The Return of Magical Thinking?

Ylva von Gerber
Christian Balkenius*
ylva.von_gerber@fil.lu.se
christian.balkenius@lucs.lu.se
Department of Philosophy, Lund University
Lund, Sweden

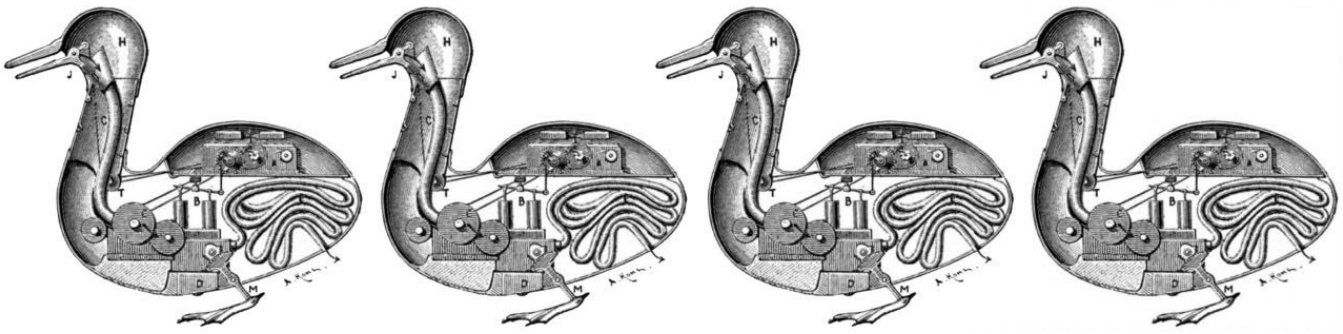

Figure 1: Canard Digérateur. Jacques de Vaucanson's mechanical duck from 1739 that faked a digestive system.

ABSTRACT

Is it possible for an artificial automated system to have agency, here understood as a *self*-governing agent that on her own has authority to initiate her actions[12]? We argue that so called autonomous systems are not autonomous in any real sense since they are bound to follow their programming and therefore do not take independent decisions. Agency, in the sense stated above, is linked to autonomy and autonomy in turn is linked to morals. It is therefore incorrect to ascribe moral agency to them. An artificial automated system does not, on its own, initiate any actions at all unless programmed to do so. Instead we propose that we distinguish between the truly autonomous (moral) agents that create the systems and the system itself that in ideal cases make decisions that are congruent with the morality or intentions of its designer.

KEYWORDS

autonomy, moral agency, autonomous systems, robots

1 INTRODUCTION

When the concept of autonomy is being used within technology, computer, and engineering science it diverges from how the concept is used within social sciences and humanities. The former makes no distinction between *automaton* (automated) and *autonomous*, whereas the latter are very precise [1, 10]. This difference matters as we now see a development within the AI discourse into a discussion about artificial moral agency (AMA), which blurs the debate.

The question is whether this discussion would ever have come about unless the concept of autonomy had not been used within the sphere of technology that develops automated AI systems in the first place; seemingly unaware about that one has been using the concept in a very different sense than the way it is understood and applied in the domain of political science or philosophy. Within the latter autonomy plays a crucial part for *moral agency* [7].

It must be emphasized that, when using a term responsibly, one should make its meaning as precise as possible, or as precise as the circumstances require [3, 4, 9]. The encompassing character of the term autonomy – similar to that of an umbrella – harbors many different meanings, some partially overlapping. Yet a term with too many points of reference, too many distinct interpretations, becomes at the same time devoid of tangible contents and may paradoxically result in too indeterminate a reference [6].

In von Gerber's work [6], the concept of autonomy was scrutinized within three contexts: in constitutional law, in the philosophy of Immanuel Kant and in the contemporary philosophical discussion about personal autonomy. When tracing the concept of autonomy historically, similarities and dissimilarities between different definitions were highlighted. It was proven that the lowest common denominator for the three different definitions was a "self". Being self-reflective and self-aware permits the actor to understand herself as an agent, who can deliberate about and choose how or if to initiate an act[8].

At least three questions about AI and AMA seem to fascinate and interest people. The classical philosophical dialogue might come handy in trying to answer these three questions as well as trying to argue for the answers given in the following.

2 WHAT IS ARTIFICIAL MORAL AGENCY?

There is no such thing as artificial moral agency - it is a *contradictio in terminis*. An artificial – here understood as a manmade (man also invented the concept) – machine, technology, application can only get its morals or ethics from someone else and can only act according to the given ethics – what someone else considers to be right or wrong.

The "artificial moral agent" does not have a self, no self-awareness, no self-consciousness, *per se* and thus, consequently, no moral agency. It does neither, on its own, choose which rules or norms it shall prescribe for itself, nor whether to follow them or not. The artificial moral agent does not possess, are not born with (they are in fact *not born*), free will and hence no *autonomy*. If moral actions (moral agency) only consist of obedience, imposed, or programmed from outside, from someone else, they lack moral value [7]. Show me a robotic lawn mower that *decides for itself* to stop doing what its programmed to do and jump into the neighbour's swimming pool, to relax, instead!

3 IS ARTIFICIAL MORAL AGENCY POSSIBLE?

Can there be such a thing as artificial moral agency? Our answer is a distinct no. Moral agency demands that someone – a *self* – can have or build an opinion about what is right or wrong, good and evil, and choose to act at one's *own* discretion. Someone chooses for oneself. From within. If anything should be said to have moral value this is the crucial part – to choose for *oneself*, on one's own, if or how one wants to, shall och should act.

To be able to claim that something has moral value or relevance there must be a choice – to choose good or evil, to say yes or no. Or, for that matter, not to act at all; e.g. if Robot created to work with facial recognition should choose to *refuse* to execute the task it has been made to fulfil.

4 WHAT WOULD ARTIFICIAL MORAL AGENCY DENOTE FOR E.G. RESPONSIBILITY OR MORAL DUTIES?

Artificial moral agency would denote nothing, as there is no such thing as artificial moral agency. To claim that a thing or, for that matter, an animal, to make another example, is accountable for an action, has both historically as well as nowadays been regulated in law – a norm system directed to the owner, hence, neither to the algorithm creator nor the robot constructor, and definitely not to the robot or an animal itself.

There are however historical examples, as well as from the present age from different parts of the world, where magical thinking, a magical view upon the world, and one can learn that stones are believed to have agency.

Given the questions above and particularly the answer given to question three, one might ask oneself if we aren't on the way (back) to a world of magical thinking?

It appears like the discussion has left the world of the living – the hungry, the passionate, the suffering – sentient beings, born with survival instinct, and starts from a world of things where one wants to apply parts of the empirical and philosophical research that tries to explain our way of being, of living, to work. But to be alive makes a difference. It is, after all, only men who concerns themselves with morals or ethical questions.

It seems like one is trying to make the moral debate, that applies to persons, to humans, and has been carried out through the centuries, applicable to things. But this falls flat if the hypothesis, what one holds as true, is based on incorrect premises. An algorithm, a computer, a robot does not emanate from an evolutionary process. It does not have a wish or a will to live; no instinct to survive. And we cannot demand obligations from a tree, a stone, a dog, or a mobile phone, even if we, as humans, might have obligations to other living (and perhaps also dead) things.

To claim that a stone or a robot has a *self* – that it feels, giggles, have opinions, can suffer, or has an instinct to survive – is absurd. Without self-awareness or self-consciousness *per se*, there can be no autonomy. A "self" that can be a "we" (people) or an "I" (person), depending on the context, is crucial for autonomy. The "artificial moral agent" does not have a "self", no self-awareness, no self-consciousness, *per se* and thus, consequently, no moral agency.

It appears to be a recurring theme in robotics and AI to believe that something that appears to have some properties does in fact have them [2], much like Vaucanson's mechanical duck that appeared to have a digestive system (Fig. 1). This goes back to the Turing test [11] according to which a machine would be considered actually intelligent if it appears to be to an independent investigator. Similarly, robotic systems are described as moral agents if they appear to make moral decisions. We believe that this view on intelligence as well as autonomy and morality is not only wrong but also leads to research being directed in a direction that is less fruitful. Even worse, it contains the risk of causing policy makers to think of robotic systems as truly autonomous and be caught up on problems such as robot rights that would better stay in the realm of science fiction.

5 CONCLUSION

We have argued that artificial moral agency is not possible as the concept of autonomy is not applicable within the AMA context.. Does that mean that the study of moral robots or ethical AI is meaningless? Fortunately, this is not the case. However, it is important to be clear about who is the moral agent, who has agency and who is autonomous in the real sense of the words. It is not the machine itself, but rather its designer [5], who in fact is the *self*-governing agent calling the shots and hence accountable for its actions. [12]. Any meaningful approach to "artificial moral agents" must thus approach how machines can be designed in such a way that they behave *as if* they were moral, reflecting the intentions of the moral agents that designed them. Consequently, the responsibility for the actions of an artificial system lies clearly with its creator and the designers of whatever regulatory framework that constraints its programming and potential actions.

ACKNOWLEDGMENTS

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program – Humanities and Society (WASP-HS) funded by the Marianne and Marcus Wallenberg Foundation and the Marcus and Amalia Wallenberg Foundation.

REFERENCES

[1] Appelgren, Jessica, Beran, Tâm, Musco Eklund, Amanda, Hagström, Martin, 2022, Autonoma vapensystem - dagens debatt och en väg framåt Tekniska, legala och etiska aspekter https://foi.se/rapportsammanfattning?reportNo=FOI%20Memo%206953
[2] Balkenius, C. and Johansson, B., 2022. Almost Alive: Robots and Androids. Frontiers in Human Dynamics. Vol 4.
[3] Carnap, Rudolf, 1934. Logische Syntax der Sprache, Wien, Julius Springer, 1934.
[4] Carnap, Rudolf, 1950. Logical Foundations of Probability, London, Routledge and Kegan Paul, Ltd., 1950.

[5] Dignum, V., 2019. Responsible artificial intelligence: how to develop and use AI in a responsible way. Cham: Springer.
[6] von Gerber, Ylva, 2014. Autonomi – realitet eller ideal?, Media Tryck, Lund University.
[7] Kant, Immanuel, 1903. KGS, GMS (Grundlegung zur Metaphysik der Sitten), Bd IV.
[8] Oshana, Marina, Personal Autonomy in Society, Aldershot, Ashgate Publishing Ltd., 20.
[9] von Pufendorf, Samuel, 1991. On the Duty of Man and Citizen, Cambridge University Press.
[10] Rantakokko, Jouni, Appelgren, Jessica, Bengtsson, Kristofer Hagström, Martin, Jansson, Ove, Kraft, Karin, Kullander, Fredrik, Nygårds, Jonas, Näsström, Fredrik, Pettersson, Magnus, Rydell, Joakim, Woltjer,Rogier, 2020, Gemensamma teknikbehov inom obemannade och autonoma system https://www.foi.se/rest-api/report/FOI-R–5096–SE
[11] Turing, A.M., 2009. Computing machinery and intelligence. Springer.
[12] ”Personal autonomy”, 2018, Stanford Encyclopedia of Philosophy.