Extending the Is-ought Problem to Top-down Artificial Moral Agents

Robert James M. Boyles

Abstract: This paper further cashes out the notion that particular types of intelligent systems are susceptible to the is-ought problem, which espouses the thesis that no evaluative conclusions may be inferred from factual premises alone. Specifically, it focuses on top-down artificial moral agents, providing ancillary support to the view that these kinds of artifacts are not capable of producing genuine moral judgements. Such is the case given that machines built via the classical programming approach are always composed of two parts, namely: a world model and utility function. In principle, any attempt to bridge the gap between these two would fail, since their reconciliation necessitates for the derivation of evaluative claims from factual premises.

Keywords: artificial moral agent, David Hume, is-ought problem, machine ethics, top-down AMA.

Introduction

In *Hume's Law as another Philosophical Problem for Autonomous Weapons Systems*, Boyles (2021) argues that military-grade autonomous weaponry are susceptible to the is-ought problem. Autonomous Weapons Systems (AWS), on the one hand, may be defined as machines that, "once activated, can select and engage targets without further intervention by a human operator." (Department of Defense 2012, 13) The is-ought problem, on the other hand, is a logical problem commonly attributed to David Hume (Gunkel 2018, 88). Also known as "Hume's Law" (Hare 1952) or "Hume's Guillotine" (Black 1964, 166), the said problem espouses the thesis that evaluative conclusions may never be inferred from factual premises alone (Restall and Russell 2010, 243).

For Boyles, grounding the purported moral judgements of AWS appear to be intractable in light of the is-ought problem, since these artifacts make use of factual data from their environments to carry out specific actions. Supposedly, the process involved in such requires the derivation of evaluative statements from a set of purely factual ones. He further explains that, "[i]f there remains a fundamental difference between the actions or behaviors of … AWS from their human counterparts – the latter being capable of arriving at genuine moral judgments, then one should be cautious in naively trusting the apparent ethical decisions of the former." (2021, 126)

Though certain distinctions between human beings and AWS were pointed out in the said article, particularly in terms of ethical decision-making and their moral standing, no supplementary explanations were offered to account for why

the latter are unable to arrive at genuine ethical decisions (i.e., beyond just appealing to Hume's no-ought-from-is doctrine). For one, the notion that there are different ways of designing intelligent machines was not considered. Thus, this paper looks into how the is-ought problem relates to the internal design of artifacts, specifically focusing on top-down artificial moral agents.

An artificial moral agent (AMA) is commonly defined as "an artificial autonomous agent that has moral value, rights and/or responsibilities." (Sullins 2009, 208) In order to realize this type of machine, the top-down method of designing AMAs subscribes to the view that moral principles may be directly encoded into its internal program (Wallach and Allen 2009, 83-97). By doing so, an artifact's actions and behaviors would, thus, be regulated by the said precepts. However, note that several challenges have also been put forward against the top-down AMA track.

Apart from the difficulty of translating and applying abstract moral principles to specific, actual situations, Misselhorn (2018, 165) holds that top-down AMAs are predisposed to the frame problem. In a nutshell, the latter problem concerns logic-based systems, specifically on how to represent the effects of their actions. Supposedly, identifying the particular conditions in modeled environments that have been affected by the actions of top-down systems pose certain hurdles, since there is an assumption that all other conditions stay fixed. The said assumption, however, is still unfounded, citing the unresolved issue of aptly sorting out all relevant information from the irrelevant ones. Dennett (1984, 130), for one, asserts that this issue eventually results in a "deep epistemological problem."

Allen, Smit, and Wallach (2005), on the other hand, explain that a major concern with top-down AMAs is that the ethical rules or commandments programmed into them often conflict with one another, especially once the said systems encounter real-world ethical dilemmas. They further maintain that such conflicts result in computationally intractable situations, and most all rule-based systems do not offer concrete ways to resolve them.

The primary aim of this paper is to raise another challenge against top-down AMAs, which is the is-ought problem. Following Boyles' (2021) use of the latter to proffer certain foundational worries against AWS, the present work extends the said strategy to top-down AMAs. In this paper, it is argued that the said systems are also prone to Hume's Law, since machines built via the said method are always composed of two parts, namely: a world model and utility function (Hall 2011, 512). In principle, any attempt to bridge these two parts would fail, since reconciling them would be the same as deriving evaluative statements from a set of factual ones. Furthermore, note that, although Hall (2011, 514) briefly mentions that the world model and utility function of classical systems are separated by 'Hume's is-ought guillotine,' no extensive explanation for such has been provided. Hence, this article also seeks to offer a more detailed analysis of the said idea.

To contend that top-down AMAs are susceptible to the no-ought-from-is doctrine, the following section initially revisits Hume's original discussion of the said problem, while also citing the two views that resulted from it (i.e., moral descriptivism and moral non-descriptivism). The main objective of this part is to highlight the idea that there seems to be no foolproof solution to the is-ought problem today. The subsequent section, meanwhile, provides a summary on the view that AWS are susceptible to the is-ought problem. Moreover, in order to further ground this notion, specifically on how it relates to top-down AMAs, the next section looks into the nature of classically programmed artifacts. In this section, it is shown that the reason why top-down technologies are unable to produce genuine moral judgements is that the world model and utility function embedded in them, in principle, cannot really be reconciled. The final section of this paper provides a few concluding remarks.

Hume's No-ought-from-is Doctrine

As mentioned earlier, the origins of the no-ought-from-is doctrine may be traced to Hume (Gunkel 2018). In his *Treatise on Human Nature*, Hume states that:

> In every system of morality, which I have hitherto met with, I have always remark'd, that the author proceeds for some time in the ordinary way of reasoning, and establishes the being of a God, or makes observations concerning human affairs; when of a sudden I am surpriz'd to find, that instead of the usual copulations of propositions, is, and is not, I meet with no proposition that is not connected with an ought, or an ought not. This change is imperceptible; but is, however, of the last consequence. For as this ought, or ought not, expresses some new relation or affirmation, 'tis necessary that it shou'd be observ'd and explain'd; and at the same time that a reason should be given, for what seems altogether inconceivable, how this new relation can be a deduction from others, which are entirely different from it. But as authors do not commonly use this precaution, I shall presume to recommend it to the readers; and am persuaded, that this small attention wou'd subvert all the vulgar systems of morality, and let us see, that the distinction of vice and virtue is not founded merely on the relations of objects, nor is perceiv'd by reason. (1739/1964, 243-244)

It could be inferred from the above quote that the is-ought problem centers on the viability of providing factual justifications for moral judgments. For Hume, no legitimate deduction[1] may be made from an 'is' to an 'ought.' (Brown 2008, 229)

Following Hume's line of thinking, it may be said that in any argument that is composed of (1) a set of purely factual premises and (2) a normative conclusion (i.e., derived from the said set), the normative judgment found in the latter would not logically follow from the factual assertions found in the series of is-statements.[2] Snare (1992, 84-85) also explains that the is-ought problem may be

[1] Hume's usage of the term 'deduction' has resulted in a debate on what he truly meant by this. See Schurz (1997, 2). The present work is neutral about this issue.

[2] The is-ought problem may also be further related to Hume's view on ethics. See Cohon (2010).

likened to an in-principle thesis – on the foundational level, evaluative conclusions may never be arrived at as long as one deduces them from factual premises alone. To further understand this, one may cite the logical relationship between the nature of 'oughts' and 'isses.'

One way to account for is-statements is to think of them as the content of assertions or descriptive statements. Note that the latter are truth-evaluable expressions, and a standard example of such are declarative sentences. [3] Conversely, ought-statements operate more like imperatives or prescriptions of actions.[4] So, in contrast to is-statements, ought-statements cannot be evaluated as either true or false, since they do not state facts.[5]

Since ought-statements generally pertain to a moral obligation or a norm of conduct (i.e., in the context of moral judgments), they naturally relate to the notion of ethical value. This is because all moral systems normally presuppose a close link between moral obligations and ethical values (Schurz 1997, 1). The idea behind this is that what is deemed as ethically good ought to be, must be, or needs to be done, which demonstrates the obligatory aspect, if not the imperative force, of an ought claim. So, ought-statements function more as prescriptions of actions, and they contrast with is-statements that bear truth claims.

Ever since Hume pointed out the apparent logical invalidity of deducing 'oughts' from 'isses,' a number of philosophers have proffered different ways to address the no-ought-from-is doctrine. Among the numerous replies to the latter include that of Hare (1952) and Searle (1964), which could be treated as standard representatives of the universal prescriptivist and moral descriptivist views, respectively.

Prescribing Ought-statements

As discussed by Boyles (2021, 118), universal prescriptivism adheres to the notion that ought claims are a kind of prescription or imperative (Gensler 2011, 56). In *The Language of Morals*, Hare contends that the "language of morals is one sort of prescriptive language." (1952, 1) So, for prescriptivists, imperatives do not really state facts, which further means that these can neither be true nor false.

Prescriptions operate like commands, basically directing someone to do or perform something. For prescriptivists, ought-statements are just expressions of impartial desires of how one should live, act, or behave (Gensler 2011, 57). In a sense, this demonstrates the normative aspect of imperatives, which also accounts for why prescriptivists believe that prescriptions are universalizable.

---

[3] As also noted by Boyles (2021, 126), not all declarative sentences are is-statements. Furthermore, not all kinds of assertions can be judged as straightforwardly true or false.

[4] Though there is an ongoing debate as to whether or not ought-statements are, in fact, truth-evaluable, this issue is well beyond the scope of this work.

[5] Several philosophers, like Hume, have argued that moral judgments do not, strictly speaking, state facts, which makes them non-truth-evaluable also. For a brief summary of Hume's view regarding this issue, see Cohon (2010).

Ought-statements, for prescriptivists, are universalizable prescriptions (Gensler 2011, 58). If one comes up with an 'ought,' then this does not merely equate to the act of making a simple prescription. Putting forward an ought-statement expresses one's utmost desire that an action be context-invariant (i.e., the prescribed course of action ought to be followed in all analogous cases). The said idea is also embodied in the logical rules for 'oughts', which are as follows:

> U. To be logically consistent, we must make similar evaluations about similar cases.
>
> P. To be logically consistent, we must keep our moral beliefs in harmony with how we live and want others to live. (Gensler 2011, 58)

Logical rules U and P are consistency rules for action (Gensler 2011, 58). Logical rule U dictates the iteration of a specific action in all analogous cases. This means that whenever an ought-statement is made, we should treat its content as context-invariant. On the other hand, logical rule P maintains that ought judgments are, in fact, imperatives, which further entails that an ought judgment is somehow devoid of its obligatory function (i.e., in the moral sense). Gensler further notes that, "[i]nstead, they tell us what we must do if we're to be logically consistent in our moral beliefs." (2011, 58) This highlights the notion that an ought-statement becomes a logical test for the consistency of our moral judgments and beliefs.

As for the issue of universal prescriptivism being a rational ethical system, even though it regards ought-statements as non-truth bearing claims, Gensler (2011, 57) maintains that it is quite possible to construct a system comprised of a set of prescriptions that does not necessarily equate to a moral system. He further asserts that among the said systems include cookbooks, the laws of a particular country, and complex computer programs, to name a few.

Gensler (2011, 61-63), however, holds that prescriptivism may further lead to the denial of the possibility of attaining moral knowledge and truths. If ought-statements are just universalizable prescriptions, then moral judgments would only be a test of consistency of prescribed actions. In a way, this highlights the idea that no further moral truths may be attained given that, for there to be further moral truths, new information must be accounted for.

In relation to the is-ought problem, prescriptivists readily affirm such. As discussed earlier, Hare was even famous for coining the phrase "Hume's Law." (Cohon 2010) Considering that they accept Hume's Law, the only recourse for prescriptivists is to show that, in all moral arguments, there is an underlying evaluative statement hidden, if not assumed, alongside the relevant factual premises (Joaquin 2012, 55-56). So, with regard to the attempt of deriving an ought-conclusion from a series of is-statements, it appears that prescriptivism does not yield a tenable solution to the is-ought problem at present. For prescriptivists, the said problem is, in fact, a live one.

Moral Talk as Factual Claims

Proponents of moral descriptivism argue that ethical language is best treated as an attempt to describe something in the world (Fisher and Kirchin 2006, 3). As per Boyles (2021, 118-120), ethical statements are, for them, simply reducible to claims about facts (i.e., under a certain set of conditions). Thus, moral statements could also be evaluated, like descriptive statements, based on their truth-value. To further grasp the descriptivist model, specifically in the context of how it deals with the no-ought-from-is thesis, consider Searle's (1964) view regarding this matter.[6]

To address the is-ought problem, Searle first challenges the notion that facts are entirely distinct from values (Joaquin 2012, 56-57). He demonstrates this by providing the following counterexample:

1) Jones uttered the words 'I hereby promise to pay you, Smith, five dollars.'
2) Jones promised to pay Smith five dollars.
3) Jones placed himself under (undertook) an obligation to pay Smith five dollars.
4) Jones is under an obligation to pay Smith five dollars.
5) Jones ought to pay Smith five dollars. (Searle 1964, 44)

As to how one may derive the evaluative claim, "Jones ought to pay Smith five dollars," from the said set of factual statements, Searle (1964, 44-49) explains that this could simply be done by adding 'empirical assumptions, tautologies, and descriptions of word usage' to the given premises (1964, 48).

Moreover, by using definitional connections between 'promise,' 'obligate,' and 'ought,' as well as including a *ceteris paribus* clause to eliminate possible contrary considerations, Searle claims that the move from premises (2) to (5) seems "relatively easy." (1964, 49) Recall that moral descriptivists, like Searle, hold that ought-statements could be reduced into fact-stating propositions under a given set of conditions. For Searle, he is able to specify such conditions by employing the concept of institutional facts (Joaquin 2012, 65).

In a nutshell, institutional facts are specific kinds of facts that depend on human convention and agreement (Searle 1995, 29). In contrast to brute facts (e.g., Water is H_20 in this world), which exist independently of human agreement (Searle 1995, 27), institutional facts presuppose human institutions, since they are responsible for creating the system of constitutive rules – those that not only regulate, but also ensure the rules' very existence.

Searle employs the idea of institutional facts to specify the scope or conditions that allows for the translation of ought-statements into descriptive

[6] It should be pointed out here that Searle was actually responding to the more modern formulation of the is-ought problem, which was put forward by philosophers such as Hare (Joaquin 2012, 56).

ones. So, for instance, it may be argued that the statement "Jones ought to pay Smith five dollars" may be considered true only if one presumes that there exists a human or social institution that states that such is the case. As ingenious as Searle's strategy may seem, however, a couple of concerns may be raised against it (Boyles 2021, 119).

First, Searle's maneuver to infer an 'ought' from an 'is' is not without problems. In fact, he tries to anticipate many objections to this.[7] For example, his supplementary premise to bridge (4) and (5), "(4a) Other things are equal," may be rendered to the following statement: "(4a) All those who are under an obligation, *ceteris paribus*, ought to fulfill that obligation." Note that this seemingly equal formulation could be treated as an (implicit) ought premise, which would put into question Searle's primary goal of deducing an 'ought' from purely is-statements.[8]

Second, even if one grants that Searle is successful in deriving an 'ought' from a set of 'isses,' it must be pointed out that such strategy appears to only work for a very particular case, specifically, to promise making (Boyles 2021, 119). Given the seemingly limited scope of the said strategy, it may be argued that it is really not that fruitful in light of the endeavor of developing autonomous machines.

With regard to the attempt of deriving an 'ought' from a series of is-statements, it appears that universal prescriptivism and moral descriptivism do not yield, as of yet, a tenable solution to this issue. As mentioned earlier, both strategies are not ironclad. Furthermore, note that the worries against these two positions have also been related to the prospect of creating autonomous weapons systems.

Hume's Guillotine, Autonomous Weaponry and Moral Judgments

As regards the development and deployment of AWS, many have already called for more research into the ethical concerns and dangers surrounding these types of technologies (Sharkey 2010; Sparrow 2016; Boyles, Dacela, Evangelista, and Rodriguez 2022, 192). Boyles (2021), for one, proffers that such are prone to the no-ought-from-is doctrine.

To establish that AWS are susceptible to Hume's is-ought, Boyles (2021, 115) first cites Boulanin and Verbruggen's (2017, 7-11) idea that the concept of autonomy in artifacts is basically operationalized by integrating three fundamental capabilities (i.e., sense, decide, and act). He further explains that the sense capacity, mainly composed of built-in sensors and sensing software, is utilized by AWS to perceive the environment (Boyles 2021, 120). So, all data generated by this capacity are particular facts about the context an AWS is presently situated, and these facts become input for the decide capability. After a

[7] See Searle 1964, 49-52.

[8] This worry against Searle's strategy may be generally classified to fall under the "objections against the *ceteris paribus* clause." (Joaquin 2012, 59-63)

specific decision has been reached, an autonomous system, then, implements a set of actions. So, in the sense-decide-act cycle, data input is critical, since the judgements and actions of AWS depend on the information gathered by its sensors and sensing software.

As per Boyles, the actions of AWS that subscribe to the sense-decide-act cycle cannot be trusted, morally speaking, since there is no direct way of reconciling their sense capacity with the decide part (2021, 120). This is because doing so requires inferring evaluative statements from a set of factual ones, which entails that all of their supposed moral judgments are of no ethical worth. He further maintains that attempts to get around the said issue, like encoding AWS with ethical precepts, would not work as such strategy is similar to the prescriptivist solution to the is-ought problem.

Recall that, for universal prescriptivists, adding an evaluative statement to a series of descriptive ones would circumvent the is-ought problem. This is because prescriptivists claim that this additional statement would enable the deduction of an evaluative conclusion from the said total set of premises. So, in the context of AWS, a machine would be capable of generating a moral judgment, *prima facie*, as long as it is pre-programmed with certain ethical principles. However, Boyles (2021, 121) forewarns that this might be deceiving.

Citing Gensler (2011, 57), Boyles explains that it is possible to develop a model consisting of a set of prescriptions, but is devoid of any ethical value. Standard examples of these kinds of systems include the laws of different nations, computer programs, and the like. So, with regard to prescriptivist-based AWS, it might be the case that their apparent ethical judgments do not really have any moral worth.

Furthermore, in terms of developing descriptivist-based AWS, Boyles claims that this strategy is problematic as well. He states that:

> Programming the decide capacity of an AWS so that it could decide which particular theory is the most relevant in a specific situation somehow issues in the frame problem. As per the said problem, tagging a theory as the most relevant one necessitates for an artifact to consider infinitely finite facts in a given situation... Note also that the task of ascertaining which ethical theory is most appropriate in a given context parallels the descriptivist solution to Hume's Law. This is because descriptivists claim that moral statements are reducible to factual ones under specific conditions, perhaps arguing that a moral theory may be considered most relevant in a given situation when it addresses the issue at hand. (2021, 124)

Since descriptivist solutions to the is-ought problem still make use of implicit ought-claims (i.e., to generate evaluative conclusions from descriptive statements), Boyles (2021, 124) notes that proposals of this type also do not stand on firm grounds. Additionally, recall that descriptivist solutions, like that of Searle (1964), appear to only work in very limited circumstances. If the objective is to create robust and autonomous machines, then making use of the descriptivist design strategy appears to be futile.

Considering the issues of both prescriptivist and descriptivist solutions to the is-ought problem (i.e., in relation to the endeavor of developing AWS), Boyles (2021) claims that there is a fundamental difference between machines and their human counterparts, especially in terms of their moral standing and ethical decision-making. He maintains that, at present, we are certain that humans could enact genuine moral judgments, but it is not that clear if AWS, or any other kinds of machines, could actually do the same (2021, 123).

In light of the said standing issues against prescriptivism and descriptivism, Boyles (2021) holds that grounding the apparent moral judgments of AWS is problematic. If the ultimate goal of creating these types of machines is for them to become autonomous agents, then there is no certainty that they would always come up with ethically-sound actions (i.e., in the context of real-life moral dilemmas). This is because the actions of AWS largely depend on the data gathered by their sensors and sensing software, and the process of generating the former from the latter parallels the move of deriving evaluative conclusions from a set of purely factual premises.

Note that Boyles (2021), however, does not really provide any further explanation as to why AWS are not capable of producing genuine ethical judgments (i.e., beyond just appealing to Hume's is-ought). For one, the notion that there are different types of artificial moral agents, which are intelligent artifacts that have the capacity to enact moral decisions (Cervantes et al. 2020), was not considered.

For Wallach and Allen, with regard to developing AMAs, three design strategies could be employed, namely: the top-down, bottom-up, and hybrid approaches. As explained earlier, the top-down AMA method adheres to the idea that moral principles may be directly encoded into an artifact's internal program (2009, 83-97). By programming certain ethical precepts from the get-go, the actions and behaviors of these kinds of machines would be regulated, morally speaking.

Bottom-up AMA approaches, on the other hand, are those that employ evolutionary, learning, or developmental methodologies (Wallach and Allen 2009, 80). This track focuses on creating environments where artifacts could consider and enact different courses of action, while also learning from them in the process. Note that bottom-up AMAs are given set rewards whenever they exhibit praiseworthy behavior. In contrast, hybrid AMAs integrate the design principles of both top-down and bottom-up options, and one way to supposedly realize this method is by using Aristotelian virtue ethics (Wallach and Allen 2009, 10). Wallach and Allen note that "[v]irtues are a hybrid between top-down and bottom-up approaches, in that the virtues themselves can be explicitly described, but their acquisition as character traits seems essentially to be a bottom-up process." (2009, 10)

Considering the different AMA design tracks, it might be the case that some further explanation is needed as to why AWS, if not all AMAs, are unable to

generate genuine ethical decisions. After all, with regard to providing moral capacities to artifacts, Misselhorn highlights such differences, explaining that:

> [An] important issue is how moral capacities can be implemented in artificial systems. This entails two questions: first, with which moral standards artificial systems should be furnished and, second, how those standards can be implemented. Both issues are related since a decision for a certain ethical framework also entails certain constraints on its realization in a software program. (2018, 165)

Furthermore, Misselhorn notes that, as regards artificial systems, there are three approaches to moral implementation (i.e., the top-down, bottom-up, and hybrid tracks), and these differing methods "bring together a certain ethical theory with a certain approach to software design." (2018, 165) So, perhaps the particular ways that the no-ought-from-is doctrine is explicitly and individually realized in the different AMA tracks should also be considered. For the purposes of the present work, the top-down AMA approach is further examined in relation to the is-ought problem.

Top-down AMAs and Classical Programming

As noted earlier, the top-down approach for designing AMAs could be realized through encoding ethical principles into a machine's internal program. The assumption here is that, once certain moral precepts have been hardwired into the latter, the actions and behaviors of artifacts would, then, be regulated by such. Wallach and Allen further explain that:

> ... a top-down approach takes an ethical theory, say, utilitarianism, analyzes the informational and procedural requirements necessary to implement this theory in a computer system, and applies that analysis to the design of subsystems and the way they relate to each other in order to implement the theory. (2009, 80)

So, for Wallach and Allen (2009, 84), the top-down track centers on the idea of having a set of rules that, in turn, could be developed into an algorithm. Note that the foundational assumptions of this track, in a way, may also be related to the direct programming or classical AI method of creating intelligent systems.

Proponents of classical AI, also known as 'Symbolic AI,' assert that artificially intelligent systems may be achieved by writing sophisticated computer programs (Haugeland 1985, 112-114). [9] This parallels Searle's idea that an "appropriately programmed computer really is a mind, in the sense that computers given the right programs can be literally said to understand and have other cognitive states." (1980, 417) Note that the said view, which Searle calls

[9] Haugeland (1985, 112) also calls the classical AI track as 'Good Old Fashioned Artificial Intelligence' or GOFAI.

'strong AI,' states that the human mind could be likened to a computer program, implemented by a brain that functions as its hardware (Mabaquiao 2014).[10]

Classical AI works under the assumption that brains are nothing more than complex machines, which entails that, in order to create autonomous agents[11] (Pfeifer and Scheier 1999, 25-27), one has to write computer programs that would serve as their central intelligence system. If one, then, grants that the foundational assumptions of the top-down AMA track are, in principle, the same with, if not grounded on, classical AI, the next issue would be how to show that the former is susceptible to Hume's is-ought problem. To address this issue, one may look into the inner workings of a top-down AI's artificial mind.

To argue that AMAs developed via the top-down method are indeed predisposed to the is-ought problem, it must be recalled that these technologies have built-in computer programs that try to mimic the human mind, specifically its ability to exercise thinking. In addition, it should be highlighted that such programs are based on the notion that an artificial mind is always composed of two parts, namely: (1) a world model and (2) utility function (Hall 2011, 512).

A world model (WM), on the one hand, is the part of an artificial mind that houses the objective knowledge regarding the world. It contains all the facts about the world that it is modeling, and this may, in turn, be used by an artifact to plan, evaluate, and predict the effects of its actions. The utility function (UF), on the other hand, "establishes a preference between world states with which to rank goals." (Hall 2011, 515) It may, thus, be said that the WM of an artifact is the one primarily responsible for providing it with the current state of affairs in the world, including the different possibilities or consequences of its actions, while its UF calculates which of these is the most desirable given a specified goal.

The problem with the said model is that there seems to be no way of bridging the gap between a machine's utility function and its world model. This is because, in principle, the move from first modeling the actual and potential state of affairs in the world, to finally deciding which among these possibilities is most preferable, is quite similar to the attempt of deriving an 'ought' from a series of factual statements about the world.[12] The is-ought problem is at work in this model, and there seems to be no clear solution to this philosophical worry at present, as discussed in the previous sections.

Recall that the WM is composed of various facts about the world. Suppose that an autonomous artifact with a central processing system, for instance, sees a speeding automobile that appears to be in a collision course with a group of tourists. Its WM would generate a number of facts about the situation, including

[10] This is in contrast to what Searle has called "weak AI," (1980, 417) which holds that computers are nothing but powerful tools for studying the mind.

[11] This may be further related to what Goertzel has dubbed as "artificial general intelligence." (2007, 1161-1163)

[12] As noted earlier, Hall (2011, 514) also mentions that the WM and UF are separated by 'Hume's is-ought guillotine,' but there appears to be no detailed analysis of the said idea.

the potential casualty count if nobody does anything, the risks involved in helping, and so on. From all of these facts, the artifact may come up with predictions of the possible effects and side effects of its action, if not its inaction.

However, note that the procedure by which a machine decides that a particular action is better over another seems to be quite untenable. This is because, citing the no-ought-from-is thesis, evaluative conclusions may never be derived from factual statements, which is said to be the manifest function of the UF in making sense of the factual data generated by a machine's WM.

It may then be argued that, if AMAs developed via the top-down route subscribe to the WM-UF model, then these artificially intelligent machines, strictly speaking, would not really be able to come up with genuine moral judgments due to the is-ought problem. Either their so-called judgements are absent of good moral grounding, or their conclusions are somehow empty in the ethical sense. Note that this further explains Boyles' (2021) point that technologies based on the sense-decide-act cycle are quite questionable in terms of their moral standing and ethical decision-making.

On top of the idea that AWS are unable to produce genuine moral judgments because their actions largely depend on the data gathered by their sensors and sensing software, it may be argued that all technologies modeled via the direct programming method would not be able to perform the said task given their internal design. These artifacts are unable to generate actions with actual moral worth because there is no way of reconciling their WM-UF parts. Such is the case given that this step would require the deduction of evaluative conclusions from purely factual premises.

If one takes into consideration the internal design of top-down AMAs, the idea that certain kinds of intelligent machines are, in principle, predisposed to the is-ought problem would make further sense. Consider, for instance, Boyles' (2021, 124-125) explanation why descriptivist-based AWS are not capable of circumventing Hume's Law. As regards the said notion, Boyles states that descriptivist strategies still make use of implicit ought-claims (i.e., to come up with an evaluative conclusion from descriptive premises). Furthermore, he also holds that, "if such is the case, then it may also be argued that the descriptivist solution to Hume's Law is nothing different from the prescriptivist idea that evaluative judgments may be uncovered in the factual premises of moral arguments." (2021, 124) In a way, this explanation regarding the said types of AWS become more intelligible if one further relates such to the fact that top-down AMAs are largely determined by their world model and utility function.

Consider the following case: suppose that a top-down AMA finds itself in a position of having to confront a modified version of Foot's (1967) trolley problem. Imagine a runaway train that is fast speeding down a railway and there are five individuals at the end of one of the tracks. The said train is headed right straight for these people, and the top-down machine could prevent their demise because it is standing by the lever that controls the tracks. If it pulls the lever, the train

would switch to a different set of tracks. However, if the said lever is pulled, the train would head directly to another person on this different track. What must our top-down AMA do? Should it pull the lever? If it does, the train would be diverted onto the new track where one person would be killed. If it does not, the train would kill the five individuals at the end of the main track. Which is the correct choice?[13]

How would a top-down AMA address the mentioned situation? Let us suppose that our artifact has built-in ethical systems (e.g., utilitarianism, deontology, etc.) from which it could choose the right course of action to take. But the said artifact has a problem: "What ethical theory should it choose to base the right course of action?" To answer this, let us suppose that, on top of its basic program, there is a second-order platform that might guide our artifact to favor or prefer one ethical theory over the rest. Here, an impasse is reached.

All considerations in the second-order platform seem to be exactly the same as those in the first. Since there is no forthright way of deciding which moral principle is better over the other regardless of the programming levels (i.e., without, of course, begging the question), this endeavor would likely lead to a problem of circularity. Boyles (2021, 123) further explains that the said strategy might even be prone to relativism, if not an arbitrary assignment of values – preferring a specific theory, but with no justified (moral) grounds. Note also that the issue of having a multitude of ethical theories that conflict with each other has already been raised previously (Tonkens 2009; Lara and Deckers 2019).

By stressing the difficulties in refereeing between opposing moral precepts, which in effect also highlights the issue of employing ought-premises in programming ethical machines, it could be said that one is left to work with only factual propositions. Note that such a case eventually results in yet another is-ought problem. However, it might still be argued by others that the adjudication process between the various competing ethical theories may be addressed by simply giving a machine a certain modification. For one, Boyles' (2021, 124) considers the possibility of this issue being "addressed by further adjusting the decide capacity" of an artifact. Actually, taking into account the WM and UF of top-down AMAs, the said modification concerns an artifact's UF, while its WM might also be affected.

Instead of just focusing on a machine's decide capacity, citing a top-down AMAs internal program would provide a better picture on why they are predisposed to Hume's is-ought. In altering a machine's UF, note that labeling a particular theory as the most appropriate one (i.e., as compared to other ethical precepts) denotes that such a theory is actually the most relevant among its competitors. So, claiming that the decide capacity of an artifact would be the one

[13] Note that the original intention of Foot's (1967) thought experiment is to show that there is a difference between letting someone die and killing a person. This, for her, has ramifications on the moral status of some abortion cases.

affected by the proposed strategy (Boyles 2021, 124) appears to have only scratched the surface.

Furthermore, with regard to Boyles' (2021, 124) view that the process of identifying something as being the most relevant would eventually run into issues, note that a top-down AMA's MW could also be examined to better understand this. For one, remember that he generally maintains that encoding the decide capacity of an AWS (i.e., so that it could adjudicate and select which particular theory is most relevant in a given situation) somehow leads to the frame problem. It may, thus, be said that this issue concerns a machine's WM, since the latter would be the one responsible for modeling each and every fact about a specific situation, resulting in an infinite regress. This is the reason why an artifact with such a program would not be able to generate genuine moral judgments; this task entails that a top-down machine would have to infinitely account for all the factual data processed inside its WM.

Recall that descriptivists claim that moral statements may be reduced to factual premises under certain conditions (Boyles 2021, 118). So, it may be argued that an ethical theory is the most relevant one if it is the most apt in a given situation. However, it must be remembered that the descriptivist approach is doomed to fail, since it still smuggles in implicit ought claims (i.e., in attempting to infer an 'ought' from a series of is-statements) as part of its starting set of facts.

Moreover, even if one grants that this strategy succeeds, it seems to only work in very narrow cases at best. It may even be argued that the possibility of actually identifying such narrow cases is close to impossible because this exercise could lead up to other issues, like the frame problem – considered by many as a technical and philosophical problem that focuses on "representing the effects of action[s] in logic without having to represent explicitly a large number of intuitively obvious non-effects." (Shanahan 2016) Note that this parallels the view of Moss, that "[d]etermining the best action at every moment would overwhelm a finite computational device." (2016, 2) Yet again, it seems that further looking into the internal design of top-down AMAs (i.e., that they are largely determined by their WM and UF) provides further grounding as to why certain types of machines are unable to produce genuine ethical decisions.

Finally, a couple of things may also be pointed out about the idea of pre-programming ethical theories into AMAs. First, it must be noted that, even if the top-down track prospers, it would not be that simple to assign concepts like 'praiseworthiness,' 'blameworthiness,' and so on to such artifacts. Since the said values were just pre-programmed to them by their designers, achieving full moral agency by means of this track is a bit questionable. For one, Krzanowski and Trombik explain that:

> Can then such a deep ethics be computed (in the Church–Turing sense), given that metaphysics is not mathematical? Ethical rule-based on Hobbesian, Kantian, utilitarian or other ethical schools can be to some extent translated into a computer algorithm and made 'computable.' But then all 'metaphysical'

> dimensions of the ethical actor are 'lost in translation.' If a machine is programmed according to 'translated' rules... this ethics would be a special type of ethics, not ethics in the deep, metaphysical sense. (2021, 143)

Krzanowski and Trombik hold that ethics, in a deep (metaphysical) sense, is non-computable, and they maintain that there really is no other way of defining and accounting for what is "computable." (2021, 143)

Recall that, although one may construct a system consisting of a set of prescriptions, like a standard computer program, this system does not really equate to a moral system (Gensler 2011, 57). So, whenever AMAs built through the top-down route initiate actions that, at first glance, appear to be ethical in nature, such as their apparent moral judgments, it might be the case that the actions of these machines are actually devoid of any ethical value.

Second, note that there is a difference between ethical reasoners and ethical decision-makers (McDermott 2008). Ethical reasoners are artifacts that model the reasoning processes of human beings (i.e., in coming up with ethical conclusions). Ethical decision-makers, in contrast, are those that duplicate or mimic what in people are classified as ethical decisions. The primary distinction between these two is that the latter really understands what is actually at stake whenever moral dilemmas or conflicts arise (e.g., the ethical thing to do in a given situation and how it seems to conflict with one's self-interest).

On the other hand, ethical reasoners, in a way, just mechanically generate moral conclusions by, say, considering the facts at hand. Similarly, Hunyadi explains that:

> As far as *machine ethics* is concerned, this means one thing: if you program a specific set of ethical principles into a machine, you do not make the machine an artificial *moral* agent, but an *executor of those specific principles*, which is an entirely different thing. This so-called 'artificial agent' will be expected to respond according to *those ethical principles*, chosen by the programmer. (2019, 62)

Hunyadi, thus, further clarifies that, as regards an artificial system, it is more apt to label such an 'artificial moral executor' instead of 'artificial moral agent,' specifically an "artificial *utilitarian*, *deontological* or *perfectionist* executor, depending on the ethical principles chosen by the programmer." (2019, 62)

In light of the differences cited above, it could be contended that, in order to actually realize the concept of moral machines, they should not only be simple ethical reasoners, but ethical decision-makers as well. Unfortunately, the suggested strategy of encoding machines with pre-programmed ethical precepts does not fall under the latter. So, the idea of assigning ethical notions such as praiseworthiness, blameworthiness, and others to artificial moral agents built via the top-down track seems like a lost cause.

Conclusion

As regards top-down AMAs, it was argued that they would not be able to come up with morally-relevant judgments, since artifacts built this way are primarily composed of two parts, namely: a world model and utility function. In principle, there is no way of reconciling these two parts because such would entail the derivation of evaluative claims from a set of purely factual ones, and this goes against the general tenets of the is-ought problem. One consequence of this is that either the so-called moral judgements of these AMAs would be absent of any good moral grounding, or their generated conclusions would be empty of ethical value. Note that, instead of simply citing Hume's is-ought, as well as the sense-decide-act cycle, one could further look into a machine's internal design in order to have a better understanding as to why they are incapable of moral reasoning.

As discussed above, there are different types of artificial moral agents, developed either via the top-down, bottom-up, and hybrid approaches. With regard to top-down AMAs, which is the focus of the present work, it appears that they may not be considered as authentic moral agents (i.e., as compared to humans) due to their internal design. Note, however, that there are those, like Nadeau (2006), who also contend that even biological humans may not be considered as moral agents.

Prospectively, further research on how the other forms of AMA technologies, specifically those designed via the bottom-up and hybrid methods, fare against the no-ought-from-is thesis may be looked into. Would these strategies be susceptible to Hume's is-ought as well? If yes, then how would this relate to other concerns put forward against such types of AMAs? For instance, in relation to bottom-up AMAs, consider Baum's view that "it is impossible for AI designers to avoid embedding certain ethics views into an AI... because there is no one single aggregate ethical view of society." (Baum 2020, 167) It is, thus, interesting to know if the in-principle thesis embedded in the is-ought problem could aid in further understanding these sorts of ideas. Of course, if it really turns out that all these AMAs encounter insurmountable issues resulting from Hume's Guillotine, then perhaps it may be high time to look for other viable alternatives.

References

Allen, Colin, Iva Smit, and Wendell Wallach. 2005. "Artificial Morality: Top-down, Bottom-up, and Hybrid Approaches." *Ethics and Information Technology* 7: 149-155.

Baum, Seth D. 2020. "Social Choice Ethics in Artificial Intelligence." *AI & Society* 35 (1): 165-176.

Black, Max. 1964. "The Gap Between 'Is' and 'Should.'" *Philosophical Review* 73 (2): 165-181.

Boulanin, Vincent, and Maaike Verbruggen. 2017. *Mapping the Development of Autonomy in Weapon Systems*. Solna: Stockholm International Peace Research Institute.

Boyles, Robert James M. 2021. "Hume's Law as another Philosophical Problem for Autonomous Weapons Systems." *Journal of Military Ethics* 20 (2): 113-128.

Boyles, Robert James M., Dacela, Mark Anthony, Evangelista, Tyrone Renzo, and Jon Carlos Rodriguez. 2022. "COVID-19 and Singularity: Can the Philippines Survive another Existential Threat?" *Asia-Pacific Social Science Review*, *22* (2), 181-195.

Brown, Charlotte R. 2008. "Hume on Moral Rationalism, Sentimentalism, and Sympathy." In *A Blackwell Companion to Hume*, edited by Elizabeth S. Radcliffe, 219-239. Oxford: Blackwell Publishing Ltd.

Cervantes, José-Antonio, Sonia López, Luis-Felipe Rodríguez, Salvador Cervantes, Francisco Cervantes, and Félix Ramos. 2020. "Artificial Moral Agents: A Survey of the Current Status." *Science and Engineering Ethics* 26: 501-532.

Cohon, Rachel. 2010. "Hume's Moral Philosophy." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. Available at: http://plato.stanford.edu/entries/hume-moral. Accessed: 11 August 2021.

Dennett, Daniel. 1984. "Cognitive Wheels: The Frame Problem of AI." *In Minds, Machines, and Evolution: Philosophical Studies*, edited by Christopher Hookway, 129-151. Cambridge: Cambridge University Press.

Department of Defense. 2012. *Autonomy in Weapon Systems*. DOD Directive 3000.09. Washington: Department of Defense. Available at: https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf. Accessed: 3 August 2021.

Fisher, Andrew, and Simon Kirchin. 2006. *Arguing about Metaethics*. Oxon: Routledge.

Foot, Philippa. 1967. "The Problem of Abortion and the Doctrine of Double Effect." *Oxford Review* 5: 5-15.

Gensler, Harry J. 2011. *Ethics: A Contemporary Introduction.* 2nd ed. New York: Routledge.

Goertzel, Ben. 2007. "Human-level Artificial General Intelligence and the Possibility of a Technological Singularity: A Reaction to Ray Kurzweil's The Singularity is Near, and McDermott's Critique of Kurzweil." *Artificial Intelligence* 171: 1161-1173.

Gunkel, David J. 2018. "The Other Question: Can and Should Robots have Rights?" *Ethics and Information Technology* 20: 87-99.

Hall, Storrs J. 2011. "Ethics for Self-Improving Machines." In *Machine Ethics*, edited by Michael Anderson and Susan Leigh Anderson, 512-523. Cambridge: Cambridge University Press.

Hare, Richard Mervyn. 1952. *The Language of Morals.* Oxford: Clarendon Press.

Haugeland, John. 1985. *Artificial Intelligence: The Very Idea*. Cambridge: MIT Press.

Hume, David. 1739/1964. *A Treatise of Human Nature*, edited by L.A. Selby-Bigge. Oxford: Clarendon Press.

Hunyadi, Mark. 2019. "Artificial Moral Agents. Really?" In *Wording Robotics: Discourses and Representations on Robotics*, vol. 130, edited by Jean-Paul Laumond, Emmanuelle Danblon and Céline Pieters, 59-69. Cham: Springer.

Joaquin, Jeremiah Joven. 2012. "John Searle and the Is-ought Problem." *Scientia* 2 (1): 53-66.

Krzanowski, Roman M., and Kamil Trombik. 2021. "Ethical Machine Safety Test." In *Transhumanism: The Proper Guide to a Posthuman Condition or a Dangerous Idea?*, edited by Wolfgang Hofkirchner and Hans-Jörg Kreowski, 141-154. Cham: Springer.

Lara, Francisco, and Jan Deckers. 2019. "Artificial Intelligence as a Socratic Assistant for Moral Enhancement." *Neuroethics*, 1-13. (please check reference, as it is inexact: https://philpapers.org/rec/LARAIA-5)

Mabaquiao, Napoleon. 2014. "Turing and Computationalism." *Philosophia: International Journal of Philosophy* 15 (1): 50-62.

McDermott, Drew. 2008. "Why Ethics is a High Hurdle for AI." In *North American Conference on Computers and Philosophy*. Bloomington, Indiana, July 9-12. Available at: http://www.cs.yale.edu/homes/dvm/papers/ethical-machine.pdf. Accessed: September 16, 2014.

Misselhorn, Catrin. 2018. "Artificial Morality. Concepts, Issues and Challenges." *Social Science and Public Policy* 55: 161-169.

Moss, Henry. 2016. "Genes, Affect, and Reason: Why Autonomous Robot Intelligence Will Be Nothing Like Human Intelligence." *Techné: Research in Philosophy and Technology* 20 (1): 1-15.

Nadeau, Joseph Emile. 2006. "Only Androids Can Be Ethical." In *Thinking about Android Epistemology*, edited by Kenneth M. Ford, Clark Glymour and Patrick Hayes, 241-248. Cambridge: MIT Press.

Pfeifer, Rolf, and Christian Scheier. 1999. *Understanding Intelligence.* Cambridge: MIT Press.

Restall, Greg, and Gillian Russell. 2010. "Barriers to Consequence." In *Hume on Is and Ought*, edited by Charles Pigden, 243-259. Basingstoke: Palgrave Macmillan.

Schurz, Gerhard. 1997. *The Is-ought Problem: An Investigation in Philosophical Logic.* Dordrecht: Springer.

Shanahan, Murray. 2016. "The Frame Problem." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. Available at: https://plato.stanford.edu/entries/frame-problem/. Accessed: 24 October 2017.

Sharkey, Noel. 2010. "Saying 'No!' to Lethal Autonomous Targeting." *Journal of Military Ethics* 9 (4): 369-383.

Searle, John R. 1964. "How to Derive 'Ought' From 'Is.'" *The Philosophical Review* 73 (1): 43-58.

————. 1980. "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3 (3): 417-457.

————. 1995. *The Construction of Social Reality*. London: Penguin Books Ltd.

Snare, Francis. 1992. *The Nature of Moral Thinking.* London: Routledge.

Sparrow, Robert. 2016. "Robots and Respect: Assessing the Case Against Autonomous Weapon Systems." *Ethics & International Affairs* 30 (1): 93-116.

Sullins, John P. 2009. "Artificial Moral Agency in Technoethics." In *Handbook of Research on Technoethics*, edited by Rocci Luppicini and Rebecca Adell, 205-221. New York: IGI Global.

Tonkens, Ryan. 2009. "A Challenge for Machine Ethics." *Minds and Machines* 19 (3): 421-438.

Wallach, Wendell, and Colin Allen. 2009. *Moral Machines: Teaching Robots Right From Wrong.* New York: Oxford University Press.