

# Virtuous AI?


## Some Considerations from the Aristotelian-Thomistic Perspective

*Mariusz Tabaczek*

**ABSTRACT** This paper offers an Aristotelian-Thomistic response to the question whether AI is capable of developing virtue. On the one hand, it could be argued that this is possible on the assumption of the minimalist (thin) definition of virtue as a stable (permanent) and reliable disposition toward an actualization of a given power in the agent (in various circumstances), which effects that agent's growth in perfection. On the other hand, a closer inquiry into Aquinas's understanding of both moral and intellectual virtues, and a more detailed analysis of the ontological status of AI, show that it is highly unlikely to envision the design of specifically human-like reason-based and/or behavioral-based ("strong") AI that would possess properly human virtues. Still, virtuous "weak AI" might be possible, although a question ought to be asked whether we should classify artifacts' virtues using categories developed in reference to specifically human dispositions and actions.

**KEYWORDS** Aquinas; Aristotle; artificial intelligence; hylomorphism; intellectual virtues; moral virtues; ontological status of AI; strong AI; virtue; weak AI

✉ Mariusz Tabaczek, Pontifical University of Saint Thomas Aquinas: Rome, Italy     mtabaczek@gmail.com  
 0000-0001-6985-8337

©  FORUM PHILOSOPHICUM 29 (2024) no. 2, 371–89  
ISSN 1426-1898    E-ISSN 2353-7043

SUBM. 2 November 2023    ACC. 16 January 2024  
DOI:10.35765/forphil.2024.2902.08

## INTRODUCTION

As commonly known, the fast-growing field of technology based on computing, neural networks, and machine learning classifies many of its products as examples of artificial intelligence (AI). Intelligence as such has been defined in various ways, in reference to logical reasoning, understanding, abstraction, learning, self-awareness, emotional knowledge, planning, creativity, adaptiveness, critical thinking, and problem solving. The category of AI reflects the motivation of many designers of contemporary computing machines to mimic or copy in their products specifically human dispositions and actions, including human intelligence (marked by complex cognitive features and high levels of self-awareness and motivation).<sup>1</sup>

Among many critical questions concerning AI that have been raised more recently we find the one asking whether AI could be considered to be a moral agent, bearing some level of responsibility for its actions. Related to this question is yet another query that is the subject of this paper. Is AI capable of developing virtue, defined as a stable and persistent disposition to a particular good action, in given circumstances?

The question about virtuous AI has been asked before. In 2022, Mihaela Constantinescu and Roger Crisp wrote an article in which they answered it negatively. However, despite their skepticism, they suggested that while AI systems cannot genuinely “be” virtuous, they can behave in a virtuous way (see Constantinescu and Crisp 2022). The opinion of Derek C. Schuurman expressed in his article published in 2023, is similar. Rejecting the suggestion that AI can be a moral (and virtuous) agent, he agrees that it can perform actions that are in accordance with moral behavior. He thus speaks about “virtue-by-proxy” in reference to AI programs that mimic human virtues (see Schuurman 2023). Most recently, this topic was addressed by Ruth Groff and John Symons (2024). They share the same opinion that artificially intelligent artifacts cannot be virtuous in terms of the classical notion of *phronimos*.

What brings together all these responses is their reference to the Aristotelian virtue ethics tradition. The goal of the research presented in this article is to approach the same question from a broader perspective of the Aristotelian-Thomistic school of thought.<sup>2</sup> In order to do so, I will first

1. The term “artificial intelligence” was coined in 1955 by emeritus Stanford Professor John McCarthy and—according to Google’s English dictionary, provided by Oxford Languages—can be defined as the theory and development of computer systems that are able to perform tasks that normally require human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages.

2. A view similar to the one presented in this article may be also found in (Xu 2024).

provide a minimalist (thin) Aristotelian-Thomistic definition of virtue, which should be in principle applicable to AI (section 1). Section 2 will offer an attempt of such application. In the following step, a deeper analysis of the classical notion of virtue will be offered, with reference to the distinction between intellectual and moral virtues (section 3). This inquiry opens the way to the question about the ontology of AI, which will be addressed in section 4. The ultimate section (section 5) will provide a preliminary attempt to formulate a Thomistic response to the question about the possibility of developing virtuous AI.

### 1. A MINIMALIST (THIN) ARISTOTELIAN-THOMISTIC DEFINITION OF VIRTUE

According to the Aristotelian-Thomistic school of thought, virtue belongs to a broader category of habits, where a habit (*habitus*) is defined as an acquired disposition that improves the agent's performance, making him/her more successful in the quest to achieve a particular goal.<sup>3</sup> As a positive or good (and thus desired) habit, virtue refers to a particular power possessed by an agent. It consists in a stable (permanent) and reliable disposition toward an actualization of a given power in the agent (in various circumstances), which effects that agent's growth to perfection. Most importantly, it is thus defined in reference to, yet distinguished from, natural inclinations of the agent who possesses it.<sup>4</sup>

Two additional remarks are in place here. First, virtue must be delineated in terms of (1) its object, and (2) the very actualization of the power that it stimulates, in order to achieve it. To give an example, the virtue of prudence, referred to the life of a community of agents, may be defined in terms of the common good of the entire group in question (ad 1). At the same time, it is

3. Thornton Lockwood says that in *Nicomachean Ethics* (hereinafter: *NE*) Aristotle defines habit (*hexis*) as "an entrenched psychic condition or state which develops through experience rather than congenitally," i.e., "that according to which, with respect to emotions, 'we are having' (*echomen*) either well or badly ([*NE* II, 5] 1105b25–26)" (2013, 23). He notes that this definition is built on a pun which "plays on the fact that the word *hexis* derives from the intransitive use of the Greek verb 'to have' (*echein*) and a *hexis* is a kind of 'having' or possession" (Lockwood 2013, 23). While Greek *echein* translates into Latin *habeo* ("to have") and *hexis* into Latin *habitus* ("habit"), more contemporary authors and translators (e.g., David Ross) prefer to render *hexis* as a state of character or a disposition. The latter lines up with Aristotle's (somewhat circular) definition of habit in *Metaphysics* (hereinafter: *Meta.*), where he sees it as "a disposition according to which that which is disposed is either well or ill disposed, and either in itself or with reference to something else" (*Meta.* V, 20 [1022b 10-12]).

4. For an introduction to the Thomistic notion of virtue see: (McInerny 1997, chapter 11: Virtue; Rhonheimer 2011, section IV: Moral Virtues; Pinckaers 2005, section IV: Passions and Virtues; Floyd 2023).

appropriate to say that prudence (defined in the same communal context) consists (is manifested) in particular actualizations of the power (faculty) of distinguishing between different particular goods, where the agent faces competing demands for attention (ad 2). Note that the first aspect of this definition highlights a teleological aspect of virtue, i.e., its orientation toward a particular end (goal), while the second emphasizes its practical orientation toward concrete action (agency).

The second important remark refers to a significant twist that the theory of virtue brings into the classical principle which states that the mode of action follows upon the mode of being (*agere sequitur esse*). Understood as a habit, distinct from natural and instinctive inclinations, virtue emerges from particular actions (actualizations or manifestations of a given power or faculty) regularly repeated over time. Consequently, as McNerny notes,

as a quality it [virtue] comes to be and is preserved in being only by action, and therefore it follows upon and is dependent on action. In that sense we can say that being, albeit accidental, follows action (McNerny 1997, 154).

## 2. AI AND THE MINIMALIST (THIN) ARISTOTELIAN-THOMISTIC DEFINITION OF VIRTUE

The subject of the proposed minimalist (thin) version of the Aristotelian-Thomistic definition of virtue is deliberately thought to be an unspecified agent. One might argue that this opens a way to speculate about the possibility of virtuous AI. In the age of machine learning and constructing neural (or connectionist) networks, it could be suggested that, at some point of their development, a new emergent feature (or a new aspect of an existing feature) might be instantiated in AI that directs it to manifest (actualize) its particular power(s)—in more or less specific circumstances—in a new way that further perfects the quality and efficiency of its operations. As an emergent feature of an AI agent, this new faculty would differ from (or further specify and adjust) its basic dispositions. It can be envisioned by the constructor who may also specify and provide for the initial conditions of its emergence. At the same time, in itself, it may remain an irreducible property/phenomenon.

To give an example, a hypothetical AI agent *A* that performs a complex computational operation *x* (let us say text editing and auto-correction) when a given set of data of the type  $d_i$  (words contained in a generalized dictionary) is fed to it, may begin to spontaneously browse for, recognize, and use sets of data falling under  $d_i$  within the broader set of information

it can analyze (e.g., words from the same language that are not contained in a generalized dictionary) or apply (and adjust)  $x$  in reference to a slightly different sets of data falling under  $d_2$  where such an operation turns out to be meaningful and useful (e.g., recognition of specific patterns in irregularly conjugated verbs and/or irregularly declined nouns fed into it by its users). Again, envisioned and planned by the constructor of  $A$ , such a new aspect (quality or feature) of  $x$  may be thought as an emergent outcome of the machine learning and/or the development of a given neural network.

With regard to the two remarks concerning virtue discussed in the previous section, one might argue that AI meets both criteria of its description. The virtuous action of our hypothetical AI agent  $A$  has a particular object (goal), i.e., performing the operation  $x$  in given circumstances. As an acquired disposition, its virtuous performance of  $x$  in the same or new context fulfills the requirement of a practical orientation toward action (agency). Again, nothing precludes  $A$ 's teleology, i.e., its orientation toward a particular end (goal), to be externally programmed and envisioned by its constructor. This should not pose a problem in reference to the minimalist (thin) version of the Aristotelian-Thomistic definition of virtue offered here.

Finally, the reversed order of the relation between action and being in the case of virtue—which highlights the role of action as foundational for the accidental being of virtue—may find support among AI constructors, who usually pay less attention to the ontology of their “creations,” while emphasizing their operational skills and effectiveness.

### 3. INTELLECTUAL AND MORAL VIRTUES

While the minimalist (thin) version of the Aristotelian-Thomistic definition of virtue may be seen as supportive of the idea (project) of virtuous AI, an exploration of further fundamental aspects of the classical notion of virtue poses considerable difficulties for this endeavor.

One of the crucial distinctions, introduced by Aristotle and discussed by Aquinas, refers to the categories of intellectual and moral virtues.<sup>5</sup> Another version of the same distinction speaks about virtues perfecting the speculative intellect and those perfecting the practical intellect (a subject matter of ethics). It is at this point that the nature of the agent of virtue needs to be specified. Both Aristotle and Aquinas leave no doubts that such agent must be a human being.<sup>6</sup>

5. See *Summa Theologica* (hereinafter: *ST*) I-II, 57–58.

6. One could obviously argue that an angel can be virtuous as well. My analysis is limited to material agents.

With respect to the speculative intellect, Aquinas states that its primary concern is the truth of things. Hence, “the virtues of the speculative intellect are those which perfect the speculative intellect for the consideration of truth” (*STI-II*, 57, 2, co). One of these virtues is knowledge (*scientia*), defined as a habit of discovering truth “through another,” i.e., through discursive reasoning, by which we draw correct conclusions from sound premises. While it could be argued that such an operation is also instantiated in AI—based on information it gathers and rules of logic it follows—we must not forget that for Aquinas the virtue of knowledge (*scientia*) is grounded in (or moves from) another virtue proper to the speculative intellect, i.e., the virtue of understanding. He defines it in terms of a habit of discovering truth “in itself,” i.e., having an intuitive grasp of first principles of things and of actions, such as the ineluctable truth that good should be pursued and evil should be avoided. Moreover, both knowledge (*scientia*) and understanding depend on the virtue of wisdom, “which considers the highest causes,” where its ultimate concern would be the first cause, i.e., God.<sup>7</sup> A reasonable doubt may be raised whether AI can develop the latter two types of virtues that perfect speculative intellect (i.e., understanding and wisdom). This leads to the conclusion that among intellectual virtues at least these two are specifically human.<sup>8</sup>

Considering the practical intellect, on Aquinas’s account it determines the right action in accordance with the moral truth of things, discovered by the speculative intellect. While one could argue that AI can perform a similar evaluation—based on the information it gathers and processes in accordance to logical procedures encoded by its designer—we must not forget that for Aquinas moral acts of the practical intellect have directly to do with the perfection of the appetites, intuitively associated with conscious human beings. However, one could argue that when defined as inclinations

7. “For it is thus that science depends on understanding as on a virtue of higher degree: and both of these depend on wisdom, as obtaining the highest place, and containing beneath itself both understanding and science, by judging both of the conclusions of science, and of the principles on which they are based” (*STI-II*, 57, 2, ad 2).

8. Contemporary virtue epistemology goes beyond the classical list of intellectual virtues described here. It introduces a number of new categories, including attentiveness, benevolence (principle of charity), creativity, curiosity (defined as studiousness/assiduity), discernment, honesty, humility, objectivity, parsimony, (rational) passion, and scrutiny—contrasting them with curiosity (defined as attraction to unwholesome things), denial/wishful thinking, dishonesty, (irrational) dogmatism, epistemic blindness, folly, hubris, laziness, (irrational) passion, obtuseness, parsimony, superstition, anti-intellectualism, and apathy. See (Turri, Alfano, and Greco 2021). It seems to me that a meaningful argument might be made that AI can/will be a subject of at least some of these categories. However, a more detailed analysis of this suggestion goes beyond the scope of this article.

of a given entity to what is in accord with its nature—possibly without any knowledge of the reason why such a thing is appetible (hence, beyond the narrow psychological meaning)—appetites could be predicated also about inanimate objects, including AI.

And yet, a further difficulty emerges once we take into account the complexity of the relationship between moral virtues and emotions (passions). On the one hand, they are clearly distinct from each other, and this is for at least three reasons. First, while emotion is a movement of the sense appetite, moral virtue is not such a movement but rather a principle of an appetitive movement. Second, unlike virtues emotions (taken in themselves) are morally neutral. And third, while the movement of emotion begins in the appetite and terminates in reason (to which it naturally tends to conform), the impetus of virtue departs from reason and terminates in appetite (see *ST I-II*, 59, 1, co).

On the other hand, however, Aquinas has no doubt that moral virtue does not exclude emotions. Quite contrary, it necessarily involves them. Thinking with Aristotle, he sees the human person as a unity of one substance that has material and spiritual aspects to it, rather than an aggregate or union of two separate, material and immaterial (spiritual), substances. If this is the case, then emotions can and should be seen as playing an important cooperative role in the development and manifestation of moral virtues. In fact, one could argue that for Aquinas passions are as closely related to virtue as the actions (operations) it inspires:

Operation and passion stand in a twofold relation to virtue. First, as its effects; and in this way every moral virtue has some good operations as its product; and a certain pleasure or sorrow which are passions (*ST I-II*, 60, 2, co).

He goes as far as to say that in certain situations—related to justice, temperance, or fortitude—“virtue must needs be chiefly about internal emotions which are called the passions of the soul” (*ST I-II*, 60, 2, co).

This brings us to one of the most often raised objections to the possibility of AI, which states that “Computers, for all their mathematical and other seemingly high-level intellectual abilities have no emotions or feelings” (Hauser 2024). Even if this might not pose a problem with regards to intellectual virtues (emotions may not be regarded as indispensable for a rational thought), emotions (passions) may be considered crucial for a general notion of intelligence and moral virtues (as mentioned above). This may lead to the conclusion that the latter can be instantiated only in human beings.

#### 4. ONTOLOGICAL STATUS OF AI

Thus, one way or the other, we arrive at the most fundamental question concerning the ontology of AI. It is not surprising that the technocratic paradigm assumed by its theoreticians inspires them to define AI in terms of its goals—i.e., reasoning and/or behavior—rather than in reference to its nature. Nevertheless, ontology enters these accounts as both goals mentioned here are usually delineated in reference to human beings. Hence, according to the classical proposal offered by Stuart Russel and Peter Norvig, (1) the reasoning-based definition comes in two versions: (1a) one that sets the goal of AI to imitate specifically human thinking and (1b) the other that refers to the idea of developing in AI a general (ideal) rationality. Similar with (2) the behavior-based definition. It also comes in two versions: (2a) one oriented toward the goal of designing AI that matches human performance and (2b) the other one that aims at developing AI that acts rationally, where rationality is understood as a general and not necessarily/specifically human feature (see Russell and Norvig 2020, section 1.1).<sup>9</sup>

Russel and Norvig strive to provide more precise characteristics (expectations or requirements) concerning all four definitions of AI. When speaking of (1b) they concentrate on the “laws of rational thinking,” i.e., yielding correct conclusions from given premises, studied in logic. They claim that the lack of certain knowledge about many phenomena (seemingly crucial for rational thinking) may be dealt with in AI in reference to the theory of probability.<sup>10</sup> Concerning (1a) they speculate that it would require AI to have—in addition to the requirements specified for (1b)—introspection and psychological experience.<sup>11</sup> In an attempt to specify the character of (2b)

9. According to Selmer Bringsjord and Naveen Sundar Govindarajulu (1a) is supported by John Haugeland who states that AI is “The exciting new effort to make computers think . . . machines with minds, in the full and literal sense,” while (2a) is represented most prominently by Turing, whose test of linguistic indistinguishability is passed only by those systems that are able to act sufficiently like a human (more on Turing’s test below in footnote 22). (1b) is preferred by Patrick H. Winston, while George Luger and William Stubblefield may be thought as representative for (2b) (see Bringsjord and Govindarajulu 2022, section 8.1; Haugeland 1985; Turing 1950, 433–460; Winston 1992; Luger and Stubblefield 1993).

10. Interestingly, one of the examples and/or characteristic features of AI is computation-based planning. In relation to this phenomenon, a reference is made to Aristotle as the precursor of AI: “Aristotle conceived of planning as information-processing over two-and-a-half millennia back; and in addition, as Glymour (1992) notes, Aristotle can also be credited with devising the first knowledge-bases and ontologies, two types of representation schemes that have long been central to AI” (Bringsjord and Govindarajulu 2022, Section 1).

11. While this requirement may seem difficult to achieve, Russel and Norvig acknowledge that “Recently, the combination of neuroimaging methods combined with machine learning techniques for analyzing such data has led to the beginnings of a capability to ‘read



they state that the agency of AI should show traces of autonomy, perception of the environment, persistence over time, and adaptation, as well as formulation and pursue of goals (Russell and Norvig 2020, section 1.1.3). Finally, speaking of (2a) they state that the computer that could pass a rigorously applied Turing test would have to possess: (i) natural language processing (to communicate successfully in a human language), (ii) knowledge representation (to store what it knows or hears), (iii) automated reasoning (to answer questions and to draw new conclusions), (iv) machine learning (to adapt to new circumstances and to detect and extrapolate patterns), (v) computer vision and speech recognition (to perceive the world), and (vi) robotics to manipulate objects and move about.<sup>12</sup>

Definitions offered by Russell and Norvig reflect a more general distinction between the efforts to design so-called “weak AI” and “strong AI,” where the former category is defined as an information-processing machine that appears to have partial or even full mental repertoire of human persons, yet lacks consciousness. The latter category refers to artificial persons, i.e., machines that have all the mental powers we have, including phenomenal consciousness.<sup>13</sup> Nevertheless, popular descriptions of AI remain ontologically vague and often do not openly side with either of these options. Hauser provides an example of such an approach:

The scientific discipline and engineering enterprise of AI has been characterized as “the attempt to discover and implement the computational means” to make machines “behave in ways that would be called intelligent if a human were so behaving” (John McCarthy), or to make them do things that “would require intelligence if done by men” (Marvin Minsky). These standard formulations duck the question of whether deeds which indicate intelligence

minds’—that is, to ascertain the semantic content of a person’s inner thoughts” (Russell and Norvig 2020, section 1.1.2). This might be perceived as the first step to designing AI defined in terms of (1a).

12. See (Russell and Norvig 2020, section 1.1.1). The Turing test (proposed in Turing 1950, 433), was designed as a thought experiment that would avoid the philosophical vagueness of the question “Can a machine think?” A computer passes the test if a human interrogator, after posing some written questions, cannot tell whether the written responses come from a person or from a computer. It is worth noting that Turing did not consider the physical stimulation as necessary for AI to be considered to be intelligent. However, other researchers claim it is necessary. They therefore add points (v) and (vi) on the list mentioned in the main text and classify thus envisioned test as the Total Turing Test (TTT). See (Harnad 1991, 43–54).

13. See Bringsjord and Govindarajulu (2022, section 8.1). Examples of “weak AI” include Alexa, Siri, Cortana, or Google Assistant. There are no real examples of “strong AI,” as it remains to be a hypothetical theory.

when done by humans truly indicate it when done by machines: that's the philosophical question (Hauser 2023, Introduction).<sup>14</sup>

Within the philosophical reflection on the ontology of AI, hopes and optimism with regard to designing AI that imitates specifically human thinking and matches 1:1 human performance (1a and 2a) is received with a certain dose of skepticism. The ambition of developing in machines an artificial, yet fully human-like introspection, psychological experience, robust autonomy, and formulation and pursue of goals, as well as natural language processing, representation of data, reasoning, and learning that would allow AI to impart common sense in a fully human measure—is thought to be rather unrealistic. And this is for at least two reasons. The first and less convincing argument refers to the current state of the development of AI. As Bringsjord and Govindarajulu note:

[T]he most articulate of computers still can't meaningfully debate a sharp toddler. Moreover, while in certain focused areas machines out-perform minds ... minds have a (Cartesian) capacity for cultivating their expertise in virtually any sphere. ... AI simply hasn't managed to create general intelligence; it hasn't even managed to produce an artifact indicating that eventually it will create such a thing. (Bringsjord and Govindarajulu 2022, section 1)

The view of Hauser is similar:

High level intelligent action, such as presently exists in computers, however, is episodic, detached, and disintegral. Artifacts whose intelligent doings would instance human-level comprehensiveness, attachment, and integration ... remain the stuff of science fiction, and will almost certainly continue to remain so for the foreseeable future. (Hauser 2023, section 3)

Moreover, Hauser seems to be skeptical about classifying as truly intelligent both specifically human-like (1a and 2a) and nonspecifically human-like (1b and 2b) reason-based and/or behavioral-based AI agents:

Do the “low-level” deeds of smart devices and disconnected “high-level” deeds of computers—despite not achieving the general human level—nevertheless comprise or evince genuine intelligence? Is it really thinking? And if general human-level behavioral abilities ever were achieved—it might still be

14. He refers to (McCarthy 1997) and (Minsky 1968).

asked—would that really be thinking? Would human-level robots be owed human-level moral rights and owe human-level moral obligations? (Hauser 2023, section 3)

Naturally, these and similar arguments might be easily dismissed as the field of AI research and design develops rapidly and what seems unrealistic today may not look impossible tomorrow.<sup>15</sup> Nevertheless, a much stronger philosophical argument against the possibility of designing specifically human-like reason-based and/or behavioral-based AI agents can be developed in ontology. We can think about at least four possible views grounding such an argument:<sup>16</sup>

1. Philosophers of mind who favor substance or property dualism would definitely claim that machines cannot think or have conscious experience, as they consider it to be ontologically different and not merely emergent, supervenient, or epiphenomenal with respect to physical entities and/or properties.
2. Proponents of mind-brain identity (both type and token identity) who hold that specifically human-like intellectual properties are identical with biological brain processes (including those who favor the position of anomalous monism) would most likely reject the idea of “strong AI” as implausible in principle.

15. One of the anonymous reviewers of the article drew my attention to the paper that has recently been published in PNAS in which the authors present the outcomes of their experiment in which—as they claim—AI chatbots passed the Turing test, the methodology of which “goes beyond simply asking whether AI can produce an essay that looks like it was written by a human or can answer a set of factual questions, and instead involves assessing its behavioral tendencies and ‘personality’” (Mei et al. 2024). The authors of the experiment prompted AI chatbots to participate in classic behavioral economics games and compared their responses and choices with tens of thousands of humans (students) who faced the same surveys and game instructions. They claim that “the chatbots’ behaviors are generally within the support of those of humans and ... [w]hen they do differ, the chatbots’ behaviours tend to be more cooperative and altruistic than the median human, including being more trusting, generous, and reciprocating” (Mei et al. 2024). This conclusion is definitely groundbreaking and may even go beyond Turing’s original question “Can a machine think?” (since it tests AI’s behavioral tendencies and “personality”). However, apart from the fact that the outcomes of this experiment require further tests and confirmation within a broader academic community working on the ontology of AI, it is rather unlikely that AI chatbots tested in it instantiate all dispositions required for classifying them as cases of “strong AI” (see the main text). Hence, despite their breathtaking qualities, they would still fall under the umbrella of “weak AI.”

16. For an introduction to philosophy of mind and analysis of all four views presented here see: (Heil 2013; Jaworski 2011; 2016; Madden 2013). Heil does not analyze the hylomorphic view, Jaworski does but in reference to its contemporary (analytic) version, while Madden reaches back to the classical notion of the hylomorphic metaphysical composition of the human person. More recently, Aquinas’s position is delineated and defended by (Wood 2020).

3. Followers of the theories of supervenience and emergence might also show similar skepticism, although their attitude would most likely depend on what sort of grounding they consider as a necessary subvenient or lower-level base for specifically human-like supervenient or emergent intellectual and psychological features.
4. Those who accept and apply classical version of hylomorphism and anthropology based on it would most likely dismiss the idea of “strong AI” as well. They see all specifically human properties related to intellect, consciousness, and free will as dispositions grounded in the immaterial human soul, which is a particular type of substantial form that actualizes primary matter in an organic human body (human being or person). As such, human soul and its proper dispositions can be neither reduced to nor developed from a purely material base.<sup>17</sup>

Moreover, according to the Aristotelian-Thomistic school of thought all specifically human (metaphysically higher) dispositions are directed at a particular and unique ultimate goal of a human person, which is defined philosophically as happiness and theologically as beatific vision (*visio beatifica*).<sup>18</sup> Again, this particular type of intrinsic teleology, consciously discovered and freely chosen, is perceived as proper to human beings only. It must be distinguished from other aspects of teleology characteristic of the human nature (many of which operate independently of intellect and will), as well as numerous instantiations of intrinsic teleology proper to sensitive yet non-intellectual beings and other non-sentient animate and inanimate entities. It should also be differentiated from extrinsic teleology characteristic of AI, which I think can be classified as an example of “teleonomy.” This term was coined by Colin Pittendrigh and favored by Ernst Mayr (in the context of evolutionary biology). It is defined as a process or behavior “that owes its goal directedness to the operation of a program”

17. At the same time, it must be emphasized that the human soul actualizes primary matter in a way that provides for the correspondence between properly formed and structured biological matter (secondary matter) and the immaterial dispositions of self-consciousness, intellect, and will. In other words, human nature is not a bundle of purely material or material and epiphenomenal, supervenient or emergent properties—but a unity of substance with both material and immaterial (spiritual) aspects to it.

18. In my account I concentrate mainly on acquired virtues since the infused theological virtues (see *ST I-II*, 62, 1), as well as infused cardinal virtues that are “corresponding in due proportion, to theological virtues” (*ST I-II*, 63, 3, co.), require/assume a cooperation with the supernatural gift of grace. I believe this remains beyond the scope of dispositions that can possibly be developed in AI.

(Mayr 1976, 403).<sup>19</sup> Again, applied to AI, teleonomy might be classified as an externally-imposed goal-directed operational program that differs qualitatively from the intrinsic teleology proper for natural entities (agents), with its most sophisticated instantiation in the highest (intellectual and volitional) dispositions of human beings.<sup>20</sup>

Finally, according to the classical Aristotelian-Thomistic school of thought, teleology has a normative aspect to it. It is defined as a tendency to the good, where all natural goods, proper for each particular entity, are ultimately grounded in God as the source of all goodness. As Aquinas notes, “[S]ince all things flow from the Divine will, all things in their own way are inclined by appetite (*per appetitum*) towards good, but in different ways” (*STI*, 59, 1, co).<sup>21</sup> The same tradition would carefully distinguish goods that are proper for artifacts, inanimate, and animate entities, with special attention paid to human beings. Even if a number of goods can be shared by AI and humans, and AI as such (in its very existence) is good and directed toward God, only human beings are destined to the contemplation of God who is the essence of all goodness.<sup>22</sup>

Having all this in mind, we shall now go back to the question of AI and virtue, approaching it once again from the point of view of the classical Aristotelian-Thomistic school of thought.

## 5. AI AND VIRTUE

In light of what has been said up to this point (especially in section 3), it becomes clear that from the perspective of the Thomistic ontology—which builds on Aristotelian hylomorphism—it is highly unlikely, if not entirely

19. Interestingly, Mayr compares his idea of an operational program in nature with a computer program: “The purposive action of an individual, insofar as it is based on the properties of its genetic code, therefore is no more nor less purposive than the actions of a computer that has been programmed to respond appropriately to various inputs. It is, if I may say so, a purely mechanistic purposiveness” (Mayr 1988, 31). This reflection seems highly relevant and useful to AI studies.

20. Even if the category of teleonomy is more appropriate with respect to AI than teleology, one might argue that new emergent dispositions of those systems may show levels of goal-directedness that would qualify as rudimentary instantiations of Aristotelian teleology. While this might be true, we should not forget that the dispositions in question derive from and depend on the principles of extrinsic teleology introduced by engineers. Moreover, taking into account the very nature of AI systems which should be classified as sophisticated aggregates of parts (showing various levels of accidental unity) and artifacts, their goal-directedness should be still considered as accidental and not intrinsic.

21. For the defense of normativity in contemporary approach to teleology see (Bedau 1992).

22. This is naturally a theological argument, grounded in the philosophical reflection that precedes it.

impossible, to envision the design or development of specifically human-like reason-based and/or behavioral-based (“strong”) AI. Consequently, for the reasons mentioned above, we must conclude that it is also in principle impossible to copy/develop properly human virtues in AI. This conclusion remains in line with the thought of Aquinas who openly states that “Reason, or the mind, is the proper subject of human virtue” (*STI-II*, 55, 4, ad 3). And because human mind is for him grounded in the human soul—for “every operation proceeds from the soul through a certain power” (*STI-II*, 56, 1)—we must agree that virtue in its proper meaning (*per se*) is a specifically human disposition (or habit).

Having said this, we should ask about the possibility of developing virtuous “weak AI,” i.e., designing artifacts which—through repeated actions grounded in machine learning and knowledge representation (embodying concepts and information in computationally accessible and inferentially traceable forms)—would be capable of developing new and stable emergent dispositions toward certain types of action in particular circumstances. It seems that such a scenario is plausible. However, it is not clear to what extent we should classify “weak AI’s” virtues using categories developed in reference to specifically human dispositions and actions.<sup>23</sup>

23. One of the anonymous reviewers of the article suggests that my distinction between the minimalist (thin or weak) definition of virtue, that potentially could be attributed to AI, and the complete (full-blown) classical definition of virtue, which is not attributable to AI, resembles Robert Audi’s differentiation between acting “in accordance with” and “from” virtue. Audi grounds his distinction in Aristotle’s explanation of the way in which virtue differs from craft: “Aristotle notes that while the products of craft determine by themselves whether they are well produced, this does not apply to the products of virtue, since ‘for actions expressing virtue to be done justly or temperately [and hence well] it does not suffice that they are in themselves in the right state. Rather, the agent must also be in the right state when he does them. First, he must know [that he is doing virtuous actions]; second, he must decide on them, and decide on them for themselves; and, third, he must do them from a firm and unchanging character. ([*NE II*, 5] 1105a29ff).” Audi adds that “In short, action from virtue is not a behavioural concept, in the sense of one defined in terms of *what* is accomplished, as opposed to *how*. Thus the adverbial forms of virtue terms—such as ‘courageously,’ ‘honestly,’ and ‘justly’—can apply to actions not performed from the relevant virtues, and even to actions aimed at pretending to manifest those virtues. Given this thin use of virtue terms, the distinction between action merely in conformity with virtue and action from it may be regarded as a special case of a distinction between conduct of a behaviourally specified type, e.g. meting out equal shares, and conduct described mainly in terms of how it is to be explained, e.g. as done from a sense of justice” (Audi 1995, 450–51). It might be the case that my minimalist (thin) definition of virtue lines up with Audi’s concept of acting “in accordance with” virtue, as distinguished from acting “from” virtue, where the latter must meet “the *selection requirement*,” i.e., consciously and freely deciding upon an action, and “the *intrinsic motivation requirement*,” i.e., being motivated by relevant virtue, rooted in a “firm and unchanging character” (Audi 1995, 451). See also my

With respect to intellectual virtues, it has already been stated in section 3 that Aquinas's definition of the virtues of understanding and wisdom seems to be appropriate only in reference to human agents. Indeed, concerning understanding, it is hard to imagine (weak) AI agents, such as aforementioned software agents (Alexa, Siri, Cortana, or Google Assistant) being capable of the highest level of speculative reasoning and imagination that is necessary to have a grasp of (1) the natures of things—which the Aristotelian-Thomistic tradition defines metaphysically in terms of (a) substantial forms actualizing primary matter and, associated with them, (b) kind-specific instantiations of intrinsic teleology (goal-directedness) characteristic of both inanimate and animate entities—and of (2) the first principles of action, understanding of which is highly intuitive and dependent on the comprehension of the transcendental categories such as goodness or truth. Again, thinking about wisdom, it is rather unlikely that “weak AI” systems could be capable of grasping the radical transcendence of God as the highest and first principle, i.e., the primordial cause (source and creator) and the ultimate end of the universe.<sup>24</sup>

Concerning the virtue of knowledge, I have speculated that—based on information it gathers and rules of logic it follows—AI could be thought as capable of developing authentic *scientia*. However, one must not forget that the classical approach to this virtue states that it is developed through discursive reasoning, which includes both syntactic and semantic information processing. When defined as such, it seems to be attributable to human beings alone.

Nevertheless, I believe that there might still be a reason and space to speak about equivalents or analogs of intellectual virtues in “weak AI” systems. Their ability of gathering, sorting out, and synthesizing information in accordance to specific rules of logic could be classified as “artificial” or “machine” (“machine-based”) knowledge. The extent to which such systems

reference to Mihaela Constantinescu and Roger Crisp's distinction between being virtuous and behaving in virtuous way, and Derek C. Schuurman's category of “virtue-by-proxy”—both mentioned in the Introduction.

24. One might argue that, according to the logic of natural theology, the notion of God as the first cause, the source, and the creator of the universe should be distinguished from the notion of God as the ultimate end of all things, where only the latter requires (can be discovered with the help of) divine revelation. If the truth about God as the first cause, i.e., the creator of the universe, can be discovered by the power of human reason alone, maybe it could be “understood” by AI as well? However, we must not forget that the notion of creation is qualified by the claim that it is *ex nihilo*. Again, it is hard to imagine that the highly speculative and counterintuitive category of absolute, i.e., metaphysically defined, nothingness—which is challenging to grasp even by humans—could be comprehensible for AI.

can abstract some general characteristics of natures of things, based on their empirically verifiable individual accidental features, could be classified as “artificial” or “machine” (“machine-based”) understanding. And the level of “weak AI’s” assessment—as a logical conclusion of the machine-based analysis—of the fact that there may/must exist the first cause of all things in the universe (without qualifying it as divine), could be defined as “artificial” or “machine” (“machine-based”) rudimentary wisdom.

Concerning moral virtues—leaving aside the question about emotions, which indeed seem to be proper only to humans and higher animals—similar strategies could be proposed. (1) Prudence (grounding memory, intelligence, docility, shrewdness, reason, foresight, circumspection, and caution), (2) temperance (grounding chastity, sobriety, abstinence, and humility), (3) courage (grounding endurance, magnanimity, patience and perseverance), and (4) justice (grounding truthfulness, gratitude, revenge, liberality,<sup>25</sup> and friendship)—when predicated with respect to “weak AI”—could be specified by adding the same qualifying category “artificial” or “machine” (“machine-based”). Naturally, each of such “artificial” or “machine” (“machine-based”) moral (as well as intellectual) virtues would have to be carefully defined with respect to “weak AI” systems or agents. In addition, their relation to and distinctiveness from specifically and uniquely human virtues would require a clear explanation as well.

To give but one example, human prudence can be defined as a stable disposition to make good judgements about one’s behavior. Aquinas characterizes it as “wisdom concerning human affairs” (*ST II–II*, 47, 2, ad. 1). As such, it requires obtaining knowledge of the future, based on the knowledge of the present and of the past, and includes not only critical reasoning and assessment, but also foresight, circumspection, and caution. Moreover, as a moral virtue developed in a conscious human agent, prudence is perceived by Aquinas as a “quasi-natural inclination” or a “second nature” of a person. Hence, the subject of this virtue must have at least an intuitive grasp of the notion of the human nature and its intrinsic teleology, which is directed toward transient natural goods and the ultimate supernatural good of *visio beatifica*.

While it becomes apparent—based on the research presented in this article—that neither “strong” nor “weak AI” is capable of developing the virtue of prudence as defined in the Aristotelian-Thomistic school of thought, one could still defend the possibility of an instantiation of “artificial” or

25. A virtue whereby we benefit others by giving or sharing with them the goods we possess.



“machine” (“machine-based”) prudence in “weak AI” systems. Grounded in the processing of data concerning the present and the past, as well as analytic modeling of the future, a “weak AI” machine—e.g., a personalized system of disease diagnosis, clinical results prediction, and drug development (designed in reference to the latest achievements of systems biology)—may be fed with the information about a particular human person and the condition of his/her flourishing. While this data would remain far from the full-blown Aristotelian-Thomistic notion of the human nature and its transient and ultimate ends, it could still count as grounding “artificial” or “machine” (“machine-based”) prudence of the “weak AI” system that makes proper judgements with respect to therapy strategies designed specifically for that person.

Having said this, I am aware of a possible skeptical reaction to the introduction of the category of “weak AI”-based equivalents or analogs of specifically human intellectual and moral virtues, coming from classically-minded thinkers. They may argue that such terminology distorts the clear distinction between humans and machines, between what is natural and what is artificial. Acknowledging this difficulty, I still argue in favor of the proposed strategy as possibly adequate and helpful in a proper analysis and classification of present and future technologies and artificial agents, which will presumably show ever more sophisticated and human-like dispositions. Alternatively, one could think about another tactic with respect to specific, emergent, and stable dispositions that could be developed in “weak AI” agents and classified as virtuous, i.e., rising the perfection of their operations in various circumstances. One could suggest that we should develop and coin a set of AI-specific (AI-exclusive) terms to name such dispositions. While such an approach may be criticized for an unnecessary multiplication of beings (categories), it could be defended as protecting AI designers and theorists from anthropomorphisms and providing a clear-cut distinction between machine-based and specifically human virtues.

## CONCLUSION

In the age of rapid development of technologies based on computing, neural networks, and machine learning, which are designed to enter into an interaction with humans, we should be all the more careful in categorizing and describing products of our ingenuity. AI, capable of perceiving, synthesizing, and inferring information, should be clearly distinguished from human intelligence, defined in reference to introspection, consciousness, and psychological experience, robust autonomy and formulation and pursuit of goals, as well as natural language processing, reasoning, and learning.

This clear distinction between artificial and human intelligence enables us, in turn, to recognize a specifically human character of intellectual and moral virtues, defined as habits disposing us to a good and desired actions that perfect our nature. While analogical accidental yet stable properties might be established as emergent features in “weak AI” agents, they should be clearly differentiated from specifically human virtues. They can be classified as “artificial” or “machine” (“machine-based”) virtues. Alternatively, developing a separate nomenclature and classification of such dispositions might be useful in the future progress of AI studies.

#### BIBLIOGRAPHY

- Aquinas, Thomas. 1947. *Summa Theologia*. Translated by the Fathers of the English Dominican Province. Benziger Bros. Online Edition: <http://dhspriority.org/thomas/english/summa/index.html>
- Aristotle. 2001. “*Ethica Nicomachea* (Nicomachean Ethics),” translated by W. D. Ross. In *The Basic Works of Aristotle*. Edited by Richard McKeon, 935–112. New York: The Modern Library.
- Aristotle. 2001. “*Metaphysica* (Metaphysics),” translated by W. D. Ross. In *The Basic Works of Aristotle*. Edited by Richard McKeon, 681–926. New York: The Modern Library, 2001.
- Audi, Robert. 1995. “Acting From Virtue.” *Mind* 104 (415): 449–71. <https://doi.org/10.1093/mind/104.415.449>.
- Bedau, Mark. 1992. “Where’s the Good in Teleology?” *Philosophy and Phenomenological Research* 52 (4): 781–806. <https://doi.org/10.2307/2107911>.
- Bringsjord, Selmer, and Naveen Sundar Govindarajulu. 2022. “Artificial Intelligence.” In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta and Uri Nodelman, Fall 2022 edition. Accessed May 10, 2023. <https://plato.stanford.edu/archives/fall2022/entries/artificial-intelligence/>.
- Constantinescu, Mihaela, and Roger Crisp. 2022. “Can Robotic AI Systems Be Virtuous and Why Does This Matter?” *International Journal of Social Robotics* 14 (6): 1547–57. <https://doi.org/10.1007/s12369-022-00887-w>.
- Floyd, Shawn. 2023. “Aquinas: Moral Philosophy.” *Internet Encyclopedia of Philosophy*. Accessed May 15, 2023. <https://iep.utm.edu/thomasaquinas-moral-philosophy/>.
- Glymour, Clark. 1992. *Thinking Things Through*. Cambridge, MA: MIT Press.
- Groff, Ruth, and John Symons. 2024. “Is AI Capable of Aristotelian Full Moral Virtue? The Rational Power of *Phronesis*, Machine Learning and Regularity.” In *Artificial Dispositions: Investigating Ethical and Metaphysical Issues*, edited by William A. Bauer and Anna Marmodoro, 219–32. London: Bloomsbury Academic.
- Harnad, Stevan. 1991. “Other Bodies, Other Minds: A Machine Incarnation of an Old Philosophical Problem.” *Minds and Machines* 1 (1): 43–54. <https://doi.org/10.1007/BF00360578>.
- Haugeland, John. 1985. *Artificial Intelligence: The Very Idea*. Cambridge, MA: MIT Press.
- Hauser, Larry. 2024. “Artificial Intelligence.” In *Internet Encyclopedia of Philosophy*. Accessed December 20, 2024. <https://iep.utm.edu/artificial-intelligence/>.
- Heil, John. 2013. *Philosophy of Mind: A Contemporary Introduction*. New York: Routledge.
- Jaworski, William. 2011. *Philosophy of Mind: A Comprehensive Introduction*. Malden, MA: John Wiley & Sons.

- Jaworski, William. 2016. *Structure and the Metaphysics of Mind: How Hylomorphism Solves the Mind-Body Problem*. Oxford: Oxford University Press.
- Lockwood, Thornton C. 2013. "Habituation, Habit, and Character in Aristotle's *Nicomachean Ethics*." In *A History of Habit: From Aristotle to Bourdieu*, edited by Tom Sparrow and Adam Hutchinson, 19–37. Plymouth: Lexington Books.
- Luger, George F., and William A. Stubblefield. 1993. *Artificial Intelligence: Structures and Strategies for Complex Problem Solving*. Redwood, CA: Benjamin Cummings.
- Madden, James D. 2013. *Mind, Matter, and Nature: A Thomistic Proposal for the Philosophy of Mind*. Washington, D.C.: CUA Press.
- Mayr, Ernst. 1976. "Teleological and Teleonomic: A New Analysis." In *Evolution and the Diversity of Life: Selected Essays*, 383–404. Cambridge, MA: Harvard University Press.
- Mayr, Ernst. 1988. *Toward a New Philosophy of Biology: Observations of an Evolutionist*. Cambridge, Mass.: Harvard University Press.
- McCarthy, John. 1979. "Ascribing Mental Qualities to Machines." In *Philosophical Perspectives in Artificial Intelligence*, edited by M. Ringle. Brighton: Harvester Press.
- McInerney, Dennis Q. 1997. *A Course in Thomistic Ethics*, 3rd ed. Elmhurst, PA: The Priestly Fraternity of Saint Peter.
- Mei, Qiaozhu, Yutong Xie, Walter Yuan, and Matthew O. Jackson. 2024. "A Turing Test of Whether AI Chatbots Are Behaviorally Similar to Humans." *Proceedings of the National Academy of Sciences* 121 (9): e2313925121. <https://doi.org/10.1073/pnas.2313925121>.
- Minsky, Marvin. 1968. *Semantic Information Processing*. Cambridge, MA: MIT Press.
- Pinckaers, Servais. 2005. *The Pinckaers Reader: Renewing Thomistic Moral Theology*. Washington, DC: The Catholic University of America Press.
- Rhonheimer, Martin. 2011. *The Perspective of Morality: Philosophical Foundations of Thomistic Virtue Ethics*. Translated by Gerald Malsbary. Washington, DC: The Catholic University of America Press.
- Russell, Stuart, and Peter Norvig. 2020. *Artificial Intelligence: A Modern Approach*. 4th ed. Hoboken, NJ: Pearson.
- Schuurman, Derek C. 2023. "Virtue and Artificial Intelligence." *Perspectives on Science and Christian Faith* 75 (3): 155–61. <https://doi.org/10.56315/PSCF12-23Schuurman>.
- Turing, Alan. 1950. "Computing Machinery and Intelligence." *Mind* 59: 433–460. <https://doi.org/10.1093/mind/LIX.236.433>.
- Turri, John, Mark Alfano, and John Greco. 2021. "Virtue Epistemology." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Winter 2021 edition. <https://plato.stanford.edu/archives/win2021/entries/epistemology-virtue/>.
- Winston, Patrick Henry. 1992. *Artificial Intelligence*. Reading, MA: Addison-Wesley.
- Wood, Adam. 2020. *Thomas Aquinas on the Immateriality of the Human Intellect*. Washington, D.C.: The Catholic University of America Press.
- Xu, Ximian. 2024. "How Virtuous Can Artificial Intelligence Become?: Exploring Artificial Moral Advisor in Light of the Thomistic Idea of Virtue." *Perspectives on Science and Christian Faith* 76 (2).

