Ximian Xu

Article

DOI: https://doi.org/10.56315/PSCF9-24Xu

How Virtuous Can Artificial Intelligence Become? Exploring Artificial Moral Advisor in Light of the Thomistic Idea of Virtue

Ximian Xu

Is artificial intelligence (AI) virtuous? Can AI become as virtuous as humans? This article is intended to explore these questions with a focus on artificial moral advisor (AMA). AMA is a proposal for the future application of AI to human moral life. Hence, this article will be dedicated to the theoretical analysis of the issues surrounding AMA. Socratic AMA will be the specific object of study. To this end, this article will examine whether or not the Socratic AMA can enhance human virtuous life through exploration of Thomas Aquinas's theology of virtue. It will argue that the Socratic AMA is not as virtuous as humans insofar as it lacks both the subject and ultimate end of virtue and is characterized as measurable. Nevertheless, the Socratic AMA can be considered to be embedded with delegated virtues and is expected to assist humans in their cultivation of virtue in certain contexts by reason of both its capacity to gather voluminous information and its tremendous processing power.

Keywords: artificial moral advisor, Thomas Aquinas, virtue, virtuous AI, teleology, technology ethics

A virtue refers to a trait or excellence that gives birth to right actions leading to flourishing. A virtue may grow or wither across time. In the sphere of ethics, virtues that grow will foster morality, but withering virtues are coupled with moral vices. This definition of virtue is deliberately made in a broad sense insofar as some scholars suggest that technology ethics should eschew an anthropocentric construal of the moral status of smart machines.[1] With considerable reservations about such a claim, I have demonstrated elsewhere, from a theological perspective, how the moral status of smart machines can be construed in the light of human morality as a starting point without running the risk of anthropocentrism.[2] That said, for the purpose of this article, a broad definition of virtue can conduce to exploration of the distinction and yet connection between human virtues and so-called artificial intelligence (AI) virtues.

The idea of virtue features in recent studies on technology ethics. A typical example of these studies is Shannon Vallor's elaborate account of contributions that virtue ethics can make to technomoral futures. That is, virtue ethics can help articulate a framework for diverse ethical narratives in relation to technological society.[3] The idea of virtue also draws much attention in recent discussions of AI ethics. Scholars such as Robert Sparrow attend to the importance of virtue ethics for the evaluation of the relationship between humans and robots as well as AI.[4] Others explore the way in

Ximian (Simeon) Xu *(PhD, University of Edinburgh) is Duncan Forrester Fellow at the Institute for Advanced Studies in the Humanities and the Centre for Theology and Public Issues at the University of Edinburgh, Edinburgh, Scotland.*

which virtue ethics can help individuals and governments use AI-powered technologies.[5]

This article is intended to address the questions of whether and how AI itself can be considered virtuous, through the lens of Thomas Aquinas's theology of virtue. Given the limited space of the article, we cannot go into a comprehensive consideration of the details related to virtuous AI. I will instead put the spotlight on the artificial moral advisor (AMA) in relation to human virtues and moral growth. Some AI ethicists are convinced that AI is expected to advise humans on moral life, including, among others, moral decision making, moral deliberation, and moral judgment. Good moral advice can be called virtuous insofar as it leads to right actions and facilitates moral flourishing. Virtuous advice cannot be made without a virtuous entity precisely because virtue must be a trait or excellence within advisors. For example, a virtuous human person is capable of offering moral advice on making moral decisions in accordance with virtuous principles while confronting a moral dilemma. If AI can, *on its own*, generate virtuous advice to make human moral life flourish, then it can be argued that AI itself is virtuous.

It is worth pausing here to clarify that AMA is a proposal for the future application of AI to human moral growth. This means that there is no existing AMA system that we can use to generate moral advice; so, there are no real-life examples available for analysis. For this reason, this article is dedicated to the theoretical analysis of the issues surrounding AMA in the relevant literature.

I will argue that despite gathering voluminous information and having tremendous processing power, AI is not as virtuous as humans because it lacks both the subject and ultimate end of virtue and is characterized as measurable. That said, the AMA can be considered to be embedded with delegated virtues; consequently, it is able to facilitate the cultivation of human virtues and moral growth. In what follows, I will first elucidate how the AMA operates to generate moral advice with a focus on the Socratic AMA, which is designed to question human agents when they are confronted with moral dilemmas. Second, I will unpack Aquinas's view of virtue, bringing to light the particularities of the human being as the virtuous agent. Finally, I will flesh out the sense in which the AMA may possess delegated virtues and may be able to promote human moral life in certain contexts.

AMA's Advice and Its Nature

The AMA was proposed against the backdrop of wide debates over biomedical moral enhancement. The proponents of biomedical moral enhancement maintain that biomedical technologies can significantly and rapidly foster human morality. All the more provocative is the claim that extensive biomedical moral enhancement is urgent and inescapable. The reason is that, as per Ingmar Persson and Julian Savulescu's claim, the exponential growth of knowledge and the advancement of science and technology can empower any individual to perform gravely immoral actions.[6] The development of human moral capacities lags behind technoscientific progress, and biomedical moral enhancement can effectively bridge the gap between morality and scientific knowledge so as to safeguard against any potential catastrophes caused by technoscientific advancements.

The proposal of biomedical moral enhancement has sparked vigorous debate. Some argue that this proposal gives us a misleading impression that morality can be biologized.[7] The proponents of AMA criticize that the effects of biomedical interventions are short term in comparison with the moral advice offered by AI systems.[8] Among these proponents, Francisco Lara and Jan Deckers raise the idea of the Socratic AMA, which is designed to assist humans in moral decision making by asking certain questions when humans are in moral dilemmas.[9] Compared with other proposals of AMA (which I will touch on later), the Socratic AMA seems to be the most viable approach to the development of AMA. For this reason, I will concentrate on the idea of the Socratic AMA before exploring the question of virtuous AI.

Lara and Deckers's proposal of the Socratic AMA aims to achieve AI-powered moral enhancement that can be made "more rapidly and successfully than traditional methods, and with fewer risks and controversies than bio-enhancement."[10] They begin with the critique of the two proposals of AMA: exhaustive and auxiliary enhancement.

Lara and Deckers take issue with the proposal of exhaustive enhancement. This proposal suggests that autonomous artificial agents can be created to make moral decisions in place of humans insofar as these AI-powered agents are more capable than human agents to deal with moral issues. Lara and Deckers contest the following idea:

> The essential aspect of this proposal is that all human participants, including the designer, would

> be expected to take a passive role after the original programming had been completed.[11]

This criticism targets Blay Whitby's claim that AI technology can create a moral machine which generates moral decisions *for* humans. First, Whitby suggests that

> the general acceptance of machine-generated moral judgments is not incompatible with either or both of these assertions. Firstly it is necessary to make the banal observation that the use of tools—from flint axes to computers—is also an essential part of what it is to be human.[12]

In this way, exhaustive enhancement by the AMA becomes constitutive of the human being as moral agent. From this, it follows that AI systems are always embedded within human social structures and, consequently, within "a network of authority."[13] As such, the decision made by the AMA is morally authoritative for human agents. For Lara and Deckers, however, exhaustive enhancement deprives humans of decision making and arrests human agenthood in that it is the autonomous machine, rather than humans who perform their moral agency, that makes moral decisions.

The second proposal that Lara and Deckers criticize is called auxiliary enhancement, advocated by Julian Savulescu, Hannah Maslen, and Alberto Giubilini.[14] In contrast to exhaustive enhancement, auxiliary enhancement refrains from making moral decisions on behalf of humans in order not to render humans *passive* moral agents. Savulescu and Maslen contend that

> the moral AI would monitor physical and environmental factors that affect moral decision making, would identify and make agents aware of their biases, and would advise agents on the right course of action, based on the agent's moral values. In being tailored to the agent, the moral AI would not only preserve pluralism of moral values but would also enhance the agent's autonomy by prompting reflection and by helping him overcome his natural psychological limitations.[15]

Even though the proposal of auxiliary enhancement pays due attention to human autonomy in moral enhancement, Lara and Deckers disapprove of this proposal for two reasons. Firstly, human agents are, by and large, passive in that humans still depend upon the AMA and are ignorant of the connection between their moral values and the AMA's moral decisions. Secondly, it is taken for granted that the AMA can provide moral advice according to programmed moral values. Yet, such fixed moral values may undermine human reflection when humans are making moral decisions.[16]

Having brought to light the flaws of the proposals of both exhaustive and auxiliary enhancement, Lara and Deckers flesh out the idea of the Socratic AMA. The rationale behind the Socratic AMA is rooted in Socrates's pedagogy.

> Socrates always presents himself and acts as a mere assistant who aims to refute the definitions he receives from his interlocutors. He … is like a midwife who only helps the other to give birth to his own knowledge. In our case, this knowledge would not reveal any hidden or common-sense truth, but consist in a moral judgment that was formed by applying conditions of empirical, logical and ethical rigor to one's beliefs. The agent should always have a privileged place, should always provide the first solution in a significant conflict, which is then submitted to staged scrutiny so that the machine, like Socrates, may ask relevant questions and reveal potential failures in the argumentation.[17]

Three observations can be made in reference to this passage. First, as a "midwife," the Socratic AMA is designed to influence human agents rather than to generate a decision forthrightly. In spite of their differing methodologies, both exhaustive and auxiliary enhancement are intended to produce moral decisions for human agents. By contrast, Lara and Deckers's "emphasis is placed on the formative role of the machine for the agent, rather than on the result."[18] Although the Socratic AMA provides moral advice, it is always the human agent who makes moral decisions. To this extent, the Socratic AMA is more capable than the previous two AMAs of enabling humans to foster their own virtues.

Second, the Socratic AMA's task is to assist human agents in the formation of moral judgments, which are characterized by "empirical, logical, and ethical rigor." Instead of using biomedical technologies to intervene in human moral judgments, the Socratic AMA is an apparatus through which to advise humans on gaining proper knowledge for moral decision making. Lara and Deckers suggest six functions of the Socratic AMA:

1. invalidate moral judgments by empirical premises,
2. clarify concepts for moral judgments,
3. pinpoint the logical deficiencies of moral judgments,

4. test the ethical plausibility of moral judgments,
5. remind humans of personal limitations, and
6. advise on the execution of moral decisions.[19]

All these functions rely on AI's rapid accumulation of voluminous information and tremendous processing system. For example, AI's voluminous data bank collects a huge number of moral concepts for moral judgments, a quantity that is far beyond human moral knowledge. Coupled with its tremendous processing ability, the AMA can identify, select, and clarify moral notions in nanoseconds for humans when making moral judgments. A second example is its ability to track human personal limitations. For instance, an AI-powered device (e.g., Apple watch) can monitor the human agent's sleep, and it has been discovered that sleep deprivation can increase intergroup bias.[20] The AMA can monitor the human agent's quality of sleep (e.g., average deep sleep, average light sleep, and average awake time) as a reminder of the potential to make wrong judgments on moral matters.

Third, the Socratic AMA operates by posing questions to human agents. It raises some questions when human agents are wrestling with moral dilemmas, such as, "Are you aware that both assertions are contradictory?"[21] These questions aim to remind human agents of errors (e.g., logical contradiction) lurking around moral reasoning and judgment. By doing so, it is expected that the human agent's "motivations and emotional dispositions" can be changed to make good moral judgments and decisions, leading to moral enhancement. Francisco Lara argues elsewhere:

> The virtual assistant will make the person aware of their possible errors and *they will feel motivated*, where appropriate, either to respond as to why they believe they are not errors or to avoid them with revised positions. It is foreseeable that, with this dialectical training, the person will acquire the capacity to make decisions *critically and self-sufficiently* in the future.[22]

Moral questioning signals the nature of the Socratic AMA. That is, through questioning human agents, the Socratic AMA is conducive to the formation of human motivations for making good moral decisions and performing good actions accordingly.

During moral questioning, the human agent's mental activities concerning morality can be nudged in the direction of moral good to give birth to good actions. As such, some moral traits and excellence must be cultivated within human agents to elicit virtuous actions. To put it differently, the Socratic AMA asks moral questions with the purpose of cultivating the virtues in human agents. From this, we can infer that the Socratic AMA's questions must be somewhat characterized as virtuous or that the Socratic AMA's questions must have virtues as their *telos*. Either of the two inferences may yield the corollary that the Socratic AMA is, to a certain degree, virtuous precisely because moral questioning per se as virtuous action should be derived from a virtuous entity.

It is beyond doubt that the Socratic AMA can somewhat assist in the cultivation of human virtues by posing questions related both to moral decision making and to virtuous actions. The moral questions generated by the Socratic AMA can alert human agents to the factors that must be taken into consideration while making moral decisions. However, this corollary raises the question of to what extent and in what sense the Socratic AMA is considered to be virtuous. To complicate this question further, recent studies on artificial general intelligence (human-level AI) may induce us to draw a hasty conclusion that the Socratic AMA is as virtuous as humans. In what follows, Aquinas's theology of virtue will provide us with a lens through which to spell out the meaning of being virtuous in relation to AI.

Being Virtuous

Moral theology occupies a crucial place in Aquinas's theological system. As an illustration of this, the elaboration on moral matters makes up the majority of the second part (*Secunda pars*) of *Summa theologiae*.[23] Virtue is a subject matter of Aquinas's moral theology. As he claims, "all moral matters are reduced to the consideration of the virtues."[24] For this reason, his view of virtue can be a conceptual tool for theological analysis of morality in relation to AI, more so when exploring whether the Socratic AMA could be considered virtuous.

A virtue is, for Aquinas, a habit (*habitus*) within an agent. The Latin word "*habitus*" literally means "to have," which can connote either "to have something" or "to have relation to itself or something else." Aquinas's use of *habitus* in relation to virtue is associated with the latter sense: "habit is a disposition whereby that which is disposed is disposed well or ill, and this, either in regard to itself or in regard to another … Wherefore … habit is a quality."[25] As a

quality, a habit is durable within the agent.[26] To put it in Bonnie Kent's words, "habits grow to be, or are habituated as, a 'second nature'" for the agent.[27]

As a quality of the human agent, "habit is that whereby we act when we will."[28] A habit implies a relation to an act and "a state of potentiality" (disposition) in respect to operation.[29] If operations of the body are caused by the body's natural qualities, then the body needs not to be disposed. To this extent, such operations have nothing to do with habits. As such, habitual operations of the body can be moved only by the soul, and so the soul is the authentic subject of habits.[30] Aquinas takes a further step to elucidate that habits are in the powers of the soul insofar as "the soul is the principle of operation through its powers [*potentias*]."[31] From this it follows that a virtue as a habit consists only in a power of the soul. Aquinas argues:

> Virtue denotes a certain perfection of a power. Now a thing's perfection is considered chiefly *in regard to its end*. But the end of power is act. Wherefore power is said to be perfect, according as it is determinate to its act. Now there are some powers which of themselves are determinate to their acts; for instance, the active natural powers. And therefore these natural powers are in themselves called virtues. *But the rational powers, which are proper to man, are not determinate to one particular action, but are inclined indifferently to many: and they are determinate to acts by means of habits* … Therefore human virtues are habits.[32]

Two points are of note here: (1) having turned to the soul as the subject of virtue, Aquinas emphasizes human rational powers; and (2) virtue carries the connotation of teleology, which means that a power of the soul is perfected toward an end. These two points will be unpacked below and set a scene for clarification on the meaning of being virtuous in relation to the Socratic AMA.

Aquinas maintains that virtues can only be in the rational part of the soul. "And therefore reason, or the mind, is the proper subject of virtue."[33] He reformulates that virtue refers to "a good quality of mind by which we live righteously."[34] Aquinas moves on to elucidate the sense in which virtue is tied up with the rationality of the soul: since the "mind is chiefly called the intellect," the subject of virtue is the intellect.[35] The intellect as the subject of virtue should be understood in reference to virtue understood in a relative sense—that is, a virtue enables humans "to have the aptness to do well." On the other hand, if virtue is understood simply or absolutely (*simpliciter*), which means that a virtue enables humans to do well *actually,* then the will is the subject of virtue. This is so because the will, "a rational power," moves all other powers to act in a rational way.[36]

It is worth noting that for Aquinas a virtue cannot consist in several powers of the soul at the same time. One virtue is in one power. That said, a virtue can belong chiefly to one power but, at the same time, diffuse to the other powers in a certain order.[37] Hence, the subject of virtue is, properly speaking, the soul as a whole, not a single power of the soul. As will be discussed, the soul as the subject of virtue draws attention to the differing meanings of being virtuous in relation to the Socratic AMA and human agents.

The second point Aquinas emphasizes carries a teleological implication: a virtue perfects a power toward an act as its end. For all that rational powers are directed toward many actions, one virtue perfects one power toward one act as its end: "diversity of ends demands a diversity of virtues."[38] In Aquinas's position, an act itself is not the ultimate end of virtue. His teleological account of virtue must be read in tandem with his view of the cause of virtue. As Jeffrey Brower remarks, Aquinas characterizes agents as "always acting for ends" due to their "teleologically directed causal powers."[39] The only ultimate end is God himself, who can completely satisfy all human desires.[40] God as the ultimate end of virtue presupposes that God is the cause of infused virtues that are necessary for the perfection of the soul in relation to things exceeding human nature.[41] No acquired virtue can, in its own right, dispose humans to the ultimate end. Rather, it is theological virtues—that is, faith, hope, and charity—that are infused by God and dispose humans toward God.[42] Furthermore, God infuses cardinal virtues—that is, prudence, justice, fortitude, and temperance—that are "corresponding, in due proportion, to the theological virtues." In this way, the soul is effectively perfected toward the ultimate end.[43]

As infused virtues, theological virtues are formed within humans without reference to human action. Given this, the prominent feature of theological virtues is the *immeasurability* of theological virtues.

> It follows that human virtue directed to the good which is defined according to the rule of human reason can be caused by human acts: inasmuch as such acts proceed from reason, by whose power

> and rule the aforesaid good is established. On the other hand, virtue which directs man to good as defined by the Divine Law, and not by human reason, cannot be caused by human acts, the principle of which is reason, but is produced in us by the Divine operation alone.[44]

Given their divine origination, theological virtues cannot be measured by human standards. Rather, as Aquinas argues elsewhere, "theological virtue has for its object the first standard itself, which is not measured by another standard."[45] While faith, hope, and charity can be measured in accordance with our condition, they cannot be measured with reference to God precisely because God, the ultimate end, is *the* rule and measure of theological virtues.[46] As will be seen, this immeasurability of theological virtues brings to light a critical distinction between virtuous human agents and the AMA in that the latter is created with *measurable mathematical models*. Furthermore, Aquinas maintains that theological virtues radically differ from and underlie both intellectual (wisdom, science, and understanding) and moral (temperance, justice, prudence, and fortitude) virtues.[47] The distinctiveness of theological virtues signals that both intellectual and moral virtues need to be oriented by theological virtues toward the ultimate end.

To summarize, Aquinas's theological account of habitual virtues shows that virtues can be habituated either by acquiring or by divine infusion. He does not restrain the idea of virtue within moral confines but rather extends it to human life as a whole. Aquinas's view of virtue is anthropocentric. As Thomas Osborne rightly notes, Aquinas concentrates on "which is proper to humans and not to other animals or bodies" while developing his theology of virtue.[48] Be that as it may, Aquinas's theology of virtue fits well into the broad concept of virtue laid out at the beginning of this article. Putting Aquinas's theology of virtue alongside the Socratic AMA, we are now turning to exploration of how virtuous AI can become, if it is considered virtuous somehow.

If the AMA Becomes Virtuous

If the Socratic AMA can advise humans on moral decision making, then moral advice that underlies questions generated by AI systems can be characterized as virtuous insofar as it directs humans toward a virtuous life. As per Aquinas's view of being virtuous, virtuous actions must be traced back to a virtuous agent. A question arises here: is the Socratic AMA as virtuous as humans? The answer to this question is entangled with the debate over whether AI can evolve to be a human-level agent. Even if the answer is negative, a further question may be brought up: how virtuous can the AMA become in its potential contribution to the flourishing of human moral life?

Some AI philosophers and ethicists claim that a human-level artificial agent will be created in the future. An example of this position is John Sullins's endorsement of the moral agency of robots. Sullins contests that robots are fully moral agents provided that they are significantly autonomous (without the direct control of others), act intentionally ("seemingly deliberate and calculated"), and behave in a way that implies moral responsibility to others.[49] By arguing so, the boundaries between human and artificial agents are obliterated.

In contradistinction to scholars like Sullins, Robert Sparrow reminds us of the discrepancy between scientific and moral truths.

> When it comes to performing a mathematical calculation or analyzing a mechanism, someone else could make "my" decision because any consideration for them is also a consideration for me and vice versa. By contrast, ethical dilemmas attach to agents in such a way that they are essentially dilemmas for particular people. The nature and role of ethical truths are correspondingly different from that of scientific truths.[50]

A major distinction drawn by Sparrow between scientific and moral truths involves the contextual variables and personal features concomitant with moral issues. Viewed in this light, AI cannot become a fully moral agent like humans because AI algorithms, as mathematical operations, cannot fully deal with moral issues related to particular people in particular contexts at particular times.[51] Considering this in the light of Aquinas's definition of virtue, it can be argued that since virtues are not habituated as a second nature of the Socratic AMA, the artificial agent per se is not disposed to virtuous actions across different circumstances and times. For this reason, the Socratic AMA cannot become virtuous in the same sense as virtuous human agents.

Aquinas's view of virtue provides another theological objection to human-level artificial agenthood and, at the same time, echoes and theologically consolidates Sparrow's stance in three respects: (1) the

subject of virtue, (2) the *telos* of virtue, and (3) the immeasurability of infused virtues. Aquinas broadly identifies the soul as the subject of virtue. Locating virtues in the rational part of the soul, he suggests that a virtue is in the mind as the proper subject.

Aquinas's connection between the mind and virtues resonates with recent studies on the link between mental activities and AI's moral status. Kenneth Einar Himma contends that inasmuch as agency is intertwined with volition, intention, belief, desire, and other mental states, AI cannot evolve to be a human-level moral agent due to the impossibility of human-level artificial consciousness.[52]

Likewise, Richard Spinello argues that

> personal actions are those consciously brought about by free will responding to the intelligible goods presented to it by reason. Only a person can make this choice for one of these goods and thereby qualify for moral agency. Moral agency requires a "free will" that is capable of voluntary action.[53]

Due to the lack of volitional actions, AI and smart machines cannot have human-level moral agency. Both Himma's and Spinello's arguments locate the human mind at the center of moral life.

Aquinas argues in a similar way that the mind—the will, in particular—moves all rational powers to enable the body to perform good actions.[54]

> Now the proper nature of a power is seen in its relation to its object … [I]f man's will is confronted with a good that exceeds its capacity, whether as regards the whole human species, such as Divine good, which transcends the limits of human nature, or as regards the individual, such as the good of one's neighbor, then does the will need virtue. And therefore such virtues as those which direct man's affections to God or to his neighbor are subjected in the will, as charity, justice, and such like.[55]

Accordingly, being virtuous implies that an entity can volitionally perform virtuous actions toward both God and humans. Recent studies have cogently demonstrated from multifaceted perspectives that the particularities of humanity, including human religious and moral nature, cannot be reduced to mathematical models or algorithms. For example, Antonio Damasio demonstrates that social emotions such as shame, embarrassment, and envy—which carry moral values—are associated with the ventral and medial aspects of the prefrontal cortex.[56]

Jobst Landgrebe and Barry Smith also underscore the uniqueness of humans as organisms vis-à-vis smart machines.

> We cannot model and engineer in a machine the human way of perceiving and interacting within a social environment. Instead, the machine must rely on sensory data, and on serial interpretation of these data in order to obtain the inputs to its algorithmic counterpart of social norms. We can imagine improvements in engineering that go beyond this sort of serial interpretation, but however far these improvements will take us, we will … still not be able to model the complex systems that enable human social behavior. Therefore, we cannot build machines that can know and apply social norms with the facility that is characteristic of human beings.[57]

Needless to say, like social norms, moral and religious norms cannot be simplified as mathematical operations. As Mark Coeckelbergh rightly argues, humans "are meaning-making, conscious, embodied, and living beings whose nature, mind, and knowledge cannot be explained away by comparisons to machines."[58] As a crucial constituent of the human being, the mind as the subject of virtue cannot be mathematized. As a result, the Socratic AMA cannot operate as a human-level virtuous agent to generate virtuous advice on human moral growth.

Aquinas's teleological account of virtue features in a theological objection to the view of Socratic AMA as a human-level virtuous agent. In discussion over virtue ethics with reference to technology as well as AI, "virtue" is often construed in a moral sense. Through exploration of Aquinas's view of virtue, especially his view of the ultimate end of virtue, it comes to be seen that virtue should be understood in a broader sense.

In Aquinas's position, virtue involves the whole human life. Being virtuous means that the human being must be disposed toward the ultimate end on all levels and in all circumstances. Within this ideal view, religion itself is a virtue. All the more important is that religion is a special, moral virtue which excels among all moral virtues in that it gives special honor to God as its end.[59] It is clear that Aquinas's view of moral virtue is broader than the notion of virtue in the present literature on technology ethics, which separates religion from morality. Aquinas's position is a good reminder to us that being virtuous must be tied to an end beyond the natural sphere where humans reside and behave morally.

It has been argued in recent literature, however, that AI can be directed toward God, the ultimate end, in the future.[60] The above analysis of the mind as the subject of virtue turns down the human-level virtuous Socratic AMA. Aquinas's teleological view of moral virtue reinforces this refutation. If the Socratic AMA is as virtuous as humans, then God as the ultimate end must be embedded within its AI system. Yet, those who endorse AI's direction toward God are oblivious to the way in which AI is made purposeful. Mihaela Constantinescu and her colleagues note that "it is not the AMA app that initiates and controls the facts leading to the human decision and action: it is the human users who decide to take benefit of the app and start using it for their own purposes."[61]

Added to this observation is that programmers and designers can endow particular *telos* to AI systems and make AMAs operate in a way to generate expected advice. As an artefact, the AMA reflects the values of programmers and designers who are involved in the creation of AI.[62] From this it follows that AI's *telos* is predicated upon human values. Sarah Lumbreras's distinction between two kinds of delegation can help us make further clarification on this point. The delegation of the first sort means that we trust in someone who shares our same values to orient our actions. The delegation of the second sort requires humans to provide "sufficiently detailed instructions, instructions that cover the full spectrum of situations that could emerge in the context of the decision."[63] The creation of AI falls into the second category of delegation. Those who are involved in the creation of AI delegate to the Socratic AMA to generate advice in certain contexts. This delegation becomes obvious when we look at the Socratic AMA's questions. The questions proposed by Lara and Deckers for the Socratic AMA include the following examples:

> Are you aware that this deduction/induction/analogy … is not valid? Do you know that this is not a common value? Are you aware that your current physical condition/environment is not the best one to make an important decision? Do you know that, in these circumstances, your decision could be best executed like this?[64]

These questions can help human agents in certain circumstances make their own moral decisions prudently after weighing up various factors. Yet notwithstanding this, the words such as "contradictory," "valid," "common value" imply that an inducing, preliminary judgment has already been made for humans. Given AI as created by design, such judgments can be considered the delegation of the second sort, being embedded with detailed instructions for asking specific questions in certain circumstances. Doubtless, AI programmers and designers can endow the Socratic AMA with a *telos* during such a delegation process. To say the least, it is widely recognized that AI can be manipulated by capitalists and technocratic elites to achieve their own purposes. Hence, it suffices to say that it is misleading to say that the Socratic AMA in itself is as virtuous as humans such that it can be well disposed to an end and even to the ultimate end. That said, it can be argued that the Socratic AMA has the potential to become virtuous by delegating virtues to AI. In this light, the Socratic AMA's *purposes* are nothing other than the *telos* of human virtues.

The third observation to the Socratic AMA as the human-level virtuous agent is a summary of, and intensifies, the above two observations. That is, Aquinas's emphasis on the *immeasurability* of divinely infused virtues underpins the claim that, unlike human agents, the Socratic AMA neither has the subject of virtue nor is teleologically virtuous. Since infused virtues—theological virtues *par excellence*—are caused by God, their measure must overstep the rules and principles established by humans. The immeasurability of divinely infused virtues echoes Sparrow's distinction between scientific and moral truths, which demonstrates the impossibility of mathematically measuring morality.

One may draw machine learning in support and argue that the Socratic AMA can be an autonomous agent who is able to ask questions when human agents are in particular contexts and moral dilemmas. There are diverse models for creating machine-learning systems. Yet all these systems are designed to perform particular tasks. As Peter Flach observes, "models lend the machine learning field diversity, but tasks and features give it unity."[65] This unity of tasks is rooted in the fact that these models, along with AI algorithms, are made according to possibilities that are foreseeable and susceptible to measurement. According to Stuart Russell,

> AI researchers simply bought into the standard model that maps our notion of human intelligence onto machine intelligence: humans have objectives and pursue them, so machines should have objectives and pursue them.[66]

From this we can infer that the Socratic AMA's virtuous actions can be measured insofar as human virtuous objectives (that is, virtuous actions) are subject to measurement.

The same applies to predictive AI systems, which use AI to anticipate what is likely to happen through analysis of gathered data. The term "predictive" should not be understood in the very sense of the word. The gathered data for predictive AI systems prove that predictive AI is reliant upon existing social contexts and decision making policies.[67] From this perspective, we can see that the Socratic AMA with a predictive system does not operate in an immeasurably predictive way. Rather, AI-powered mathematical and statistical methods for data analysis make it evident that the Socratic AMA's virtuous actions are not the same as human actions flowing from divinely infused virtues. As such, the Socratic AMA can never become as virtuous as humans in that immeasurable infused virtues cannot be embedded within AI systems through measurable models and algorithms. Having received detailed instructions from humans through measurable models and algorithms, the Socratic AMA becomes virtuous in the sense that its moral questions are imbued with the purposes of virtues to guide human agents toward the cultivation of virtues through addressing moral dilemmas.

The above three observations made in light of Aquinas's theology of virtue foreground the particularities of virtuous human agents in contrast to AI, showing that the Socratic AMA cannot be as virtuous as humans. Given that Aquinas's view of virtue is anthropocentric, one may ask whether the Socratic AMA is virtuous in a broader sense. To be sure, the Socratic AMA is expected to possess powerful capacities to gather information related to our moral life and rapidly process this information to identify certain lurking issues (e.g., logical contradiction) and to generate moral advice. Yet, the Socratic AMA's moral questioning related to limited circumstances can be of help to the cultivation of virtues. We also need to bear in mind the role that human agents play in designing, programming, using AI, through which, as stressed above, virtue-related purposes can be imposed upon the Socratic AMA. In this way, the Socratic AMA becomes virtuous when its embedded purpose-oriented virtues are actualized through questioning human agents for the sake of their virtuous lives.

Conclusion

Is the Socratic AMA or AI virtuous? This article is not intended to leverage an anthropocentric definition of virtue to combat the qualification of AI as virtuous. What has been demonstrated above shows that the Socratic AMA's moral questioning is considered virtuous insofar as it can be created with purpose-oriented virtues for the cultivation of human virtues.

Be that as it may, the Socratic AMA is not virtuous in the same sense as virtuous human agents. Aquinas's theology of virtue brings to light the conceptual particularities of virtue with reference to human agents. Since the subject of virtue is the soul and the mind, the technological impossibility of reproducing the mind turns out to be the impossibility of emulating the function within AI to habituate virtues as its own nature. From the theological perspective, the teleological and immeasurable features of virtues intensify the technological impossibility of creating a human-level AI to offer humans moral advice while humans are caught up in moral dilemmas.

In a nutshell, Aquinas's theology of virtue underscores the distinctiveness of humans as virtuous agents while leaving open the possibility that the Socratic AMA can be considered virtuous in the sense of delegation. It can be anticipated that AI with delegated virtues can assist humans in their cultivation of virtues provided that it does not operate as a moral decision maker in place of humans. After all, it is the virtues habituated as the second nature of human beings that need to be cultivated through addressing moral dilemmas.

Acknowledgment

A short version of this article was presented at the conference 'Virtuous AI?: Cultural Evolution, Artificial Intelligence, and Virtue,' hosted by the Centre for Theology and the Natural Sciences of the Graduate Theological Union in Berkeley, California. I am grateful to Prof. Bob Russell (the Principal Investigator of the Program) and Dr. Braden Molhoek (Program Director) for organizing this conference.

Notes

[1]See, e.g., Luciano Floridi and J. W. Sanders, "On the Morality of Artificial Agents," *Minds and Machine* 14, no. 3 (2004): 349–79, http://dx.doi.org/10.1023/B:MIND.0000035461.63578.9d.

[2]Ximian Xu, "A Theological Account of Artificial Moral Agency," *Studies in Christian Ethics* 36, no. 3 (2023): 642–59, https://doi.org/10.1177/09539468231163002.

[3]Shannon Vallor, *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting* (Oxford: Oxford University Press, 2016).

[4]Robert Sparrow, "Virtue and Vice in Our Relationships with Robots: Is There an Asymmetry and How Might It Be Explained?," *International Journal of Social Robotics* 13 (2021): 23–29, https://doi.org/10.1007/s12369-020-00631-2.

[5]See, e.g., Michael Cuellar, "A Virtue Ethical Approach to the Use of Artificial Intelligence," *Data and Information Management* (2023), https://doi.org/10.1016/j.dim.2023.100037 (withdrawn article in press).

[6]Ingmar Persson and Julian Savulescu, "The Perils of Cognitive Enhancement and the Urgent Imperative to Enhance the Moral Character of Humanity," *Journal of Applied Philosophy* 25, no. 3 (2008): 162–77, https://doi.org/10.1111/j.1468-5930.2008.00410.x; also see Parker Crutchfield, *Moral Enhancement and the Public Good* (New York: Routledge, 2021), 42–73.

[7]See, for example, Harris Wiseman, "The Sins of Moral Enhancement Discourse," *Royal Institute of Philosophy Supplement* 83 (2018): 35–58, https://doi.org/10.1017/S1358246118000280; and Celia Deane-Drummond, "The Myth of Moral Bio-Enhancement: An Evolutionary Anthropology and Theological Critique," in *Religion and Human Enhancement: Death, Values, and Morality*, ed. Tracy J. Trothen and Calvin Mercer (Cham, Switzerland: Palgrave Macmillan, 2017), 175–90.

[8]Interestingly, Julian Savulescu endorses this criticism; see Julian Savulescu and Hannah Maslen, "Moral Enhancement and Artificial Intelligence: Moral AI?," in *Beyond Artificial Intelligence: The Disappearing Human-Machine Divide*, ed. Jan Romportl, Eva Zackova, and Jozef Kelemen (Cham, Switzerland: Springer, 2015), 80.

[9]Francisco Lara and Jan Deckers, "Artificial Intelligence As a Socratic Assistant for Moral Enhancement," *Neuroethics* 13 (2020): 275–87, https://doi.org/10.1007/s12152-019-09401-y.

[10]Ibid., 276.

[11]Ibid., 277. Following this major criticism, they raise further objections to exhaustive enhancement: moral pluralism, human or nonhuman fallibility, doubts on the autonomous status of machines, impossibility of moral progress, implying the death of morality (pp. 277–80).

[12]Blay Whitby, "Computing Machinery and Morality," *AI & Society* 22 (2008): 559, https://doi.org/10.1007/s00146-007-0100-y.

[13]Ibid., 561.

[14]Savulescu and Maslen, "Moral Enhancement and Artificial Intelligence"; and Alberto Giubilini and Julian Savulescu, "The Artificial Moral Advisor. The 'Ideal Observer' Meets Artificial Intelligence," *Philosophy & Technology* 31, no. 2 (2018): 169–88, https://doi.org/10.1007/s13347-017-0285-z.

[15]Savulescu and Maslen, "Moral Enhancement and Artificial Intelligence," 80.

[16]Lara and Deckers, "Artificial Intelligence As a Socratic Assistant for Moral Enhancement," 281.

[17]Ibid., 281–82.

[18]Ibid., 282.

[19]Ibid., 283–84.

[20]Jinxiao Zhang, Yang Yang, and Ying-Yi Hong, "Sleep Deprivation Undermines the Link between Identity and Intergroup Bias," *Sleep* 43, no. 2 (2020): zsz213, https://doi.org/10.1093/sleep/zsz213.

[21]Lara and Deckers, "Artificial Intelligence As a Socratic Assistant for Moral Enhancement," 284.

[22]Francisco Lara, "Why a Virtual Assistant for Moral Enhancement When We Could Have a Socrates?," *Science and Engineering Ethics* 27, no. 4 (2021): 42, emphasis added, https://doi.org/10.1007/s11948-021-00318-5.

[23]Given the limited space, this article does not offer a detailed treatment of the important place of moral theology in Aquinas's thought. For a helpful overview of his moral theology, see Stephen J. Pope, "Overview of the Ethics of Thomas Aquinas," in *The Ethics of Aquinas*, ed. Stephen J. Pope (Washington, DC: Georgetown University Press, 2002), 30–53.

[24]Thomas Aquinas, *Summa Theologiae*, ed. The Aquinas Institute and trans. Fr. Laurence Shapcote, Latin/English Edition of the Works of St. Thomas Aquinas, vols. 13–20 (Green Bay, WI: Aquinas Institute, 2012), II–II, pr. Hereafter *ST*.

[25]*ST*, I–II, q.49, art.1, resp.

[26]*ST*, I–II, q.49, art.1, s.c.; and art.3, ad.3.

[27]Bonnie Kent, "Habits and Virtues (Ia IIae, Qq. 49–70)," in *The Ethics of Aquinas*, ed. Pope, 116.

[28]*ST*, I–II, q.49, art.3, s.c.

[29]*ST*, I–II, q.49, art.3, ad.1. Thus, Aquinas contends, 'Wherefore habit is called first act, and operation, second act.'

[30]*ST*, I–II, q.50, art.1, resp. It is beyond the scope of this article to discuss Aquinas's view of the connection between human body and morality. On a critical inquiry into this view in *Summa Theologiae*, see Marika Rose, "The Body and Ethics in Thomas Aquinas' *Summa Theologiae*," *New Blackfriars* 94, no. 1053 (2013): 540–51, https://doi.org/10.1111/nbfr.12016. In any case, Aquinas's view of ethics cannot be simply construed as rationalist or being preoccupied with the soul. Rather, his view of habit and virtue pays due attention to the whole being of humans.

[31]*ST*, I–II, q.50, art.2, resp.

[32]*ST*, I–II, q.55, art.1, resp; emphasis added.

[33]*ST*, I–II, q.55, art.4, ad.3.

[34]*ST*, I–II, q.55, art.4, resp; also see Thomas Aquinas, "On the Virtues in General," in *Disputed Questions on the Virtues*, ed. E. M. Atkins and Thomas Williams, trans. E. M. Atkins (Cambridge, UK: Cambridge University Press, 2005), art.1, s.c.

[35]*ST*, I–II, q.56, art.3, s.c.

[36]*ST*, I–II, q.56, art.3, resp; and q.50, art.5, resp.

[37]*ST*, I–II, q.56, art.2, resp.

[38]*ST*, I–II, q.54, art.2, ad.3.

[39]Jeffrey E. Brower, "First Principles: Hylomorphism and Causation," in *The New Cambridge Companion to Aquinas*, ed. Eleonore Stump and Thomas Joseph White (Cambridge, UK: Cambridge University Press, 2022), 50.

[40]*ST*, I, q.12, art.7, ad.1; and *ST*, I–II, q.11, art.3, ad.3.

[41]Aquinas asserts that "there are some habits by which man is disposed to an end which exceeds the proportion of human nature, namely, the ultimate and perfect happiness of man … Wherefore such habits can never be in man except by Divine infusion, as is the case with all gratuitous virtues" (*ST*, I–II, q.51, art.4, resp.).

[42]*ST*, I–II, q.62, art.1, s.c. Aquinas maintains that these virtues are characterized as theological for the following reason:

> it is necessary for man to receive from God some additional principles, whereby he may be directed to supernatural happiness, even as he is directed to his connatural end, by means of his natural principles, albeit not without Divine assistance. Such like principles are called *theological virtues*: first, because their object is God, inasmuch as they direct us aright to God: second, because they are infused in us by God alone: third, because these virtues are not made known to us, save by Divine revelation, contained in Holy Writ. (*ST*, I–II, q.62, art.1, resp.)

[43]*ST*, I–II, q.63, art.3, resp. The term "cardinal virtues" connotes that these virtues imply the rectitude of appetite, which "not only confers the faculty of doing well, but also causes the good deed done." Hence, these virtues are principal among moral virtues (*ST*, I-II, q.61, art.1, resp.).

[44]*ST*, I–II, q.63, art.2, resp.

[45]Aquinas, "On Hope," in *Disputed Questions on the Virtues*, ed. E. M. Atkins and Thomas Williams, trans. E. M. Atkins (Cambridge, UK: Cambridge University Press, 2005), art.1, ad.7.

[46]*ST*, I–II, q.64, art.4, resp. As Justin Anderson notes, for Aquinas, the "measures themselves arise according to the different ends toward which one moves," *Virtue and Grace in the Theology of Thomas Aquinas* (Cambridge, UK: Cambridge University Press, 2020), 32.

[47]*ST*, I–II, q.62, art.2, resp; also see Thomas Aquinas, *On Love and Charity: Readings from the Commentary on the Sentences of Peter Lombard*, trans. Peter A. Kwasniewski, Thomas Bolin, and Joseph Bolin (Washington, DC: Catholic University of America Press, 2008), III, dist.23, q.1, art.4, s.c. In *Aquinas's Ethics: Metaphysical Foundations, Moral Theory, and Theological Context* (Notre Dame, IN: University of Notre Dame Press, 2009), Rebecca Konyndyk DeYoung, Colleen McCluskey, and Christina Van Dyke, observe that "[t]he three theological virtues function as the roots of all other virtues, shaping the deepest orientation of our person and informing every other inclination and movement relevant to moral action" (p. 142).

[48]Thomas M. Osborne Jr., *Thomas Aquinas on Virtue* (Cambridge, UK: Cambridge University Press, 2022), 23.

[49]John Sullins, "When Is a Robot a Moral Agent?," in *Machine Ethics*, ed. Michael Anderson and Susan Leigh Anderson (Cambridge, UK: Cambridge University Press, 2011), 151–61.

[50]Robert Sparrow, "Why Machines Cannot Be Moral," *AI & Society* 36, no. 3 (2021): 688, https://dl.acm.org/doi/10.1007/s00146-020-01132-6.

[51]For a recent in-depth study on this front, see Yuxin Liu et al., "Artificial Moral Advisors: A New Perspective from Moral Psychology" (paper presented at the Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society [AIES '22], Oxford, UK, August 1–3, 2022), 436–45.

[52]Kenneth Einar Himma, "Artificial Agency, Consciousness, and the Criteria for Moral Agency: What Properties Must an Artificial Agent Have to Be a Moral Agent?," *Ethics and Information Technology* 11, no. 1 (2009): 19–29, https://dl.acm.org/doi/10.1007/s10676-008-9167-5.

[53]Richard A. Spinello, "Karol Wojtyla on Artificial Moral Agency and Moral Accountability," *The National Catholic Bioethics Quarterly* 11, no. 3 (2011): 496, https://www.pdcnet.org/C1257D43006C9AB1/file/FF4F20AE689EB31785257D8E0061BA79/$FILE/ncbq_2011_0011_0003_0061_0083.pdf.

[54]*ST*, I–II, q.56, art.3, resp.

[55]*ST*, I–II, q.56, art.6, resp.

[56]Antonio R. Damasio, "Neuroscience and Ethics: Intersections," *The American Journal of Bioethics* 7, no. 1 (2007): 3–7, https://doi.org/10.1080/15265160601063910.

[57]Jobst Landgrebe and Barry Smith, *Why Machines Will Never Rule the World: Artificial Intelligence without Fear* (New York: Routledge, 2022), 248.

[58]Mark Coeckelbergh, *AI Ethics* (Cambridge, MA: The MIT Press, 2020), 37.

[59]*ST*, II-II, q.81, art.2-6. For Aquinas, "religion is not a theological virtue whose object is the last end, but a moral virtue which is properly about things referred to the end" (q.81, art.5, resp.).

[60]For example, see Yong Sup Song, "Religious AI As an Option to the Risks of Superintelligence: A Protestant Theological Perspective," *Theology and Science* 19, no. 1 (2021): 65-78, https://doi.org/10.1080/14746700.2020.1825196; and Eugene A. Curry, "Artificial Intelligence and Baptism: Cutting a Gordian Knot," *Theology and Science* 20, no. 2 (2022): 156–65, https://doi.org/10.1080/14746700.2022.2051248.

[61]Mihaela Constantinescu et al., "Blame It on the AI? On the Moral Responsibility of Artificial Moral Advisors," *Philosophy & Technology* 35 (2022): 14, https://doi.org/10.1007/s13347-022-00529-z. Furthermore, they observe that "it is always up to the users to decide whether to act upon the advice offered by the AMAs" (p. 15).

[62]For some scholars, this value-loaded feature of AI underpins AI ethics. For example, in "Responsibility and Artificial Intelligence," in *The Oxford Handbook of Ethics of AI*, ed. Markus D. Dubber, Frank Pasquale, and Sunit Das (Oxford, UK: Oxford University Press, 2020), Virginia Dignum argues:

> Artificial intelligence systems use data we generate in our daily lives and as such are a mirror of our interests, weaknesses, and differences. Artificial intelligence, like any other technology, is not value-neutral. Understanding the values behind the technology and deciding on how we want our values to be incorporated in AI systems requires that we are also able to decide on how and what we want AI to mean in our societies. (p. 221)

[63]Sara Lumbreras, "Lessons from the Quest for Artificial Consciousness: The Emergence Criterion, Insight-Oriented AI, and *Imago Dei*," *Zygon: Journal of Religion and Science* 57, no. 4 (2022): 973, https://doi.org/10.1111/zygo.12827.

[64]Lara and Deckers, "Artificial Intelligence As a Socratic Assistant for Moral Enhancement," 284.

[65]Peter Flach, *Machine Learning: The Art and Science of Algorithms That Make Sense of Data* (Cambridge, UK: Cambridge University Press, 2012), 13.

[66]Stuart Russell, *Human Compatible: Artificial Intelligence and the Problem of Control* (New York: Viking, 2019), 176.

[67]Further, see Michael L. Littman et al., *Gathering Strength, Gathering Storms: The One Hundred Year Study on Artificial Intelligence (AI100) 2021 Study Panel Report*, Stanford University (Stanford, CA, September 2021), 63–64, http://ai100.stanford.edu/2021-report.