FILOSOFIJA. SOCIOLOGIJA
2024. T. 35. Nr. 1, p. 14–22

ISSN 0235-7186 eISSN 2424-4546
DOI: https://doi.org/10.6001/fil-soc.2024.35.1.3

Can't Bottom-up Artificial Moral Agents Make Moral Judgements?[1]

ROBERT JAMES M. BOYLES
Department of Philosophy, Southeast Asia Research Center and Hub, De La Salle University, Manila, Philippines
Email: robert.boyles@dlsu.edu.ph

This article examines if bottom-up artificial moral agents are capable of making genuine moral judgements, specifically in light of David Hume's is-ought problem. The latter underscores the notion that evaluative assertions could never be derived from purely factual propositions. Bottom-up technologies, on the other hand, are those designed via evolutionary, developmental, or learning techniques. In this paper, the nature of these systems is looked into with the aim of preliminarily assessing if there are good reasons to suspect that, on the foundational level, their moral reasoning capabilities are prone to the no-ought-from-is thesis. The main hypothesis of the present work is that, by conceptually analysing the notion of bottom-up artificial moral agents, it would be revealed that their seeming moral judgements do not have proper philosophical basis. For one, the said kinds of artifacts arrive at the understanding of ethically-relevant ideas by means of culling data or facts from the environment. Thus, in relation to the is-ought problem, it may be argued that, even if bottom-up systems seem *prima facie* capable of generating apparent moral judgments, such are actually absent of good moral grounding, if not empty of any ethical value.

Keywords: David Hume, is-ought problem, artificial intelligence ethics, machine ethics, artificial moral agent, bottom-up AMA

INTRODUCTION

Along with the various breakthroughs in the field of artificial intelligence (AI), the dangers that advanced technologies pose have also become hot-topic buttons, especially those that relate to ethics. These kinds of concerns are the focus of AI ethics, an area that looks into the different ethical issues related to artificially intelligent systems (Siau, Wang 2020). This research field may be further divided into two subareas, namely: (1) roboethics and (2) machine ethics.

Roboethics, short for 'robot ethics', inquires about the moral behaviours of human beings as they develop, employ and interact with AI systems, while also considering how these technologies could potentially affect the world (Siau, Wang 2020: 76). Considering that robots have already permeated several different aspects of people's lives (e.g. assistive or care robots, entertainment robots, military robots, etc.), those who have been doing research in

[1] The title pays homage to Midgley (2017), but the ideas in the present study differ from the said work.

the fields of computer science, philosophy, etc. have put forward numerous ways of addressing the specific ethical problems present in each of them (Tzafestas 2018: 3). At present, some typical topics within this subfield include the ethical worries of employing social assistive robots, wherein the proper assistance or care to vulnerable end-users, like the elderly, children with disorders, etc., is one of the key priorities (Boada et al. 2021). With regard to the potential negative impacts of AIs in business, on the other hand, there have also been discussions on whether or not automated machines are inherently evil, specifically from the perspective of management studies (Beltramini 2019). Yet another set of AI-related concerns centre on the possibility of creating artifacts that could exhibit ethical behaviour, if not also embody moral values, in order to mitigate certain potential safety threats (Siau, Wang 2020: 76). To this end, proponents of the subfield of machine ethics purport a novel way of addressing these kinds of AI risks.

Machine ethics, in general, focuses on developing AIs with moral precepts, which enables them to confront moral dilemmas that they might find themselves in (Anderson, Anderson 2007). For one, Wallach and Allen (2009) proffer that the realisation of these kinds of ethical technologies takes the form of artificial moral agents (AMAs). In brief, these systems are those that are considerate of what humans deem as valuable.

Here it must be pointed out that there are also those who question the very motivations in building AMAs. For instance, van Wynsberghe and Robbins (2019: 731) forewarn that, although AI systems and autonomous robots must be employed in morally salient situations, it does not necessarily follow that these technologies should, in fact, possess moral reasoning abilities. Others, meanwhile, have even challenged the entire machine ethics project from the outset (Yampolskiy, Fox 2013). But for Wallach and Allen (2009: 10), AMAs may be developed via the following strategies: (1) top-down or classical AI programming methods, (2) developmental or bottom-up tracks and (3) the hybrid of the said techniques. For purposes of the present study, the main focus would be on bottom-up AMAs.

Bottom-up methods presuppose that agents could develop a sense of morality on their own (Cervantes et al. 2020: 506). Thus, this method puts emphasis on designing environments for artifacts to explore or test their different courses of action, enabling them to learn and get rewarded whenever they exhibit morally meritorious behaviours (Wallach, Allen 2009: 80). Note that various issues have already been raised against bottom-up technologies in the past.

The concerns levelled against bottom-up systems include (1) the lack of pre-set safeguards that renders them quite risky to let out in the open (Wallach, Allen 2009: 114), (2) the difficulty of adopting social choice ethics (Baum 2020), etc. The simple aim of the present work is to preliminarily assess if bottom-up AMAs are also predisposed to D. Hume's is-ought problem, which espouses the idea that evaluative claims could never be derived from purely factual assertions (Restall, Russell 2010: 243). Specifically, it shows that these kinds of artifacts are at risk to the said problem, making them incapable of producing genuine moral judgements. The main hypothesis of this study is that, by conceptually analysing the notion of bottom-up AMAs, it would be revealed that their seeming moral judgements do not have a proper philosophical basis in view of Hume's is-ought.

To show that bottom-up AMAs are subject to the is-ought problem, the ensuing section recounts some studies which argue that certain intelligent systems are prone to its in-principle thesis. The succeeding section, on the other hand, briefly discusses the nature of bottom-up artifacts. The next section, then, provides an analysis to the conjecture that the apparent ethical judgments of bottom-up AMAs remain unaccounted for in light of the is-ought problem.

This is because bottom-up artifacts arrive at the understanding of ethically-relevant notions from culling data or facts from their situated environments. The final section ends with some concluding remarks.

IS-OUGHT PROBLEM AND INTELLIGENT SYSTEMS

In his *A Treatise on Human Nature* (1739/1964), Hume posits that moral conclusions, if not all evaluative assertions, could never be derived from factual premises alone. So, in light of what has come to be known as the 'is-ought problem', it could be said that no valid ought-conclusions may be inferred from purely is-premises (Frankena 2006: 49). Citing the said problem, some studies have further argued that certain types of intelligent systems are prone to its in-principle thesis.

Predicated on Boulanin and Verbruggen's (2017: 7–11) suggestion that the (1) sense, (2) decide, and (3) act capacities in artifacts must be integrated in order to operationalise the concept of autonomy, it has been contended that Autonomous Weapons Systems (AWS) do not have the ability to induce real moral judgements (Boyles 2021). This is because the data or facts collected by the sense part become the input for the decide and act parts. However, if the completion of the sense-decide-act cycle would require the inference of evaluative statements from factual ones, then any apparent moral judgement that AWS generate would have no ethical value. Any prescriptivist and descriptivist solutions to salvage the moral reasoning capacities of AWS do not appear to work as well considering the respective objections posed against them.

In general, prescriptivism permits the development of systems that consist of prescriptions, yet lacking ethical worth (Gensler 2011: 57). In fact, these types of systems include standard computer programs. Descriptivism, on the other hand, has received its fair share of objections as well (Joaquin 2013). For instance, Searle's (1964) offered stratagem still uses deducible ought-claims (i.e. so that one could derive evaluative assertions from descriptive propositions).

On a separate note, the objection cited above regarding AWS has also been re-appropriated for top-down AMAs (Boyles 2022). In brief, top-down systems are those created via programming ethical principles into them (Wallach, Allen 2009: 79–80). Hence, in light of these ethical precepts, which are stored in computer programs that serve as the artificial minds of artifacts, the actions of top-down systems are regulated to elicit moral behaviour. Hall (2011: 512) holds that the artificial minds of these AMAs generally have two components (viz. a world model and a utility function).

An artifact's world model (WM) is that which contains the objective knowledge of the world (Hall 2011: 512). This means that this part stores the disparate facts of the world that an artifact is attempting to model. The utility function (UF), in contrast, is the part that oversees the determination of preferences or proclivities among the different world states wherein specific goals are ranked (Hall 2011: 515). In short, the WM part provides top-down systems with the state of affairs of where they are presently situated, as well as the outcomes of their actions, while the UF part computes which action is most apt in a given context (Boyles 2022: 181). It must be pointed out, however, that this architecture cannot be relied on when it comes to accounting for the moral judgements of top-down artifacts, since there is no way of harmonising the UF with the WM part. This is because doing so requires the inference of ought-claims from factual premises.

NATURE OF BOTTOM-UP ARTIFACTS

The general notion behind bottom-up methods of realising AMAs is not to program artifacts with preset ethical principles; rather, to let these systems develop moral behaviour on their own (Wallach, Allen 2009: 97). Here it is assumed that ethical behaviour would emerge via a system-environment interaction. Some standard examples of design strategies that employ bottom-up methods include artificial life (ALife) and machine learning.

In general, ALife may be understood as life created by human beings instead of nature (Langton 1998: ix) as it intends to craft artificial systems via simulating evolution in virtual environments. Advocates of Alife maintain that, if blind, unintelligent evolutionary chance was able to produce human-level intelligent beings, then humans would also be capable of doing the same, if not at a much faster pace, as the said species are already intelligent from the onset (Chalmers 2010: 10). It must be underscored, however, that evolutionary processes may also be applied in designing physical artifacts, specifically in the discipline of evolutionary robotics (Pfeifer, Scheier 1999: 229).

Evolutionary robotics basically utilises evolutionary computation methods in the process of developing simulated and real-world robots (Vargas et al. 2014: ix). In relation to the motivation of creating AMAs via ALife, it may be said that evolutionary robotics similarly suggests that artificial systems could evolve into intelligent beings that possess moral discernment capacities.

On a separate note, another bottom-up technique is machine learning – a developmental strategy that operates on the assumption that moral notions could be incrementally learned through experience (Wallach, Allen 2009: 107–111). In brief, machine learning is the research area focused on providing computers with learning capabilities even without programming them explicitly (Park et al. 2018: 1).

In the context of developing ethical machines, Wallach and Allen (2009: 106–107) explain that, in general, bottom-up artifacts may need to undergo an educational process analogous to that of children. Hence, much like how children[2] acquire values, norms and proper behaviour through the process of, say, socialisation, advocates of this strategy propose to design systems with the capacity of integrating ethical sensitivity into themselves by means of some form of interaction.

In view of the characterisation of bottom-up AMAs above, Brooks' (1990) embodied AI approach may also be cited here. He explains that this AI approach follows the physical grounding hypothesis, which asserts that the representations of intelligent systems must be grounded on physical reality (Brooks 1990: 5). So, this type of design strategy also makes use of bottom-up methods.

BOTTOM-UP SYSTEMS AND HUME'S IS-OUGHT

In relation to the is-ought problem, bottom-up systems seem to run up against it as well. Given that the manner by which these AMAs acquire an understanding of ethically-relevant ideas is via collecting data from their interactions, it may be said that all these data would be fact-based information from their situated environments. Thus, these preconditions leave an opening for Hume's is-ought to be a worry, since evaluative statements cannot be derived from factual ones.

[2] Perhaps the origin of this idea may be attributed to Turing (1950), who proposed to simulate the minds of children, rather than those of adults. It must be pointed out, however, that this proposal was still based on symbol-processing methods, and machine learning has taken a completely distinct form since then (Wallach, Allen 2009: 219).

If one grants that bottom-up AMAs are, indeed, prone to the is-ought problem, it may be argued that this supposition relates to, if not fleshes out, other ideas or issues related to the said kinds of systems. Consider, for instance, Baum's (2020) worries concerning bottom-up artifacts. Specifically, he poses certain qualms as regards the proposal of adopting social choice ethics.

Social choice ethics subscribes to the suggestion of aggregating various ethical frameworks within a society in order to arrive at a correct theory of ethics (Baum 2020: 166). As Baum explains (2020: 167), implementing social choice ethics requires making three choices concerning the following: (1) standing, (2) measurement and (3) aggregation. Standing underscores the issue on whose ethics it should be taken into consideration by bottom-up AMAs (Baum 2020: 168). Baum (2020: 168–169) has already stressed various hitches related to this issue, like whether to include the ethical notions of children, non-human animals and antisocial psychopaths, to name a few. In a way, this issue somehow parallels Daley's (2021) idea that, instead of creating artificial general intelligence (AGI)[3] that is human-friendly, it might be better to foster impartial moral artifacts.

The issue of measurement, on the other hand, basically relates to the predicament on which procedure should be employed in obtaining the various values from all members of selected groups (Baum 2020: 167). Note here that there are different ways by which the ethical beliefs of concerned individuals may be collected, and it does not help that humans beings do not really have a universal set of consistent moral values (Baum 2020: 170).

Meanwhile, the third choice concerns the issue of aggregation. Baum (2020: 172) explains that once all the measurements have been completed (i.e. regardless which entities have standing and how their ethical notions were collected), the last step would be to aggregate different ethical ideas to come up with one social ethics framework. However, Baum (2020: 173) contends that aggregation is not really that straightforward, especially if non-humans, like bacteria, are granted standing as this may result in the non-anthropocentric ethics that bottom-up AMAs must consider due to the large number of non-humans in the planet (Baum 2020: 173).

In a way, Baum's (2020) issues on standing, measurement and aggregation relate with Hume's is-ought. For one, in terms of standing, arriving at the decision on whose ethical views should be considered presents worries for bottom-up systems. Consider, for instance, a bottom-up AMA that has the option of deciding which moral entities have standing.

For a bottom-up artifact to decide which entities deserve moral status, it seems that the only way that it could accomplish this task is if it possesses certain ways of assessing and, afterwards, coming up with an evaluative judgment that some beings are worthy of moral consideration, while others are not. However, the lone way by which bottom-up AMAs could acquire an understanding of morally-relevant notions is by culling facts or data. The problem with this is that there seems to be no reasonable grounds for any evaluative claim that the said AMAs would generate. So, it seems that a resolution to Hume's is-ought is needed first to overcome this hurdle, but, unfortunately, there is no such thing at present.

Furthermore, with regard to the problem of measurement, this disputably boils down to deciding upon a particular procedure that could be used to obtain values from different concerned groups (Baum 2020: 167). Considering that there are numerous ways by which the ethical views of individuals could be collected, like via voting in democracies, buying and

[3] Goertzel (2014: 1–2) clarifies that, unlike narrow AIs that simulate specific intelligent behaviours, AGIs pertain to artificial systems with broad generalisation capabilities.

selling in capitalism, etc. (Baum 2020: 170), it seems that what is needed for bottom-up AMAs to decide which is most apt among the many options is an impartial procedure that would definitively show that one is more preferable than the other.

Arguably, the same may also be said whenever these technologies are tasked to address the aggregation problem. This is because the issue on how to best collate the disparate ethical notions, so that it may be recast into one societal ethics view, could only be solved with pertinent evaluative decisions. Unfortunately, there seems to be no clear impartial procedure in sight to help address the measurement and aggregation issues due to Hume's is-ought.

In view of these discussions, one could further cash out Baum's (2020: 174) stance that the proposal to employ social choice ethics only invites open questions that deal with aggregation, measurement and standing, making it difficult for AI designers not to impose their personal convictions in addressing these issues; social choice ethics only mitigates the said impositions. For AMAs to function properly, morally speaking, it seems that they need to be programmed with apt ethical theories. However, doing so would only bring forth issues related to prescriptivism that have been mentioned earlier.

If bottom-up AMA design strategies do not subscribe to the proposal of pre-programming ethical theories into AI systems, then they only have descriptivist solutions at their disposal. But as discussed above, descriptivism is also not worry free. Perhaps a couple of things could be further said about bottom-up strategies.

First, if bottom-up AMAs were to learn, blend, and adapt to the accepted norms, customs and practices of the other entities around it, then this somehow overturns the general idea regarding ethics that it concerns 'what ought to be the case' and not 'what is the case'. Considerably, this alone is a valid reason not to consider bottom-up tracks.

Second, let us assume that the bottom-up strategy is indeed correct. Notice that, even if one accepts this view, it may still be argued that the issue of justifying why a particular ethical theory should be favoured over another has not been resolved yet. So, for instance, why should a bottom-up artifact choose Confucian ethics over, say, utilitarianism, care ethics, or any other normative ethical theory?

Unfortunately, there appears to be no definitive solution to the long-standing philosophical issue on which ethical theory is the correct one. While others, like Shaw et al. (2018), proffer suggestions that could somehow sidestep this issue,[4] it may be said that things still ultimately boil down to preferring one theory of ethics among the numerous live candidates. So, in the context of bottom-up artifacts, there seems to be no clear justification on why these AMAs should favour an ethical precept over another.

Conceivably, bottom-up technology could encounter a multitude of ethical theories as it interacts with other agents and its environment. For one, some might argue that arriving at certain values or moral standards is possible as one interacts with reality. In a way, this line of reasoning is reminiscent of theories that appeal to some natural law present in the world.

A central thesis of natural law theory (NLT) is that the grounds of morality are based on the nature of the world, if not also on human nature itself (Himma 2001). Himma (2001) further clarifies that all variants of NLT follow the thesis that certain moral laws or standards do not depend on human convention and agreement.

[4] In relation to bottom-up technologies, Shaw et al. (2018) claim that the possession of certain meta-moral qualities – that an artifact's learned moral principles must be (1) robust, (2) consistent, (3) universal and (4) simple – enables any artificial system to adapt to particular moral contexts, making them appropriate moral agents comparable to humans.

So, under NLT, it initially appears that it is possible for bottom-up systems to stumble upon moral standards or values in nature. However, it must be pointed out that natural law theories are not that worry free. For one, though somewhat debatable, Hume's is-ought problem has also been cited as a potential worry against the said theory (Moschella, George 2015: 320).

One reason why the is-ought problem has been cited as a criticism against NLT is that the latter reputedly conflates (1) 'what is the case' with (2) 'what ought to be the case' (Evangelista, Mabaquiao 2020: 71). So, at the very least, the idea that it is possible to arrive at specific moral notions by encountering some natural laws does not seem fully absent of issues.

Nevertheless, even if one considers natural law theory as a possible means to create bottom-up AMAs, a further issue that has not been settled yet is which among its several variants is the most viable one. Himma (2021) holds that there are different theories that may be classified to fall under the natural law type. Also, Evangelista and Mabaquiao (2020: 66) even point out that earlier variants of the said theory possess similarities with the divine command theory,[5] since these older versions attribute the natural law to God.

If there are numerous types of natural law theories, it somehow begs the question which among these bottom-up AMAs should subscribe to. This is because there appears to be no clear way of preferring one version of this ethical theory over another. Thus, much like the issue raised against top-down AMA methods – that randomly choosing an ethical theory leads to a problem of circularity or an arbitrary assignment of values (Boyles 2022: 183), it may be said that bottom-up systems fall into the same trap as well.

CONCLUSIONS

In this article, it was argued that there are good reasons to suspect that bottom-up AMAs are also prone to the no-ought-from-is thesis. So, even if these artifacts seem *prima facie* capable of coming up with moral judgments, these are actually absent of good moral grounding, if not empty of any ethical value.

If one grants that bottom-up artifacts are, in fact, prone to the is-ought problem, perhaps other concepts, ideas, or issues related to this method could be revisited. For instance, it has been said that any design strategy for bottom-up systems must consider the frame-of-reference problem, which centres on conceptualising the connection between the following: the observer, subject, designer, environment and the artifact (Pfeifer, Scheier 1999: 650). So, it is quite intriguing how Hume's is-ought could further cash out, if not affect, bottom-up strategies, especially that certain considerations relating to artifacts, AI designers, and the environment, among others, must all be taken into account.

Moreover, considering that specific arguments have already been put forward against top-down and bottom-up systems, perhaps the possibility on how the is-ought problem could affect hybrid AMAs must be further explored. Note that hybrid technologies try to fuse certain attributes from both top-down techniques and bottom-up strategies (Wallach, Allen 2009: 117–118). So, research on these kinds of artifacts might shed some light if there still remains a feasible, practicable way to develop moral machines.

Received 12 August 2023
Accepted 23 October 2023

[5] The divine command theory follows the idea that actions are only considered morally correct if they align with the will of God (Evangelista, Mabaquiao 2020: 61).

References

1. Anderson, M.; Anderson, S. L. 2011. *Machine Ethics*. New York: Cambridge University Press.
2. Baum, S. 2020. 'Social Choice Ethics in Artificial Intelligence', *AI & Society* 35: 165–176. DOI: 10.1007/s00146-017-0760-1.
3. Beltramini, E. 2019. 'Evil and Roboethics in Management Studies', *AI & Society* 34: 921–929. DOI: 10.1007/s00146-017-0772-x.
4. Boada, J. P.; Maestre, B. R.; Genís, C. T. 2021. 'The Ethical Issues of Social Assistive Robotics: A Critical Literature Review', *Technology in Society* 67(101726): 1–13. DOI: 10.1016/j.techsoc.2021.101726.
5. Boulanin, V.; Verbruggen, M. 2017. *Mapping the Development of Autonomy in Weapon Systems*. Solna: Stockholm International Peace Research Institute.
6. Boyles, R. J. M. 2021. 'Hume's Law as Another Philosophical Problem for Autonomous Weapons Systems', *Journal of Military Ethics* 20(2): 113–128. DOI: 10.1080/15027570.2021.1987643.
7. Boyles, R. J. M. 2022. 'Extending the Is-ought Problem to Top-down Artificial Moral Agents', *Symposion* 9(2): 171–189. DOI: 10.5840/symposion20229213.
8. Brooks, R. 1990. 'Elephants Don't Play Chess', *Robotics and Autonomous Systems* 6(1–2): 3–15.
9. Cervantes, J. A.; López, S.; Rodríguez, L. F.; Cervantes, S.; Cervantes, F.; Ramos, F. 2020. 'Artificial Moral Agents: A Survey of the Current Status', *Science and Engineering Ethics* 26: 501–532. DOI: 10.1007/s11948-019-00151-x.
10. Chalmers, D. 2010. 'The Singularity: A Philosophical Analysis', *Journal of Consciousness Studies* 17(9–10): 7–65.
11. Daley, K. 2021. 'Two Arguments Against Human-friendly AI', *AI and Ethics* 1: 435–444. DOI: 10.1007/s43681-021-00051.
12. Evangelista, F. J. N.; Mabaquiao, N. M. Jr. 2020. *Ethics: Theories and Applications*. Mandaluyong: Anvil Publishing, Inc.
13. Frankena, W. 2006. 'The Naturalistic Fallacy', in *Arguing About Metaethics*, eds. A. Fisher and S. Kirchin. New York: Routledge, 47–58.
14. Gensler, H. J. 2011. *Ethics: A Contemporary Introduction (Second Edition)*. New York: Routledge.
15. Goertzel, B. 2014. 'Artificial General Intelligence: Concept, State of the Art, and Future Prospects', *Journal of Artificial General Intelligence* 5(1): 1–46. DOI: 10.2478/jagi-2014-0001.
16. Hall, J. S. 2011. 'Ethics for Self-improving Machines', in *Machine Ethics*, eds. M. Anderson and S. L. Anderson. Cambridge: Cambridge University Press, 512–523.
17. Himma, K. E. 2001. 'Natural Law', in *Internet Encyclopedia of Philosophy*, eds. J. Fieser and B. Dowden. Available at: https://iep.utm.edu/natlaw/ (accessed 07.07.2023).
18. Hume, D. 1739/1964. *A Treatise of Human Nature*, ed. L. A. Selby-Bigge. Oxford: Clarendon Press.
19. Joaquin, J. J. 2013. 'John Searle and the Is-ought Problem', *Scientia: The International Journal on the Liberal Arts* 2(1): 53–66.
20. Langton, C. G. 1998. 'Editor's Introduction', in *Artificial Life: An Overview*, ed. C. G. Langton. Cambridge: MIT Press, ix-xi.
21. Midgley, M. 2017. *Can't We Make Moral Judgements?* London: Bloomsbury Academic.
22. Moschella, M.; George, R. 2015. 'Natural Law', in *International Encyclopedia of the Social & Behavioral Sciences (Second Edition)*, ed. J. D. Wright. Oxford: Elsevier Ltd, 320–324.
23. Park, C.; Took, C. C.; Seong, J. K. 2018. 'Machine Learning in Biomedical Engineering', *Biomedical Engineering Letters* 8: 1–3. DOI: 10.1007/s13534-018-0058-3.
24. Pfeifer, R.; Scheier, C. 1999. *Understanding Intelligence*. Cambridge: MIT Press.
25. Restall, G.; Russell, G. 2010. 'Barriers to Consequence', in *Hume on Is and Ought*, ed. C. R. Pigden. Basingstoke: Palgrave Macmillan, 243–259.
26. Searle, J. R. 1964. 'How to Derive "Ought" From "Is"', *The Philosophical Review* 73(1): 43–58. DOI: 10.2307/2183201.
27. Shaw, N. P.; Stöckel, A.; Orr, R. W.; Lidbetter, T. F.; Cohen, R. 2018. 'Towards Provably Moral AI Agents in Bottom-up Learning Frameworks', in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. New York: Association for Computing Machinery, 271–277. DOI: 10.1145/3278721.3278728.
28. Siau, K.; Wang, W. 2020. 'Artificial Intelligence (AI) Ethics: Ethics of AI and Ethical AI', *Journal of Database Management* 31(2): 74–87. DOI: 10.4018/JDM.2020040105.

29. Turing, A. M. 1950. 'Computing Machinery and Intelligence', *Mind* 59(236): 433–460. DOI: 10.1093/mind/LIX.236.433.
30. Tzafestas, S. G. 2018. 'Roboethics: Fundamental Concepts and Future Prospects', *Information* 9(6): 1–25. DOI: 10.3390/info9060148.
31. van Wynsberghe, A.; Robbins, S. 2019. 'Critiquing the Reasons for Making Artificial Moral Agents', *Science and Engineering Ethics* 25: 719–735. DOI: 10.1007/s11948-018-0030-8.
32. Vargas, P. A.; Di Paolo, E. A.; Harvey, I.; Husbands, P. (eds.). 2014. *The Horizons of Evolutionary Robotics*. Cambridge: MIT Press.
33. Wallach, W.; Allen, C. 2009. *Moral Machines: Teaching Robots Right from Wrong*. New York: Oxford University Press.
34. Yampolskiy, R.; Fox, J. 2013. 'Safety Engineering for Artificial General Intelligence', *Topoi* 32: 217–226. DOI: 10.1007/s11245-012-9128-9.

ROBERT JAMES M. BOYLES

Ar dirbtiniai moralės agentai negali priimti moralinių sprendimų „iš apačios į viršų“ principu?

Santrauka

Straipsnyje nagrinėjama, ar dirbtiniai moraliniai agentai „iš apačios į viršų“ principu gali priimti tikrus moralinius sprendimus, atsižvelgiant į Davido Humeʼo problemą. Pastarasis pabrėžia mintį, kad vertinamosios išvados niekada negali būti daromos remiantis vien faktinėmis prielaidomis. Kita vertus, technologijos „iš apačios į viršų“ yra sukurtos naudojant evoliucinę, mokymosi ar plėtojimo metodiką. Šiame darbe nagrinėjamas šių sistemų pobūdis, siekiant preliminariai įvertinti jų moralines samprotavimo galimybes. Pirma, minėtų rūšių artefaktai leidžia suprasti etiškai svarbias idėjas atsižvelgiant į iš aplinkos gaunamus duomenis ar faktus. Taigi galima teigti, kad net jei sistemos „iš apačios į viršų“ atrodo *prima facie* galinčios priimti akivaizdžius moralinius sprendimus, jos iš tikrųjų neturi gero moralinio pagrindo, vien tuščią etinę vertę.

Raktažodžiai: Davidas Humeʼas, turėjimo būti problema, dirbtinio intelekto etika, mašinų etika, dirbtinis moralės agentas, moralė „iš apačios į viršų“ principu