). Another important part of Aristotle's science is the way things, changes, non-changes in Nature are described. To be more exact, the famous doctrine of four causes is meant to provide explanation of why things are as they are or why changes happen the way they do. The four causes can be grouped into two pairs of causes. The more fundamental pair of causes are the material and formal causes.

Material causes provide potentials, which are or can be actualized by the formal cause. Thus, these two are related rather than independent aspects of explanations. The other two causes are more closely related to changes in Nature. The efficient and final causes are beyond but not necessarily independent of the material and formal causes. The efficient causes initiate processes that lead to changes, while the final cause is the end of this process of changes, what efficient causes intended to achieve.

Our proposal is in line with Aristotle's organic view of Nature and conforms to everyday intuition. Accordingly, anything that is initiating a change (motion) can be called an agent. Thus, agents could be not only biological entities but non-living things such as a volcano that may cause tremor or eruption, a chemical that can be poisonous for humans, acid that causes corrosion of metals, etc. Similarly, a moral agency does not have to be defined based on human moral standards. Instead, we propose that compliant well-trained pets could be the prototypes for moral agents. This way those animals or artificial agents that are behaving friendly with humans because they are trained that way and they do not pose any danger or threat for humans, those can be considered as moral agents for human interaction. This definition does not require mental life and knowledge of moral codes. This is a pragmatic approach to eliminate the obstacle of stringent requirement for being a moral agent similar to humans. For those, who would like to keep the standard definition of moral agent, it might be a pragmatic solution to call our proposed definition for moral agency as a definition of *compliant moral agency*.

6. Conclusions

In this paper we critically reviewed the standard view of agency and moral agency. In robotics, there seems to be a consensus on the definition of autonomous agents despite the fact that there are many variants in use. The present paper provided a critical review of this "standard view" and argued for a broader, more inclusive definition of agency and moral agency. Our definition of agency is based on everyday use of the term agent rather than its modern scientific version. Any source of change or motion can be considered as an agent. This is a pragmatic approach and can be related to Aristotle's organic view of Nature and his four types of causes. This definition eliminates the stringent requirements of intentionality and goal directedness of action. Similarly, our proposed definition of moral agency is derived from everyday intuition of our interaction with pets such as dogs and cats. The relevance of animal-human interaction for better understanding and designing artificial agents have already been recognized but was constrained by the standard view of moral agency and its human model (Calverly, 2006; Levy, 2009). Our proposal also eliminates the stringent requirement of consciousness and awareness of moral codes and norms from the definition of moral agency. Instead, we define moral agency based on compliance in interaction with humans. Accordingly, those biological and artificial agents that can interact with humans and do not pose any safety risk or ethical concerns those are ethical agents. One may call this as a definition a compliant moral agent to differentiate from the standard view on moral agency.

These definitions can simplify the requirements for robot designers. For the current purpose of safe and ethical designs these definition are for everyday use and ordinary humans should understand and be able to use artificial agents based on these definitions, because these designs of artificial agents are for fluent interaction with humans without any safety and ethical concerns.

References

Anscombe, G. E. M., (1963). *Intention*, 2nd ed., Ithaca, NY: Cornell University Press.

Beckoff, M. & Pierce, J. (2010). *Wild Justice: The Moral Lives of Animals*. University Chicago Press.

Calverley, D. J. (2006). Android Science and animal rights: Does an analogy exist? *Connection Science, 18*, 4003-417.

Carucci, R. (2016). Why ethical people making unethical choices. *Harvard Business Review*. December. 16.

Darwin, C., Ekman, P., & Prodger P. (1998/1872). *The Expression of the Emotions in Man and Animals,* 3rd edn, London: Harper Collins.

Davidson, D. (1963). Actions, Reasons, and Causes, *Journal of Philosophy, 60*, 685–699.

De Waal, F. B. M. & Preston, S.D. (2017). Mammalian empathy: behavioural manifestations and neural basis. Nature Review Neuroscience. 18, 498-509.

Haidt, J. (2012). *The Righteous Mind: Why Good People Are Divided by Politics and Religion*, New York: Pantheon Books.

Haidt, J. & Craig, J., (2004). Intuitive Ethics: How Innately Prepared Intuitions Generate Culturely Variable Virtues, *Daedalus, (Fall*): 55–66.

Himma, K. E. (2009). Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent? *Ethics and Information Technology, 11*. 19-29.

Kadar, E.E. (2019). Mind the gap: A theory is needed to bridge the gap between the human skills and self-driving cars. In Aldinhas Ferreira, M.I., Silva Sequeira, J., Gurvinder, V., Tokhi, O., Kadar, E. E., *Robotics and Well-being*. (pp. 55-65). Springer.

Kadar, E. E., Köszeghy, A., & Virk, G. S. (2017). Safety and ethical concerns in Mixed Human-Robot Control of Vehicles. In M. I. S. Ferreira, J. S. Sequeira, M. O. Tokhi, E. E. Kadar, G.S. Virk, *A World with Robots: International Conference on Robot Ethics: ICRE 2015*. (pp. 135-144). Springer.

Levy,D. (2009).The ethical treatment of artificially conscious robots. *International Journal of Social Robotics, 1*. 209-216.

Papi, F. (1992). Animal homing. Springer. The Netherlands.

Searle, J. R. (1983). *Intentionality: An Essay in the Philosophy of Mind.* Cambridge University Press.

Webster's New Encyclopedic Dictionary (1993). New York: Black Dog & Leventhal.