Φιλοσοφια
Volume 26, 2: 2025

CONJECTURES ON THE POSSIBILITY OF RECASTING HUME'S GUILLOTINE FOR HYBRID-DESIGNED ARTIFICIAL MORAL AGENTS

Robert James M. Boyles
De La Salle University, Philippines

This article explores the prospects of re-appropriating Hume's Guillotine to argue against the suggestion of developing artificial moral agents via a hybrid approach. Hume's Guillotine states that any type of evaluative claim may never be exclusively drawn from factual statements. Hybrid routes of constructing artificial moral agents, meanwhile, are those that fuse two kinds of design approaches, namely: (1) top-down programming, which entails encoding artifacts with a set of moral precepts, and (2) bottom-up strategies that are characterized by learning or developmental techniques. In this essay, it is argued that there seems to be some reasonable grounds for believing that artifacts made by means of hybrid methods are vulnerable to Hume's Guillotine, especially in terms of their ability to conjure moral judgments. To preliminarily lay the groundwork for this claim, certain objections posed against top-down and bottom-up strategies are revisited in the hopes of further recasting them (i.e., to show that the moral reasoning faculty of systems built through hybrid processes are also suspect).

Keywords: artificial intelligence ethics, Hume's Guillotine, hybrid design approach, machine ethics, moral reasoning

INTRODUCTION

In the area of artificial intelligence (AI) ethics, one of the key concerns is how to aptly address the differing types of ethical problems surrounding intelligent technologies (Siau and Wang 2020, 74). These ethical worries range from more practical matters, like how technology entrepreneurs supposedly dictate which notion of the good or moral standard to adopt (Zaidi 2020, 96), to more speculative issues, such as the extermination of humankind (Chalmers 2010, 4). In relation to these AI-related ordeals, the subfields under AI ethics also proffer distinctive ways how to deal with them. One of these domains is machine ethics.

Machine ethics is concerned with creating AI systems that have moral precepts, allowing them to properly face ethical issues (Anderson and Anderson 2007, 1). In general, it is focused on providing artifacts with a set of moral values, including the

capability to exhibit ethical behavior (Siau and Wang 2020, 76). These kinds of artifacts are commonly known today as artificial moral agents (AMAs) (Wallach and Allen 2009, 4).

Wallach and Allen (2009, 10) explain that AMAs may be developed by means of (1) *top-down*, (2) *bottom-up*, and (3) *hybrid design techniques*. Top-down methods, on the one hand, subscribe to the idea of building AMAs by means of encoding ethical precepts to them (i.e., in order to regulate their actions) (Wallach and Allen 2009, 79-80). Bottom-up paths, on the other hand, are those that make use of design strategies that are characterized by learning or developmental approaches. Wallach and Allen (2009, 80) further clarify that, in bottom-up methods, greater priority is given to the design of environments where agents may freely survey and discern different plans of action, learning and being rewarded for executing ethically meritorious behavior.

Lastly, hybrid tracks are those that merge top-down and bottom-up AMA methods (Wallach and Allen 2009, 117-118). For purposes of the present work, the emphasis would be on technologies built via the hybrid path, specifically on how they relate with Hume's Guillotine. The latter suggests that evaluative claims cannot be drawn from purely fact-based propositions (Frankena 2006, 49). As regards systems that are based on top-down and bottom-up methodologies, it must be pointed out that the said philosophical problem has already been raised in the past.

In separate instances, it has been shown that AMAs fashioned by means of top-down and bottom-up approaches appear to be vulnerable to Hume's Guillotine (Boyles 2022, 180-185; Boyles 2024, 17-20). In light of the principle underlined by this problem, it has been demonstrated that these systems are only capable of producing moral judgments that are of no moral worth. Thus, the primary goal of this research is to explore the possibility of doing the same for hybrid-designed AMAs. However, note that the present study makes no commitments if (or how) the resulting artifacts from top-down, bottom-up, and hybrid strategies would be fundamentally and actually distinct from each other. This essay merely speculates on the possibility of their differences.

To lay the groundwork for the claim that AMAs developed through hybrid routes are also subject to Hume's Guillotine, the following section provides a quick overview of the said philosophical problem. The ensuing section then gives a synopsis of studies that have used this problem to challenge the tenability of different design approaches of crafting artificially intelligent systems, specifically in terms of the ability to conjure moral judgments. The next section, on the other hand, briefly defines the hybrid method of constructing AMAs, focusing on connectionist models. Meanwhile, the following section contends that artifacts based on hybrid methods are, in fact, unshielded against Hume's Guillotine. This is demonstrated by conceptually analyzing the nature of the hybrid design strategy, arguing that the objections put forward against the top-down and bottom-up tracks could be re-appropriated to show that the moral reasoning faculty of hybrid-based systems remains unaccounted for if one grants Hume's (1739/1964) worries. The last section of this paper ends with a short summary and a few closing remarks.

TWO KINDS OF STATEMENTS AND HUME'S GUILLOTINE

Black (1964, 166) elucidates that Hume's Guillotine, also known as "Hume's Law" (Hare 1952) or the "is-ought problem," has been generally attributed to the Scottish thinker David Hume. The said problem highlights the idea that evaluative statements cannot be drawn exclusively from factual claims (Frankena 2006, 49). To better grasp this problem, it would be best to first differentiate the two types of statements involved in it, namely: (1) *descriptive* and (2) *evaluative* ones.

Descriptive statements are those that have truth-value (Mohanty 1984, 37). Basically, this means that the said propositions could either be true or false. Evaluative statements, on the other hand, cannot be appraised in the same manner, since they operate more like prescriptions of actions.

Searle (2021, 5) explains that the dissonance between evaluative and descriptive statements has led many to believe that the former cannot be inferred from the latter.[1] For instance, Hume maintains 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 (Hume 1739/1964, 243-244).

Thus, following the differences between descriptive and evaluative statements, it has been said that the inference of an evaluative claim cannot be validly done using factual propositions (Frankena 2006, 49).

In relation to Hume's Guillotine, a couple of views have also been put forward in order to bypass it. These responses are (1) *moral descriptivism* and (2) *moral non-descriptivism*. Moral descriptivism subscribes to the suggestion that ethical claims could be considered as factual ones under certain parameters. (Kim 2006) This view generally employs the strategy of converting moral claims into descriptive propositions, so that the truth-value of the former would be revealed. Consider, for instance, Searle's (1964) approach on how to derive an "ought" from an "is."

According to Searle (1964, 44), statement 5 below, which is evaluative in nature, could be drawn from the first four propositions in the following argument:

(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.

However, it must be pointed out that Searle's (1964) strategy is not completely problem-free. For one, it has been pointed out that it still employs an auxiliary statement, "(4a) Other things are equal" (Searle 1964, 46), in order to finally bridge proposition 4 to 5, and this additional premise does not appear to be purely factual in nature (Joaquin 2012, 60). Also, it is not that evident how this maneuver would work apart from the sole act of promise-making (Boyles 2021, 119). So, it may be said that moral descriptivism remains contentious at present.

Moral non-descriptivism, on the other hand, suggests that moral judgments cannot be treated as actual depictions of the world, since they are best seen as mere expressions of commands (Kim 2006). A common example of a moral non-descriptivist framework is universal prescriptivism (Gensler 2011, 56). Under this framework, ought-statements are considered as prescriptions.

Prescriptivists construe ought-statements as one's impartial desire on how to live (Gensler 2011, 57). Hence, for them, these claims take the form of universalizable prescriptions. However, by regarding them as such, Gensler (2011, 61-63) further elucidates that this leads to the negation of the possibility of attaining moral truths. This is because, under prescriptivism, ought-statements are basically reformulated into a form of logical test that just inspects the consistency of an individual's moral views (Gensler 2011, 58). Furthermore, Gensler (2011, 57) clarifies that, if one treats it as an ethical theory, prescriptivism does not have any value in the ethical sense. This is because it is possible to develop systems without any moral value. Among these kinds of systems include the legal provisions of a state, computer algorithms, and so on.

In a way, it may be said that prescriptivism accepts the implications of Hume's Guillotine. This is because prescriptivists resort to uncovering the concealed evaluative claims in moral arguments (Joaquin 2012, 55-56). Thus, they assume that evaluative statements are hidden in every moral argument, and the goal is to bring them forth for purposes of evaluation.

In view of the foregoing, one could argue that there seems to be no definitive solution to Hume's Guillotine today. Moral descriptivism and non-descriptivism have their own fair share of problems, both of which are without final resolutions. Hence, the worries proffered by Hume (1739/1964) may be treated as live problems at present.

DISPUTING THE MORAL REASONING CAPACITY OF ARTIFACTS

Following the suggestion that evaluative statements may not be drawn from purely factual ones (Frankena 2006, 49), some studies have used Hume's Guillotine to argue against the ability of AIs to induce moral judgments. For one, it has been

contended that Autonomous Weapons Systems (AWS) run up against the said problem (Boyles 2021, 120-125). This is because operationalizing the concept of autonomy in AWS requires the integration of the (1) sense, (2) decide, and (3) act aspects.[2] The problem with such, however, is that, in order to complete the sense-decide-act cycle, evaluative claims need to be deduced from factual statements, given that the facts or data harnessed by the sense part eventually become the input for the other two.

Furthermore, in a more general sense, Hume's Guillotine has also been utilized to argue against AMAs developed by means of top-down approaches (Boyles 2022, 180-185). As discussed earlier, the latter focuses on programming or embedding artifacts with ethical principles, which aid in regulating their behavior (Wallach and Allen 2009, 79-80). These artifacts also possess artificial minds that are standardly made up of two parts—a world model (WM) and a utility function (UF) (Hall 2011, 512).

The WM of top-down-based systems is the part that stores all the objective knowledge about the world (Hall 2011, 512). So, this segment houses the different facts of the environment that an AMA is trying to model. Conversely, Hall (2011, 515) states that the UF determines the "preference between world states with which to rank goals." Hence, this part is arguably responsible for calculating which particular action is most preferable in a given situation.

In terms of their ability to come up with genuine moral judgments, however, it has been argued that artifacts built via top-down routes are actually incapable of doing such, given their WM-UF architecture (Boyles 2022, 181). This is because there is no definite way of linking these two parts. Recall that, according to Hume's Guillotine, evaluative statements could never be derived from factual premises (Frankena 2006, 49), and bridging the UF part with the WM necessitates such kind of inference. Thus, the consequence is that even if AMAs based on top-down strategies appear capable of producing moral judgments, these discernments do not really have actual moral worth.

Lastly, Hume's Guillotine has also been employed to show that AMAs developed from bottom-up tracks are also unable to conjure real moral judgments (Boyles 2024, 17-20). As mentioned earlier, bottom-up approaches place emphasis on developing environments where autonomous agents are accorded the freedom to consider various kinds of behaviors, rewarding them whenever they exhibit morally good ones (Wallach and Allen 2009, 80). However, the moral judgments that bottom-up-designed systems discharge apparently do not have a good philosophical basis, since they acquire an appreciation of morally relevant ideas via gathering particular facts or data from the environment. So, in relation to Hume's (1739/1964) worries, it has been said that although AMAs produced from bottom-up paths seem to have the capacity to generate moral judgments, these are actually empty of any ethical value as well.

Considering how it affects both top-down and bottom-up AMA approaches, there appears to be a need to look into the particular way that Hume's Guillotine arises in hybrid techniques. In order to lay the groundwork for this task, perhaps a number of arguments that were utilized against top-down and bottom-up methods may be revisited. Thus, the aim of the ensuing sections is to give a brief overview of hybrid-based AMAs, providing a preliminary analysis of how to potentially challenge their foundational status by means of recasting the arguments put forward against top-down and bottom-up AMA tracks.

UNDERSTANDING HYBRID METHODS AND CONNECTIONISM

In a nutshell, hybrid-designed AMAs may be defined as systems that make use of top-down and bottom-up techniques for their ethical decision-making procedures (Cervantes et al. 2020, 507). In light of these two design strategies, artifacts built via the hybrid manner supposedly possess the capacity to come up with varying and non-rigid moral judgments on a case-to-case basis. Given the worries in both top-down and bottom-up methods, Wallach and Allen (2009, 117) elucidate that:

> If neither a pure top-down approach nor a bottom-up approach is fully adequate for the design of effective AMAs, then some hybrid will be necessary. Furthermore, as noted, the top-down, bottom-up dichotomy is somewhat simplistic. Engineers commonly start with a top-down analysis of complex tasks to direct the bottom-up assembly of components.

Wallach and Allen (2009, 118) further clarify that the use of connectionist networks is thought to be one of the possible ways to realize hybrid kinds of AMAs.

In brief, connectionism may be defined as an AI design strategy that subscribes to the suggestion that the path towards computationally replicating the cognitive functions of human beings is through the creation of artificial neural networks (Carter 2007, 186). So, instead of serially processing data by means of, say, a central intelligence system, which is common in traditional AI methods, connectionism employs parallel and distributed processing units or "nodes." The latter functions like the neurons in human brains. The processing units may also be taken as the basic elements of a connectionist system, also known as a "net" or "network" (Mabaquiao 2012, 124).

Nodes, which are connected to one another, have the capability to receive, process, or send data based on their classification (Mabaquiao 2012, 124). On the one hand, input nodes possess the ability to obtain information from outside the network. Hidden nodes, on the other hand, collect information from the input units, processing data from the latter before forwarding them to the output ones. Output nodes are those that carry out the data that have been processed. In terms of information processing, specifically referring to neural networks, Pfeifer and Scheier (1999, 139) further hold that they are composed of a huge set of neurons that can process information at the same time. They maintain that one key advantage of this kind of setup is that it allows systems to accomplish a significant amount of processing in spite of some particular components being slow. It must also be pointed out that biological neurons supposedly operate in the same manner.

Moreover, nodes in one network have different levels of activation pertaining to the directional flow of information (Mabaquiao 2012, 124-125). Specifically, these two activations may either be the (1) *feed-forward* or (2) *recurrent type*. The former adheres to the standard input-hidden-output directional flow or pattern, while the latter kind loops back processed data from the output units to other nodes, so as to facilitate added processing. Both activations are also influenced by other units attached to them. In effect, the information flow from one unit to the other would be contingent on the connection type that links various nodes.

The connections between nodes may also be characterized based on their particular strengths or weights (Mabaquiao 2012, 125-126). This means that connections could either have a positive numerical value or a negative one, affecting the amount of information that flows within them. Since each unit has a threshold—the maximum data that it may retrieve, nodes are activated when these limits are met. So, the total data that a node receives would depend on the size of the input sent to it, including the connection strength that oversees the flow of information.

The reaction of a particular node to the data it acquires may be classified into two types, namely: (1) *excitatory* and (2) *inhibitory* (Carter 2007, 188). The first kind, or the excitatory-type, on the one hand, increases the activation levels of the nodes that have been interconnected to it, which, in turn, results in positive weights. Inhibitory ones, meanwhile, decrease the activation levels to produce negative weights. Thus, the connectionist model also accedes to the view that information processing transpires by means of the interactions of the nodes in a specific network (Mabaquiao 2012, 126).

Furthermore, to account for learning, it must be underlined that there are two different kinds of connection weights in neural networks, which are classified as (1) *hardwired* and (2) *learned* (Mabaquiao 2012, 126). The former type of connection weight remains unchanged or unadjusted, while the latter is the one that has already been altered. In neural networks, it is said that learning occurs whenever the weights of the networks are adjusted.

Mabaquiao (2012) additionally explains that, under the connectionist framework, learning takes place as follows:

> The connection weights or the strengths of the connections among the units may be hardwired, learned, or both… When a network's connection weights are adjusted, it is said that the network is learning. Learning in this sense can be unsupervised or supervised (Mabaquiao 2012, 126).

Thus, networks are categorized depending on the type of learning that they espouse (Pfeifer and Scheier 1999, 165-172). On the one hand, supervised networks are those that require some sort of instructor to facilitate learning. Meanwhile, non-supervised ones do not need a guide in the learning process.

As regards employing connectionism to develop moral artifacts, Wallach and Allen (2009) explain that a number of thinkers have already suggested that this is the right direction. They further clarify that the motivation behind this belief partly stems from the following:

> The view in ethics called "particularism" holds that moral reasons and categories are richly context-sensitive… The contextual details of when actions are permissible may be so rich that it is impossible to summarize them in universal moral principles. Connectionist models are good at capturing context-sensitive information without explicit or general rules (Wallach and Allen 2009, 122-123).

Wallach and Allen (2009, 122-123) claim that connectionism seems to fit well with particularism as it is more sensitive to specific situations, especially as compared

to top-down methods to ethics. They explain that the challenges of employing standard normative ethical theories in a top-down manner have prompted many to consider other ways of imparting morality to artifacts, like using Aristotelian virtue ethics (Wallach and Allen 2009, 10). This is because the concept of virtues may be considered as a fusion between top-down and bottom-up techniques, as they could be straightforwardly defined, but the mode of obtaining them follows bottom-up principles. For Wallach and Allen (2009, 118), one way of actualizing systems via hybrid methodologies is by embedding them with certain virtues, while also enabling artifacts to further cultivate such as it moves about in the world. The latter mechanism enables hybrid-based AMAs to learn and acquire new virtues along the way. At the same time, this would allow them to enact disparate, pliable moral judgments from one context to another.

To date, numerous architectures have already been identified as adhering to the hybrid path. For instance, LIDA is a model of human cognition that takes into account key ethical features in carrying out decision-making tasks, partly employed on the mobile assistance system CareBot (Cervantes et al. 2020, 519). This architecture operates with two cognitive processes, namely: (1) top-down methods that enact ethical principles by means of certain rules and (2) bottom-up techniques that use learning mechanisms. Other examples of architectures that follow the hybrid paradigm include the (1) ethical decision-making model, (2) MedEthEx, and (3) ethical multiple-agent system (Cervantes et al. 2020, 520-523).

APPLYING HUME'S GUILLOTINE TO HYBRID AGENTS

To demonstrate that there seems to be some reasonable grounds for believing that AMAs built through hybrid options are prone to Hume's Guillotine, especially with regard to their capability of conjuring moral judgments, it would be best to scrutinize the very nature of this design strategy. Recall that, in general, artifacts developed via hybrid approaches are those that fuse both top-down and bottom-up methods (Wallach and Allen 2009, 117-118). Additionally, considering that the general motivation behind this technique revolves around the harmonization of these two approaches, there is a school of thought that one way of actualizing hybrid-based systems is by giving them virtues.

Wallach and Allen (2009, 119) maintain that the hybrid option (i.e., by means of a connectionist model) could use specific virtues as focal points to elicit the ethical behavior of AIs. This coincides with the notion that fostering one's character is more fundamental than merely possessing ethical rules. For one, virtue ethicists such as Aristotle believe that individuals who are considered to be virtuous are those who embody ideal character traits (Athanassoulis 2010). In addition, virtues emanate from natural, internal tendencies, which, in turn, could be nurtured through proper upbringing and the formation of good habits (Kraut 2022). As regards the concept of virtues, Hursthouse and Pettigrove (2022) further elucidate that:

> A virtue is an excellent trait of character. It is a disposition, well entrenched in its possessor—something that, as we say, goes all the way

> down, unlike a habit such as being a tea-drinker—to notice, expect, value, feel, desire, choose, act, and react in certain characteristic ways. To possess a virtue is to be a certain sort of person with a certain complex mindset (Hursthouse and Pettigrove 2012).

Thus, it may be said that a virtuous person is someone who does not only perform morally good actions but repetitively does so.

In general, virtue ethicists follow the principle that "what one is, takes precedence over what one does" (Wallach and Allen 2009, 118). This means that possessing good character traits already provides a virtuous person with a kind of disposition or guide on how to behave aptly in a given situation. So, having certain virtues would, in a way, deter individuals from acting immorally.

For Aristotle, there are two classes of virtues: (1) intellectual and (2) moral (Evangelista and Mabaquiao 2020, 91). Although there are different kinds of excellence that correspond to these two virtues, an individual's reasoning capabilities must nevertheless be utilized to excel in both thinking and acting (Gensler 2011, 141). Thus, it is also important to know that there are varying ways of how intellectual and moral virtues can be acquired. Evangelista and Mabaquiao (2020, 91) clarify that intellectual virtues may be developed through instruction. On the other hand, moral virtues could be gained via habit formation. They further hold that intellectual virtues equip individuals with the capacity to think reasonably, while moral virtues allow one to manage their emotions and desires in a rational manner. So, in relation to the task of fostering AMAs based on hybrid routes, perhaps these characterizations could provide some sort of scaffolding.

Similar to Wallach and Allen's (2009, 117) idea that the engineering process usually begins with top-down processes, it appears that intellectual virtues may be given to AMAs directly. In contrast, moral virtues more closely relate to learning and habit formation, resembling bottom-up paths. Thus, the fusion of these two apparently parallels the synthesis between top-down and bottom-up approaches. Citing Aristotle's differentiation between intellectual and moral virtues, Wallach and Allen (2009) explain that…

> … moral virtues are distinct from practical wisdom and intellectual virtues. Aristotle thought that the intellectual virtues could be taught, whereas the moral virtues had to be learned through habit and practice. This suggests that a different approach might be needed for different virtues if they are to be implemented in AMAs (Wallach and Allen 2009, 119).

So, top-down approaches align with the ways to achieve intellectual virtues as they are translatable into rule-based precepts, which entails that they may be embedded into artifacts. Bottom-up techniques, on the other hand, appear to coincide with processes dedicated to acquiring moral virtues.

In view of the characterizations above, one may argue that it is quite possible to incorporate the Aristotelian notion of virtues into AMAs, specifically by means of employing artificial neural networks as suggested by Wallach and Allen (2009). As

promising as this may seem, however, hybrid-designed artifacts appear to be prone to the complications related to Hume's Guillotine, especially if one cites the worries that this problem has posed for both top-down and bottom-up methods. Consider, for instance, how Aristotelian AMAs would have to be realized.

In order to account for their intellectual virtues, hybrid-based artifacts would have to be imparted with a set of good character traits from the start. Practically, this means that human designers would first have to embed or engineer pre-selected virtues into AMAs. The assumption here is that the behaviors of these systems would be regulated by the virtues given to them, which may include integrity, empathy, and the like. So, for example, if an AMA were to be accorded with, say, the virtue of courage, one could imagine that it would always be guided by this in the face of moral dilemmas. To some extent, this virtue would serve as a default trait in a connectionist-based agent before it gets updated, if not replaced, via its bottom-up capabilities.

However, the task of engineering default virtues into AMAs arguably resembles the moral non-descriptivist strategy to address Hume's Guillotine. Specifically, this approach is quite reminiscent of prescriptivism. Recall that the latter prescribes the inclusion of evaluative premises to a set of descriptive ones to go past Hume's Guillotine (Boyles 2022, 178). Supposedly, the addition of an evaluative premise would consequently allow the inference of an evaluative conclusion in a given moral argument, enabling hybrid-designed AMAs (i.e., equipped with pre-set virtues) to conjure apparent moral judgments.

Unfortunately, prescriptivism is not totally free of philosophical worries, as highlighted earlier. Recall that, taken as a moral framework, it is devoid of any ethical value as it is quite possible to develop systems that have no moral import (Gensler 2011, 57). In fact, computer algorithms have been identified as prime examples of systems without any moral worth. In addition, prescriptivism grants the in-principle idea in Hume's Guillotine. Thus, advocates of this theory just recourse to ferreting out the hidden evaluative statements in a given moral argument (Joaquin 2012, 55-56). The consequence of this is that the suggestion of giving artifacts default virtues seems to be a losing proposition since their supposed moral judgments would be devoid of any moral value.

In a way, it may be argued that employing the strategy of imparting default virtues into artifacts is nothing different from the top-down path of modeling AMAs. This is because supplanting programmed ethical principles with default virtues practically amounts to the same thing. Hume's Guillotine seems to be at work in hybrid design strategies as well.

Similar to the proposal of programming AMAs with moral precepts, choosing to give Aristotelian virtues into intelligent systems appears to yield certain biases. Recall that there is no decisive way of assessing and selecting which specific moral framework is the best one among the many live candidates (Boyles 2021, 123). This entails that encoding ethical theories into artifacts either leads to some form of relativism or an arbitrary assignment of values (i.e., choosing one theory over another, but without any good moral grounds). Perhaps the same thing may be said about the proposal of conferring default Aristotelian virtues into AIs, as in the first place, there are many other forms of ideal character traits.

It must be highlighted that there are various theories that fall under virtue ethics, each of which espouses contrastive characterizations of good character traits. Among these kinds of frameworks are Aristotelian, Buddhist, and Confucian theories, as well as the various theories of distributive justice and care ethics (Evangelista and Mabaquiao 2020, 89). In light of such, it may be said that giving a default set of Aristotelian virtues into AMAs does not seem to be an elementary task, since there is no definite way of proving that it is the most appropriate one among the rest.

There appears to be no conclusive reason why a specific virtue ethical theory must be preferred over others. Comparably, this is very reminiscent of previous studies, like that of Lara and Deckers (2020, 277), which have already underscored the issues in determining which ethical theory should be imparted to AI systems. So, it seems that efforts to develop AMAs by means of the prescriptivist route are not that inviting.

Furthermore, it may be argued that the proposal of instilling AMAs with default virtues could be further correlated with the moral descriptivist efforts to solve Hume's Guillotine. This is because the said kind of approach, like that of Searle (1964), still employs concealed ought-premises, so that an evaluative conclusion may be inferred from a set of factual propositions. However, it has been demonstrated before that, given its employment of hidden evaluative premises, it appears that there is no practical difference between the moral descriptivist strategies and prescriptivist ways of bypassing Hume's Guillotine (Boyles 2021, 124).

Apparently, both prescriptivism and the moral descriptivist approach still make use of additional evaluative premises to permit the derivation of a moral judgment in a given argument. Perhaps the only difference between the two is that the former openly admits that evaluative claims are assumed in every moral argument, allowing one to expose them for evaluation purposes. If this is the case, however, then all the arguments proffered above against prescriptivism would arguably hold for moral descriptivism as well. The further implication of this is that Hume's Guillotine would remain as a drawback for any moral descriptivist approach to fashion hybrid-designed AMAs.

Moreover, apart from trying to address the problem of why the Aristotelian framework should be preferred over other virtue ethical theories, it must be pointed out that there are certain internal issues within Aristotle's concept of virtues that have already been raised in the past. For one, Aristotelian virtue may be understood as the midpoint or mean of two extremes—excess and deficiency (Gensler 2011, 150). So, for instance, courage could be deemed as the midpoint between foolhardiness and cowardice. However, Evangelista and Mabaquiao (2020, 93) explain that Aristotle's notion of the mean seems to be quite relative from one moral agent to another.

If the Aristotelian concept of the mean is relative, then the decision to consider something to be deficient, excessive, or moderate would largely depend from person to person. So, for instance, going to war could be an expression of courage for a disciplined soldier, but this very gesture may be construed as an act of recklessness for anyone without decent military training (Evangelista and Mabaquiao 2020, 94). In relation to the suggestion of imparting virtues into AMAs, this potentially becomes a problem as an evaluative judgment is needed to identify if something is a mean or not.

Unfortunately, there appears to be no reasonable basis for saying that a particular default virtue is, in fact, an appropriate trait to be imparted to a specific type of hybrid-

based AMA, while its correlates within a spectrum are not. The decision to identify something as an initial virtue for someone seems to require a reasonable basis for doing so, but this is tantamount to asking for a solution to a longstanding problem in ethics. Thus, there is some risk that the said endeavor would lead to an arbitrary assignment of values again.

Here, it should also be pointed out that other virtue ethical theories do not appear to be immune from the same worries mentioned above. For example, within the Confucian ethical framework, there is no definitive consensus on the list of key virtues. Evangelista and Mabaquiao (2020, 104) further explain that:

> …Confucius also spoke of basic or primary virtues, often called the five cardinal virtues of Confucianism, on which other virtues are based. It is widely recognized that foremost of these virtues are ren (or jen) and li. Scholars, however, are not in agreement with regard to the three other virtues that should also be in the list of cardinal virtues.

Thus, if there is no clear agreement on which key virtues must be considered, then there seems to be no conclusive grounds on why Confucianism, if not any other theory that advances its own set of virtues, should be preferred. Reminiscent of the suggestion of encoding AMAs with a set of moral principles, the last thing that anyone would want is to again become partial towards a particular ethical framework with no good moral grounds.

Additionally, Wallach and Allen (2009, 118-119) openly admit that there are varying opinions on the definitive number of fundamental virtues. For one, they state that, according to Plato, there are four cardinal virtues (viz., wisdom, courage, moderation, and justice). Contrariwise, they hold that the French thinker André Comte-Sponville advanced 18 types of good character traits. Thus, there appears to be no consensus on which conceptualization of virtues one must adhere to. This problem again is a normative issue in itself.

If one accepts the aforementioned objections against the idea of giving default virtues to artifacts, it may be said that this somehow further cashes out other worries put forward against the said AMA technique. Consider, for instance, Wallach and Allen's pronouncements regarding the latter. In particular, they state the following:

> The task of programming virtues into a computational system runs into problems similar to those of the rule-based approaches: conflicts between virtues, incomplete lists of virtues, and especially difficulties with definitions. Virtues affect how people deliberate and how they motivate their actions, but an explicit description of the relevant virtue rarely occurs in the content of the deliberation (Wallach and Allen 2009, 119-120).

Perhaps one reason why providing AMAs with virtues has been considered problematic is that a rational explanation is needed to justify certain key decisions in the AMA design process. Unfortunately, however, there is none, especially if one takes into account the problem cited by Hume (1739/1964).

Apart from the worries presented above, it might be argued that one way of

overcoming these is by imparting all kinds of virtues into AMAs. So, instead of zeroing in on one virtue ethical theory, perhaps the multitude of frameworks could aid them in enacting morally praiseworthy behavior. Ingenious as this proposal may seem, there are foreseeable obstacles to its implementation.

It must be underscored that the suggestion of embedding hybrid-based technologies with various kinds of virtues from different ethical theories is reminiscent of the idea of programming AMAs with numerous moral frameworks. As previously demonstrated, a problem with the latter strategy is that there seem to be no reasonable grounds for saying that one theory is better than another (Boyles 2022, 182-183). Thus, either this endeavor would again yield a problem of circularity if not an arbitrary assignment of values once more.

For example, consider an AMA that has been conferred with multiple virtues coming from diverse forms of ethical theories. So, among the said kinds of frameworks are Confucian, Aristotelian, and Buddhist ethics, as well as the ethics of care and the different theories of distributive justice (Evangelista and Mabaquiao 2020, 89). Since there are various types of virtues emanating from these theories, it begs the question which among them should an AMA favor, say, in a specific circumstance. If one admits that there is no decisive way of choosing which ethical theory is the most apt among the many possible candidates, then the said artifact would have no solid grounds for realizing one among the many others.

Furthermore, it does not help that some of these virtue ethical frameworks seem to be in direct opposition to each other. For instance, endorsers of the ethics of care have criticized the kind of good character traits espoused by the Aristotelian model (Evangelista and Mabaquiao 2020, 131). Specifically, one of their objections is that the latter promotes virtues that are too masculine.

Considering their notable differences, it is not that clear how to reconcile the various virtue ethical theories. Thus, it is also not that apparent how they could be concurrently implemented, entailing that a verdict has to be made first on which among the frameworks should be followed. This decision, however, would require an evaluative judgment.

Here, it must be reiterated and underscored that, as regards the top-down approach of incorporating virtues into AMAs, one way that Hume's Guillotine becomes more apparent is when an artifact is embedded with all the different frameworks of virtue ethics. This is because such kind of system would have no clear way of deciding which among the several variants of virtue ethical theories, if not which specific character trait, should it espouse (i.e., a kind of evaluative judgment) in relation to a given moral dilemma presented as a set of facts in the world. Note that the idea of programming AMAs with numerous kinds of virtue ethical frameworks seems to be the more prudent approach as compared to randomly favoring one without any good moral basis. In a way, this also coincides and goes back to Wallach and Allen's (2009, 119-120) claim that human designers programming AMAs with virtues would likely yield to issues that are resonant with those stemming from rule-based ethical approaches.

For Wallach and Allen (2009, 120), virtues are inherently made up of intricate patterns of desires and motivations. In underscoring the complexities surrounding the proposal of human developers directly programming virtues into AMAs, Wallach and Allen (2009, 120) further provide the following anecdote:

> A particular virtue—for example, being kind—can affect almost any activity a person engages in; it has system-wide effects. An artificial agent applying virtues in a top-down fashion would need to have considerable knowledge of psychology to figure out how to apply them in a given situation. For instance, what should one do when an action seems both to apply and to violate a virtue? Imagine that you—or your (ro)bot—have been asked by two people for a favor, but you can only help one of them. The other will perceive your rejection as unkind. One might feel that being unkind is unacceptable, but how does one determine which party's request to honor? A virtue-based AMA, like its deontological cohorts, could get stuck in endless looping when checking if its actions are congruent with the prescribed virtues, then reflecting on the checking, and so on.

Hence, relying on top-down methodologies to impart virtues into AMAs does not appear to be a straightforward exercise, since the resulting AIs would have no means of adjudicating which virtue ethical framework is most appropriate in a given situation in view of Hume's Guillotine. Nevertheless, even if one questions the foregoing arguments against the top-down methods of accounting for virtues, it must be recalled that connectionist models also employ bottom-up techniques, which arguably have their own set of problems.

In terms of the suggestion of using bottom-up strategies to further cultivate the existing virtues of AMAs, if not allow them to acquire new ones, Hume's Guillotine appears to be a persistent issue for this as well. Recall that this approach relies on the facts or data harnessed by an artifact in its present setting, allowing it to update or supersede its initial virtues. In consonance with the objections leveled against AMAs that employ bottom-up techniques (Boyles 2024, 17-20), this scheme ostensibly falls prey to the same form of issues as well.

Considering that bottom-up design methodologies do not adhere to the suggestion of encoding ethical notions into AMAs, it has been demonstrated before that these approaches could only make use of descriptivism (Boyles 2024, 19). However, as discussed earlier, this solution to Hume's Guillotine is not without problems. For one, it must be remembered that Searle's (1964) purported strategy still employs concealed evaluative claims. Moreover, apart from promise-making, there is no clarity on how this move would be applicable in other contexts and cases (Boyles 2021, 119).

Navigating through the complexities of dynamic moral environments would foreseeably present issues for hybrid-designed AMAs as well, especially since these artifacts are bound to encounter numerous kinds of virtue ethical theories. So, in a situation wherein the said artifacts are to choose between, say, Aristotelian and care ethics (i.e., which virtues among these frameworks should they imbibe), they would be left with no good moral basis to reasonably resolve this conundrum. As discussed above, there would presumably be issues in fusing these two since the kinds of virtues that the Aristotelian model promotes have been censured by the supporters of care ethics for being overly male-oriented.

It also does not help that the ethics of care, in principle, goes against most of the universalist and rationalist ethical frameworks (Evangelista and Mabaquiao 2020,

131). Advocates of care ethics contend that theories, like utilitarianism and the Kantian categorical imperative, are removed from actual human experience by treating moral actors as individuals who are isolated from each other. So, if this is the case, it is not clear how systems developed via hybrid means could adjudicate between these disparate ethical theories if they could only work with the facts from their environment. It seems that a resolution to Hume's Guillotine is needed to settle this problem, but there is none as of the moment.

In addition, similar to the problem cited earlier for the top-down aspect of connectionist models, recall that certain virtue ethical frameworks have internal issues among themselves. Another example to demonstrate this setback may be drawn from the many theories of distributive justice, all of which could potentially nurture the virtue of justice (Evangelista and Mabaquiao 2020, 125-130). It should be pointed out that some of these theories are also in conflict with one another, like the incompatible assumptions of, say, capitalist justice with socialism. So, if AMAs are to encounter these clashing frameworks, it is not evident which among these they should acquire, if not be guided by, when faced with particular moral dilemmas.

Considering the idea that hybrid-based AMAs would only be left to work with the facts from their current environment (i.e., in relation to their bottom-up aspect), there appears to be no conclusive basis on how to arbitrate between the numerous theories of distributive justice. Ultimately, this stems from the notion that evaluative statements may never be validly drawn from factual claims alone (Frankena 2006, 49). So, the said artifacts would seemingly be left with no concrete mechanism to decide which theory is the best one without a resolution to Hume's (1739/1964) worries.

Lastly, if AMAs are to hone their default virtues, if not acquire new ones, as they interact with their environment, note that this endeavor arguably goes against some of the basic assumptions in ethics as a branch of philosophy. Recall that ethics is mainly concerned with "what ought to be the case" and not with how the present world is like (Boyles 2024, 19). So, the idea of hybrid-designed AMAs using bottom-up strategies to favor a particular virtue ethical theory, because of, say, its predominance in society seems to run counter to the very motivations behind ethics.

If one grants that it is possible to re-appropriate the arguments purported against top-down and bottom-up methods of constructing AMAs (i.e., specifically those relating to Hume's Guillotine), then the moral reasoning abilities of artifacts built through hybrid means apparently become suspect as well. In a way, this somehow closes the loop in terms of employing the three AMA design strategies discussed by Wallach and Allen (2009). It looks like the perils caused by Hume's Guillotine for both the top-down and bottom-up techniques, taken independently, cut across the hybrid, combined form as well.

CONCLUSION

In this article, some preliminary groundwork has been laid to defend the claim that hybrid-built AMAs are also prone to Hume's Guillotine. This is because the arguments based on the latter, which have been previously utilized against top-down and bottom-up methods, could arguably be re-appropriated and redeployed to show

that the moral reasoning faculty of artifacts created by means of hybrid methods is suspect as well. Their ability to conjure moral judgments appears greatly affected by the said philosophical problem, making the hybrid route an impractical approach to fashion moral AIs.

In the future, however, it would be interesting to explore other possible realizations of the hybrid method. For one, the present work mainly focuses on Wallach and Allen's (2009, 122-123) suggestion to employ connectionism in developing hybrid AMAs. Thus, if other more robust hybrid strategies are introduced over the course of time, then the initial conjectures presented here should also be revisited and reevaluated (i.e., if they would remain true for prospective hybrid techniques).

Furthermore, it must be pointed out that Hume's Guillotine has only been cited as an argument for one form of AI system, specifically the non-human-based type. Chalmers (2010) elucidates that there are actually two ways of developing AIs, namely: (1) *human-based* and (2) *non-human-based strategies*. Human-based techniques, on the one hand, are those that upgrade and extend the biological makeup of human beings (Chalmers 2010, 32-33). On the other hand, non-human-based AI routes are approaches that build non-human systems from scratch, including the creation of learning machines, computer programs, and the like. Thus, following the said differentiation, it is also intriguing if Hume's (1739/1964) worries would also be a problem for human-based AIs.

NOTES

1. As presented later in this paper, Searle (1964) believes otherwise, arguing that there is a way to deduce evaluative statements from descriptive propositions.
2. For context, see Boulanin and Verbruggen (2017, 7).

REFERENCES

Anderson, Michael and Susan Leigh Anderson. "Machine Ethics: Creating an Ethical Intelligent Agent." *AI Magazine* 28, 4 (2007): 15-26.

Athanassoulis, Nafsika. "Virtue Ethics." In *Internet Encyclopedia of Philosophy*. Edited by James Fieser and Bradley Dowden, 2010. Available at https://iep.utm.edu/virtue/. Accessed: 12 February 2024.

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

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

Boyles, Robert James M. "Extending the Is-ought Problem to Top-down Artificial Moral Agents." *Symposion* 9, 2 (2022): 171-189.

Boyles, Robert James M. "Can't Bottom-up Artificial Moral Agents Make Moral Judgements?" *Filosofija. Sociologija* 35, 1 (2024): 14-22.

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

Carter, Matt. *Minds and Computers: An Introduction to the Philosophy of Artificial Intelligence.* Edinburgh: Edinburgh University Press, 2007.

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

Chalmers, David. "The Singularity: A Philosophical Analysis." *Journal of Consciousness Studies* 17, 9-10 (2010): 7-65.

Evangelista, Francis Julius N. and Napoleon M. Mabaquiao Jr. *Ethics: Theories and Applications*. Mandaluyong: Anvil Publishing, Inc., 2020.

Frankena, William K. "The Naturalistic Fallacy." In *Arguing about Metaethics*. Edited by Andrew Fisher and Simon Kirchin, 47-58. Oxon: Routledge, 2006.

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

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

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

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

Hursthouse, Rosalind and Glen Pettigrove. *Virtue Ethics.* In *The Stanford Encyclopedia of Philosophy*. Edited by Edward N. Zalta and Uri Nodelman, 2022. Available at https://plato.stanford.edu/entries/ethics-virtue/. Accessed: 12 February 2024.

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

Kim, Shin. Moral Realism. In *Internet Encyclopedia of Philosophy*. Edited by James Fieser and Bradley Dowden, 2006. Available at http://www.iep.utm.edu/moralrea. Accessed: 20 January 2024.

Kraut, Richard. "Aristotle's Ethics." In *The Stanford Encyclopedia of Philosophy*. Edited by Edward N. Zalta and Uri Nodelman, 2022. Available at https://plato.stanford.edu/entries/aristotle-ethics/. Accessed: 12 February 2024.

Lara, Francisco and Jan Deckers. "Artificial Intelligence as a Socratic Assistant for Moral Enhancement." *Neuroethics* 13 (2020): 275-287.

Mabaquiao, Napoleon, Jr. *Mind, Science and Computation.* Manila: Vibal Publishing House, Inc, 2012.

Mohanty, Jitendra Nath. "Philosophical Description and Descriptive Philosophy." *Research in Phenomenology* 14, 1 (1984): 35-55.

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

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

Searle, John R. "How to Derive 'Ought' from 'Is' Revisited. In *Revisiting Searle on Deriving 'Ought' from 'Is.'* Edited by Paolo Di Lucia and Edoardo Fittipaldi, 3-16. Cham: Palgrave Macmillan, 2021.

Siau, Keng and Weiyu Wang. "Artificial Intelligence (AI) Ethics: Ethics of AI and Ethical AI." *Journal of Database Management* 31, 2 (2020): 74-87.

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

Zaidi, Leah. "The Only Three Trends That Matter: A Minimum Specification for Future-Proofing." *Journal of Futures Studies* 25, 2 (2020): 95-102.

S