A theological account of artificial moral agency

A Theological Account of Artificial Moral Agency

Ximian Xu
Kenneth and Isabel Morrison Post-doctoral Research Fellow in Theology and Ethics of A.I.
School of Divinity and Centre for Technomoral Futures, Edinburgh Futures Institute
University of Edinburgh
Email: Simeon.Xu@ed.ac.uk

The terms 'moral agency' and 'moral agent' have a weighty tradition and ubiquity in usage such that they become clichés and are understood in different senses. It is hard to make a unanimous definition of 'moral agency' or 'moral agent.' However, we more often than not come across general accounts of moral agent. *Routledge Encyclopedia of Philosophy*, for example, aims to set forth a generic idea of moral agent.

> Moral agents are those agents expected to meet the demands of morality. Not all agents are moral agents. Young children and animals, being capable of performing actions, may be agents in the way that stones, plants and cars are not. But though they are agents they are not automatically considered moral agents. For a moral agent must also be capable of conforming to at least some of the demands of morality.[1]

This passage points us to the three idiosyncrasies that needed to be mulled over before defining 'moral agency' and 'moral agent.' First, the expectation of the moral agent shows that the one who expects knows that the agent has the potential to meet moral demands. In this sense, the moral status of an agent is, to a certain extent, presupposed. Second, the agent's moral agency is determined by moral demands, which means that the agent is morally responsible to meet these demands. Although moral demands vary in different communities and across time, it holds true that moral demands *per se* make agents responsible for their own actions. Third, moral agency reflects a capability to act in a moral manner in order to meet these demands. Moral agency does not mean that the agent is capable of acting in accordance with all moral demands. An agent would fail to live up to some moral demands, and this is her in*capability* to act morally.

This generic portrayal of moral agency is vague and, consequently, opens up a way for the extensive usage of 'moral agent' and 'moral agency' not only in speaking of humans and animals but also in the representations of machines, computational artefacts, especially artificial intelligence

[1] Vinit Haksar, 'Moral Agents,' in Edward Craig (ed.), *Routledge Encyclopedia of Philosophy* (London: Routledge, 1998), doi: 10.4324/9780415249126-L049-1.

(AI). Whatever moral demands are, computational artefacts are expected to meet these demands, and many believe that computational artefacts are capable of acting morally.

The idea that computational artefacts qualify as agents is not novel. More than two decades ago, Ian Kerr suggested that computational artefacts and systems can be considered agents in a legal sense in electronic commerce.[2] Yet, the idea of artificial moral agent (AMA) is not unanimously approved. Kerr maintains that these electronic agents are not moral agents.[3] As with Kerr, Aimee van Wynsberghe and Scott Robbins contest that the idea of AMA is delusive because machines can never fully emulate human ethical reasoning.[4] On the contrary, some scholars contend that computational artefacts can be fully moral agents. To cite an instance, John Sullins asserts that smart machines and computational artefacts are fully moral agents when they perform human-level duties, are autonomous and intentional, and fully understand their responsibilities in performing their duties.[5] To further complicate the debates over AMA, others suggest that moral agency is tangled up with consciousness. Scholars like Richard Spinello stress that AMA is untenable because computational artefacts cannot have human-level consciousness.[6] By contrast, some insist that AMA is theoretically defensible by virtue of the possibility of artificial consciousness.[7]

It is beyond the scope of this article to unpack the idea of artificial consciousness in relation to moral agency. In this article, I narrow down the scope and aim to examine the theories of AMA *per se* from a theological perspective and, by doing so, seek to develop a theological and ontological framework within which AMA can be conceived of. A rationale behind this method is the conviction that ethics are closely related to and predicated upon ontology. That is to say, an ontological understanding of AI should be articulated before addressing ethical issues. The Reformed notions of archetype and ectype, which carry a strong ontological implication, will be used as the conceptual apparatus through which to flesh out both the moral status of computational artefacts and the inextricable relationship between human moral agents (HMA) and

---

[2] Ian R. Kerr, "Spirits in the Material World: Intelligent Agents as Intermediaries in Electronic Commerce," *Dalhousie Law Journal* 22.2 (1999): pp. 190–249.
[3] Kerr, "Spirits in the Material World," p. 216.
[4] Aimee van Wynsberghe and Scott Robbins, "Critiquing the Reasons for Making Artificial Moral Agents," *Science and Engineering Ethics* 25.3 (2019): p. 722; doi: 10.1007/s11948-018-0030-8.
[5] John Sullins, "When Is a Robot a Moral Agent?," in Michael Anderson and Susan Leigh Anderson (eds.), *Machine Ethics* (Cambridge: Cambridge University Press, 2011), pp. 151–161.
[6] Richard A. Spinello, "Karol Wojtyla on Artificial Moral Agency and Moral Accountability," *The National Catholic Bioethics Quarterly* 11.3 (2011): pp. 469–501.
[7] For example, 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 (2009): pp. 19–29.

AMAs. Archetype (ἀρχέτυπος) literally means the ultimate exemplar or pattern, and ectype (ἔκτυπος) literally refers to a copy, replica, or reflection of the ultimate pattern. Based on some latest studies on AMA, I shall argue that computational artefacts are embedded with partial moral agency in comparison with human moral agency. This partial moral agency is the extension of human moral agency and conduces to deploying carebots in pastoral care while maintaining the human caregiver's uniqueness and responsibility in pastoral care.

It is worth pausing here to clarify that this article will pivot on computational artefacts, particularly on AI, in relation to humanity. Broadly speaking, AI is not confined to computational artefacts that simulate human intelligence. Both humans and animals are intelligent because they have psychological skills such as perception, predication, planning, and so forth.[8] For example, both humanoid and dog robots are created with inbuilt AI systems. In this article, I will focus on AI that is designed to achieve human-like (not human-level) capacities for three reasons. First, emulating human intelligence has become part of AI from the inception.[9] Second, the human being is a significant origin of AI precisely the human researcher is inclined to have herself as a model of intelligence while designing AI. We will return to this subject later. Third, the theme 'moral agency' discussed in this article is generally conceptualised in relation to humanity. As will be seen, questions surrounding AMA are usually formulated mindful of human agency. Furthermore, 'pastoral care', another theme discussed in this article, presupposes a human-and-human relationship or communication. Hence, human-designed and human-like computational artefacts as well as AI is conducive to the exploration of issues surrounding AMA and AI-powered pastoral care.

In what follows, I will first examine Luciano Floridi and Jeff Sanders's endorsement of AMA through exploration of their Method of Abstraction, followed by critical analysis of such computerisation of morality with a particular eye on Deborah Johnson's contribution to debates over AMA. Second, the theology of archetype-ectype will be unfolded in relation to theological anthropology. By doing so, it will come to be seen that the question of AMA is the question of God's creation and God-human relationship writ large. Third, I will expand on the idea of partial artificial moral agency in the sense of ectype and on the moral connection between computational artefacts and humans. Finally, I will demonstrate how the idea of ectypal artificial moral agency

[8] Further on this see Margaret A. Boden, *AI: Its Nature and Future* (Oxford: Oxford University Press, 2016), pp. 1–20. I leave aside the discussion on AI as designed by an alien and on artificial general intelligence as well as artificial superintelligence. Such fictional AI is out of tune with current mainstream AI ethics which focuses on humans as the designers of AI and on ethical issues surrounding the application of AI in human daily lives.
[9] A typical example is the Turing Test, Alan M. Turing, "Computing Machinery and Intelligence," *Mind: A Quarterly Review of Psychology and Philosophy* 59.236 (1950): pp. 433–460.

offers some guiding principles for the deployment of computational artefacts into Christian pastoral care.

I. Artificial Moral Agency in Debate

The question of AMA touches off wide debates among scholars. Luciano Floridi and Jeff Sanders are two major proponents of AMA. Their argumentation begins with an attempt made to redefine the term 'moral agent.' Floridi and Sanders observe that the traditional account of moral agent is anthropocentric and individualistic such that ethical discourse often has nothing to do with nonhuman, corporate entities. They instead suggest that artificial artefacts can qualify as moral agents by stretching the meaning of moral agent to include all 'entities that can in principle qualify as sources of moral action.'[10] Hence, moral agents can be both natural and artificial, and there is also natural and artificial evil as long as moral agents bring forth immoral effects.[11]

Floridi and Sanders develop the idea of AMA with recourse to the Method of Abstraction. The Method of Abstraction comes from the field of Computer Science, stressing the method that 'discrete mathematics is used to specify and analyse the behaviour of information systems.' It seeks to construct a model where the variables are in accordance with observables in reality. A variable, which is common to scientific modelling, refers to a symbol that points to 'an unknown or changeable referent.' When this symbol 'hold[s] only a declared kind of data,' it is named typed variable. An observable is composed of a typed variable and the connotation the latter carries of features of the system that it depicts.[12]

Under the auspices of the Method of Abstraction, one can build and formalise the model of a system, developed at levels of abstraction. Each level of abstraction 'is a finite but non-empty set of observables, which are expected to be the building blocks in a theory characterised by their very choice.'[13] Floridi and Sanders argue that the level of abstraction at which we discuss moral agents largely relies on the conviction that human beings are moral agents. However, this level of abstraction is lower and includes too many details about HMAs. In order to steer clear of the anthropocentrically defined meaning of moral agent, they claim that a higher level of abstraction must be adopted so that fewer details about moral agents need be considered.

---

[10] Luciano Floridi and J. W. Sanders, "On the Morality of Artificial Agents," *Minds and Machine* 14.3 (2004): pp. 349–350. This paper can also be found on Luciano Floridi, "On the Morality of Artificial Agents," in Michael Anderson and Susan Leigh Anderson (eds.), *Machine Ethics* (Cambridge: Cambridge University Press, 2011), pp. 184–212.

[11] Luciano Floridi and Jeff W. Sanders, "Artificial Evil and the Foundation of Computer Ethics," *Ethics and Information Technology* 3.1 (2001): pp. 55–66.

[12] Floridi and Sanders, "On the Morality of Artificial Agents," p. 354.

[13] Floridi and Sanders, "On the Morality of Artificial Agents," p. 355.

Floridi and Sanders argue that the level of abstraction at which AMA can be conceived of needs to be upgraded through considering another three criteria: interactivity (interaction with environments), autonomy (capability to change state independently), and adaptability (learning to operate in a new way).[14] Machine Learning is cited in support. Machine Learning can interact with its environment, is autonomous and non-deterministic, and can learn to change its model of operation to adapt to new circumstances.[15] In light of this upgraded, higher level of abstraction, a moral agent can be defined as an agent that 'is capable of morally qualifiable action' which 'can cause moral good or evil.'[16]

By identifying computational artefacts as moral agents, Floridi and Sanders rightly recognise their moral importance for human life. Needless to say, technology has a bearing on humans and changes the way of human life. Yet, their methodology of recasting the concept of moral agent invites criticism.

First, presupposed in Floridi and Sanders's methodology is the computerisation of morality. Taking the Method of Abstraction from Computer Science, they implicitly equate the essence of morality with information processing. They seem to believe that moral observables can unequivocally and forthrightly reveal moral nature, and that moral models can be built every bit as similarly as scientific models. The essential difference between morality and science falls through the cracks. Morality is complex, and moral rules that underlie moral observables may vary across time. Hence, computerising or modelling morality at the levels of abstraction is nothing other than simplifying the agent's moral life.

Second, Floridi and Sanders's redefinition of moral agent is a non-starter since they make one particular level of abstraction dominant among others and blur the boundaries between levels of abstraction by adopting univocal senses of the criteria for formalising models. As noted earlier, they draw on Machine Learning and stress its interactivity, autonomy, and adaptability as criteria for justifying the idea of AMA. In this regard, they are oblivious to the distinction between the meanings of these criteria in understanding AMAs and HMAs. Joanna J. Bryson's observation on the design of Machine Learning can help us here:

> The mere fact that part of the process of design has been automated does not mean that the system itself is not designed. The choice of an [Machine Learning] algorithm, the data fed into it to train it, the point at which it is considered adequately trained to be released,

[14] Floridi and Sanders, "On the Morality of Artificial Agents," pp. 357–358.
[15] Floridi and Sanders, "On the Morality of Artificial Agents," pp. 361–362.
[16] Floridi and Sanders, "On the Morality of Artificial Agents," p. 364.

> how that point is detected by testing, and whether that testing is ongoing if the learning continues during the system's operation—all of these things are design decisions that not only must be made but also can easily be documented.[17]

At one level of abstraction, Machine Learning is interactive, autonomous, and adaptive. Be that as it may, it is worth noting that behind the scenes of these criteria is the process of design. That is to say, there are different judgments about these criteria because the criteria can be formed from the respective perspectives of designers and users of computational artefacts. As Frances Grodzinsky and his colleagues note, despite that computational artefacts could be viewed as moral agents at the user's level of abstraction, they are never deemed to have moral agency at the designer's level of abstraction and, consequently, designers always take on moral responsibility.[18]

In contradistinction to Floridi and Sanders's methodology, which leads to the homogenisation of AMAs and HMAs, Deborah Johnson suggests that the Method of Abstraction fails to capture the full sense of moral agency. All technologies, including computation systems, 'are produced, distributed, and used by people engaged in social practices and meaningful pursuits,' but certain levels of abstraction have nothing to do with human social practices.[19] To illustrate this point, Johnson and her colleague Keith Miller compare the human person's and the machine's actions of opening the door. Both a human person and an electronic locking system with an electric eye can open the door for someone who is carrying a large and heavy package. The two kinds of opening the door are thoroughly different: 'The function performed is equivalent, but the underlying processes (voluntary, autonomous act versus mechanical operations) are significantly different.'[20] That is, the human person's action for other human persons is embedded with social meaning, but the electronic locking system's operation *per se* is not social, though we can say the mechanical operation is the extension of human social practices. We will return to this point later while elaborating on ectypal artificial moral agency.

Johnson and Miller stress that '[a]bstractions are the work of humans and the abstractions themselves do not exist separately from humans.' As such, it is specious to consider computational

---

[17] Joanna J. Bryson, "The Artificial Intelligence of the Ethics of Artificial Intelligence: An Introductory Overview for Law and Regulation," in Markus D. Dubber, Frank Pasquale, and Sunit Das (eds.), *The Oxford Handbook of Ethics of AI* (Oxford: Oxford University Press, 2020), p. 6.

[18] Frances S. Grodzinsky, Keith W. Miller, and Marty J. Wolf, "The Ethics of Designing Artificial Agents," *Ethics and Information Technology* 10 (2008): pp. 115–121; doi: https://doi.org/10.1007/s10676-008-9163-9.

[19] Deborah G. Johnson, "Computer Systems: Moral Entities But Not Moral Agents," *Ethics and Information Technology* 8 (2006): pp. 197–198.

[20] Deborah G. Johnson and Keith W. Miller, "Un-making Artificial Moral Agents," *Ethics and Information Technology* 10 (2008): p. 129.

artefacts as having moral agency by virtue of one particular level of abstraction.[21] Computational artefacts cannot operate completely independently of humans and should not be considered full moral agents. To flesh out the difference between the AMA and the HMA, Johnson distinguishes between intentionality and intendings to act: 'While computer systems do not have intendings to act, they do have intentionality.'[22] She continues:

> Computer systems and other artifacts have intentionality, the intentionality put into them by the intentional acts of their designers. The intentionality of artifacts is related to their functionality. Computer systems (like other artifacts) are poised to behave in certain ways in response to input. … The output (the resulting behavior) is a function of how the system has been designed and the input I gave it.[23]

Accordingly, the intentionality of artefacts rests with the human designer's intentionality, which is actualised through intendings to act. At the same time, this inbuilt intentionality of computers means that they can, to a certain degree, operate independently of humans.

Johnson contests that computational artefacts do not have intendings to act precisely because intendings to act arises from freedom. 'The intending to act is the locus of freedom; it explains how two agents with the same desires and beliefs may behave differently.' [24] Computational artefacts are designed and produced in a standardised way and are expected to operate in a specific way. As such, computational artefacts can never be fully moral agents since they cannot be completely extricated from the human designer's intendings to act.

That said, Johnson categorically refuses to deny the moral status of computational artefacts and goes to the stake for the view that computational artefacts belong to the human moral world.

> Computer systems (and other artifacts) can be part of the moral agency of humans insofar as they provide efficacy to human moral agents and insofar as they can be the result of human moral agency. In this sense, computer systems can be moral entities but not alone moral agents.[25]

Johnson unfolds this viewpoint elsewhere in two respects. Firstly, being part of human moral agency means that computational artefacts should always be 'conceptually tethered to human agents' in such a sense that it is humans who create, design, and use computational artefacts.[26] Secondly, computational artefacts as moral entities have surrogate agency. Surrogate agents are

---

[21] Johnson and Miller, "Un-making Artificial Moral Agents," p. 132.
[22] Johnson, "Computer Systems," p. 201.
[23] Johnson, "Computer Systems," p. 201.
[24] Johnson, "Computer Systems," p. 200.
[25] Johnson, "Computer Systems," p. 203.
[26] Johnson and Miller, "Un-making Artificial Moral Agents," pp. 131–132.

employed to perform tasks on behalf and in the interests of clients. Human surrogate agents and computational surrogate agents differ in that the former 'have a first-person perspective independent of their surrogacy role,' whereas computational artefacts 'do not have interests, properly speaking, nor do they have a self or a sense of self.'[27] Hence, as surrogate moral agents, computational artefacts only pursue the interests of their human users. As such, humans should constantly take on moral responsibility for the operations of their artificial surrogate agents.

Johnson rightly hedges morality against computerisation since morality is more complex than constructing models. Furthermore, she cautions us against the view that technology as well as computational artefacts are morally neutral. In fact, artefacts are embedded with moral values while being designed. Nonetheless, Johnson does not make clear two points. In what metaphysical sense shall we understand the tether between computational artefacts and humans? This question concerns the metaphysical foundation for drawing the distinction between AMAs and HMAs. Is it possible to understand the connection between AMAs and HMAs in a non-utilitarian sense? Johnson meticulously delineates how humans are morally intertwined with computational artefacts while designing and using artefacts for utilitarian purposes, that is, human intendings to act through artefacts. However, the term 'surrogate' entails the impression that the only sense in which humans and AMAs are connected is utilitarian. In this light, it seems impossible for us to uncover the ontological connection between what the human being is and what the AMA is, and the divide between HMAs and AMAs overwhelms their connection and resemblance.

Mahi Hardalupas recently raises the idea of partial moral agency, which keeps the close ties between HMAs and AMAs while differentiating these two kinds of moral agenthood. She suggests that there are four conditions for judging full moral agenthood: (1) action evaluated by moral rules; (2) act according to moral rules; (3) possibility to follow different rules; (4) moral motivators, which means either believing an action as moral or the rules to follow as moral.[28] Machines and computational artefacts are currently partial moral agents because they can only fulfil parts of these conditions, especially the first three conditions.

Hardalupas does not unfold the four conditions; neither does she discuss whether or not humans can create machines that are able to fulfil all of these conditions in the future. It is also unclear whether or not her four conditions would eventually result in a rule-based morality, a

[27] Deborah G. Johnson and Thomas M. Powers, "Computers as Surrogate Agents," in Jeroen van den Hoven and John Weckert (eds.), *Information Technology and Moral Philosophy* (Cambridge: Cambridge University Press, 2008), 257 http://doi:10.1017/CBO9780511498725.014.

[28] Mahi Hardalupas, "A Systematic Account of Machine Moral Agency," in Vincent C. Müller (ed.), *Philosophy and Theory of Artificial Intelligence 2017* (Cham: Springer, 2018), p. 253.

variant of computerisation of morality. That said, the idea of partial moral agency is a better conceptual apparatus than surrogate agency through which to construe the fact that computational artefacts are part of human moral agency. This is so for two reasons. Firstly, unlike 'surrogate agency' that implies more separation than connection, 'partial moral agency' intensifies that AMAs are part of human moral agency. Secondly, 'partial moral agency' stresses that AMAs can never escape moral responsibility. I proceed to tease out the concept of partial moral agency from the theological perspective. By doing so, the two ambiguous points in Johnson's thought can be clarified.

II. Theology of Archetype-Ectype

In the Reformed tradition, the ideas of archetype and ectype are not esoteric but appeared in tandem with the rise of Reformed prolegomena.[29] From the sixteenth century onwards, the archetype-ectype thinking occupied a significant place in Reformed theology and other Protestant traditions.

Franciscus Junius (1545-1602), who studied in Geneva with John Calvin (1509-1564), was the first Protestant theologian to distinguish between archetypal theology (*theologia archetypa*) and ectypal theology (*theologia ectypa*). Junius contended that while archetypal theology refers to God's self-knowledge, ectypal theology to all knowledge of God revealed to creatures.[30] Moreover, he stressed that the distinction between the archetypal and the ectypal rests in the qualitative distinction between the Creator and creatures.

> For this one [ectypal theology] is created, it is dispositional; nor is it absolute except in its own mode, but rather finite, discrete, and divinely communicated. It is, as it were, a true and definite image of that theology [archetypal theology] which we have explained is uncreated, essential or formal, most absolute, infinite, at once complete, and incommunicable.[31]

In this passage, Junius brings into explicit a crucial rationale that underlies the distinction between archetypal and ectypal theology, that is, the ontological distinction between the created and the uncreated. His idea of archetype-ectype, along with this rationale, was formative to Protestant

---

[29] Willem J. Van Asselt, "The Fundamental Meaning of Theology: Archetypal and Ectypal Theology in Seventeenth-Century Reformed Thought," *Westminster Theological Journal* 64.2 (2002): pp. 320–321. On the background to the idea of archetypal and ectypal theology, see Richard A. Muller, *Post-Reformation Reformed Dogmatics, Volume One: Prolegomena to Theology* (2nd ed.; Grand Rapids, MI: Baker, 2003), pp. 225–228.

[30] Franciscus Junius, *A Treatise on True Theology: With the Life of Franciscus Junius*, trans. David C. Noe (Grand Rapids, MI: Reformation Heritage Books, 2014), pp. 107–113.

[31] Junius, *A Treatise on True Theology*, p. 117.

theology. Protestant orthodox theologians, including both Lutheran and Reformed theologians, took note of the idea of archetype-ectype while developing their own theology.[32]

However, most theologians of the post-Reformation era wrote of the ideas of archetype and ectype in theological prolegomena. Francis Turretin (1623–1687) was one of the few theologians who deployed the archetype-ectype thinking in constructing theological anthropology. In his *Institutio Theologiae Elencticae* (1679-1685), one of the greatest works on Reformed dogmatics in the Reformed tradition, Turretin contends:

> image signifies either the archetype (*archetypon*) itself (after whose copy something is made) or the things themselves in God (in the likeness of which man was made); or the ectype itself, which is made after the copy of another thing, or the similitude itself (which is in man and the relation to God himself). In the former sense, man is said to have been made in the image of God; in the latter, however, the very image of God.[33]

It is clear that Turretin correlates the *imago Dei* with the notion of archetype-ectype so as to articulate an ontological distinction yet connection between God and human beings. Even so, he does not expand on how this ontological implication underpins the being of humans.

The turn-of-the-century Dutch theologian Herman Bavinck (1854–1921) took a further step to use the ontological implication of the archetype-ectype thinking to account for the being of humans. He spells out the archetype-ectype thinking in conjunction with the *imago Dei*. First of all, Bavinck argues that the *whole* human being, encompassing both the soul and the body, does not have or bear the *imago Dei* but rather *is* the *imago Dei*.[34] In order to flesh out the ontological meaning of 'is,' he draws on the archetype-ectype thinking: "'Image" expresses that God is the *archetype* and the human being is the *ectype*; "likeness" adds that this image corresponds in all parts to the original.'[35] It is clear that Bavinck trades on the archetype-ectype thinking to highlight the ontological chasm between God and humans. By the notion of archetype-ectype, he attempts not to make the human being on a par with God. He argues elsewhere that God is 'the *imago increate* or archetype' and that the human being is 'the *imago creata* or ectype.'[36]

---

[32] The Lutheran theologian John Gerhard uses the idea of archetypal and ectypal theology to articulate what true theology is; Johann Gerhard, *On the Nature of Theology and on Scripture*, ed. Benjamin T. G. Mayes, trans. Richard J. Dinda (Saint Louis: Concordia Publishing House, 2009), pp. 22–24; a helpful analysis of Gerhard's idea of archetypal and ectypal theology, see Robert D. Preus, *A Study of Theological Prolegomena*, The Theology of Post-Reformation Lutheranism, Volume I (St. Louis, Missouri: Concordia Publishing House, 1970), pp. 112–114.
[33] Francis Turretin, *Institutes of Elenctic Theology*, ed. James T. Dennison Jr, trans. George Musgrave Giger (3 vols.; Phillipsburg: P&R Publishing, 1992-1997), 5.10.3.
[34] Herman Bavinck, *Reformed Dogmatics, Volume 2: God and Creation*, ed. John Bolt, trans. John Vriend (Grand Rapids, MI: Baker, 2004), p. 530.
[35] Bavinck, *God and Creation*, p. 532.
[36] Herman Bavinck, *Gereformeerde Dogmatiek, Tweede Deel* (4th ed.; Kampen: J. H. Kok, 1928), p. 493.

Bavinck reformulates this ontological distinction between the archetype and the ectype with 'being' and 'becoming.' He asserts: "The idea of God itself implies immutability. … He cannot change for better or worse, for he is the absolute, the complete, the true being. *Becoming is an attribute of creatures*, a form of change in space and time."[37] To Bavinck's mind, human becoming is related to human morality insofar as the *imago Dei* refers primarily to the spiritual and moral quality of human nature, albeit that the *imago Dei* includes both spiritual and physical dimensions.[38] As God's ectype, human beings should continue to become moral in order that they can correspond in all parts to God by displaying God's attributes. In this vein, the ontological chasm between God and humans is concomitant with their moral connection and resemblance.

The fact that the human being *is* the *imago Dei* and the ectype that corresponds in all parts with God also means that the human being does emulate God's creativity in an ectypal sense. To put this viewpoint in Philip Hefner's words, human creativity exhibits that the human being is God's created co-creator 'whose purpose is to be the agency, acting in freedom, to birth the future that is most wholesome for the nature that has birthed us.'[39] It is worth noting that, given the ontological chasm between the archetype and the ectype, there must be essential differences between the creative activities of God and humans—that is, God creates out of nothing, but humans create out of something. Viewed in this light, human artefacts are always derived from what God has already created. The qualitative distinction between divine creation and human creation turns out that there is the essential difference between humans as the consequence of God's creation and artefacts as the consequence of human creation. As will be seen, this distinction between the consequences of divine and human creation lay a metaphysical and moral foundation for the concept of the AMA's partial moral agency.

To sum up, this archetype-ectype thinking shows the inseparable bond between ontology and morality. Being God's ectype carries the connotations of both simulating God's creation and becoming moral throughout human life. As such, human action, including human creation in an ectypal sense, bears moral implications.

III. Artificial Moral Agency as Ectypal

---

[37] Bavinck, *God and Creation*, p. 158.
[38] Bavinck, *God and Creation*, pp. 549–554.
[39] Philip Hefner, *The Human Factor: Evolution, Culture, and Religion* (Minneapolis: Fortress, 1993), p. 27. A criticism has been levelled against 'created co-creator' in that this idea seems to blur ontological boundaries between the divine and the human; see, for example, Gregory R. Peterson, "The Created Co-Creator: What It Is and Is Not," *Zygon: Journal of Religion and Science* 39, no. 4 (2004): p. 829. A detailed discussion on this is beyond the scope of this article. Yet, Hefner makes it clear that 'the co-creator has no equality with God the creator;' see *The Human Factor*, pp. 38–39.

Both the idea of AMA and the archetype-ectype thinking emphasise the intrinsic link between what the human being is and what the human being acts. They differ in that the archetype-ectype thinking grounds this intrinsic link in God and his creative work, whereas the idea of AMA articulated in Floridi's, Sanders's, and Johnson's works depicts this intrinsic link from perspectives of a utilitarian purpose and the functions of computational artefacts.

That said, I do not mean that the archetype-ectype thinking contradicts the notion of AMA. In fact, the theology of archetype-ectype can lay an ontological and moral foundation on which we can conceive of *partial* moral agency attributed to computational artefacts. This theological account of partial moral agency can be developed according to the following two syllogisms. The first syllogism is related to the difference between divine and human creation and to the distinction between AMAs and HMAs.

(1) the major premise: human moral agency is the consequence of God's creation;

(2) the minor premise: the moral agency of artefacts is the consequence of human creative work;

(3) the conclusion: artificial moral agency differs from human moral agency due to the essential difference between divine and human creative work.

This syllogism takes issue with Floridi and Sanders' computerisation of morality through modelling and levels of abstraction in that the latter methodology is rooted in the conviction that AMAs equate to HMAs.

The second syllogism is derived from human creative work and its moral significance in relation to God's creation.

(1) the major premise: the human being is the ectype of God and thus imitates God's creation;

(2) the minor premise: God's creation is coupled with the mediation of morality to the human being as his ectype;

(3) the conclusion: human creation of computational artefacts is concomitant with the mediation of morality.

This syllogism tallies with Johnson's argument that AMAs should conceptually be tethered to HMAs. Yet, admittedly, this syllogism expands and enriches the meaning of 'tethered.' That is, the mediation of morality in the human creation of computational artefacts conveys a more dynamic rather than mechanic connection between AMAs and HMAs. At the same time, this syllogism is not content with the AMA as a surrogate agent. Rather, the AMA mediates human morality.

In light of these two syllogisms, we can unpack the idea of partial artificial moral agency in three aspects. First, partial artificial moral agency is predicated upon the fact that the computational artefact is the ectype of humanity. The meaning of ectype epitomises how computational artefacts take shape in the human mind. Anne Foerst puts it well:

> Researchers under the engineering goal who attempt to construct "smart" gadgets have to use a model of intelligence that is somehow familiar to them; the obvious choice would be themselves, as they know their own intelligence best. Choosing oneself as a model of intelligence for one's project influences the whole process of construction, and self-understanding and technological success reinforce each other.[40]

This is all the more so in the creation of AI (robots). In the 1980s, researchers were unsatisfied with virtual AI systems but instead sought to design embodied AI, such as humanoid AI robots. This progress in AI research was partly due to the failure to deal with object manipulation, sensations, and locomotion at the time. As such, physical embodiments become necessary for the performance of such functions by AI systems. Needless to say, the human embodiment is the most important model for designing the embodied AI that is capable of interacting with its environments.

The idea of computational artefacts as the ectype of humanity means that humans mediate their moral values into these artefacts while creating them. As Philip Hefner notes, technology is a mirror of humanity, showing human seeking for survival, the reality of human nature, the human desire for the other world, and human values.[41] Understanding artefacts in the ectypal sense indicates that computational artefacts are not merely part of human moral agency but also the extensions of human morality. As will be seen, this extension implies that human pastoral care can be mediated through AI-powered pastoral carebots. It is in this sense of extension that human-machine relationships can be properly understood. For example, the desire for artificial companions at bottom exhibits the lack of human companions. Seen from this perspective, the extension of human moral agency in AMAs also helps us to explore the role of AI in pastoral care. We will turn to this subject later.

Second, artificial moral agency is partial because it is ectypal, limited, and consequently only related to particular moral issues. A clarification needs to be made here. Conjoining the ectypal and the partial (limited) never implies that humans as the ectype of God have only partial rather

---

[40] Anne Foerst, *God in the Machine: What Robots Teach Us about Humanity and God* (New York: Plume, 2005), p. 67.
[41] Philip Hefner, "Technology and Human Becoming," *Zygon: Journal of Religion and Science* 37.3 (2002): pp. 657–660.

than full moral agency. As noted earlier, theologically speaking, full artificial moral agency means that human creation is on a par with God's creation out of nothing. Artificial moral agency as partial shows that humans cannot fully mediate their moral agency to computational artefacts in their creative work in the same way as God did. Partial AMA reveals the limitations of human creative work. One of the limitations of human creation is that the personal nature of human morality cannot be programmed into artificial moral agency. Robert Sparrow, Professor of Philosophy based in Monash Data Futures Institute in Australia, draws a distinction between scientific and moral matters:

> Scientific questions are objective in the familiar sense that the true value of scientific claims does not depend on who is making them. This means that such questions are fundamentally impersonal. … [E]thical decisions are tied to particular people—they are decisions for them in a non-contingent sense.'[42]

Computational artefacts are standardised and, therefore, are unable to deal with contextual variables and human different reactions across time. Likewise, as will be unpacked later, AI-powered pastoral carebots are incapable of addressing all personal dilemmas. Any attempts made to offer a standard ethical decision about all ethical dilemmas are oblivious to the personal nature of moral issues and thus doomed to fail.

Third, partial artificial moral agency as ectypal brings to light the fact that it is always the HMA who is responsible for ethical decisions by virtue of the ontological connection between the archetype and the ectype. This ontological connection is a desideratum in response to the controversial notion of 'responsibility gap.' In his well-known essay, Andreas Matthias turns our attention to automated machines and AI systems (especially Machine Learning) that do not need human interventions. He argues that the automated operation of computational artefacts casts doubts on our understanding of moral responsibility.

> Now it can be shown that there is an increasing class of machine actions, where the traditional ways of responsibility ascription are not compatible with our sense of justice and the moral framework of society because nobody has enough control over the machine's actions to be able to assume the responsibility for them. These cases constitute what we will call the responsibility gap.[43]

---

[42] Robert Sparrow, "Why Machines Cannot Be Moral," *AI & Society* 36 (2021): p. 689.
[43] Andreas Matthias, "The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata," *Ethics and Information Technology* 6 (2004): p. 177.

Johnson's concept of surrogate agency cannot bracket out the questions of whether a surrogate agent is out of control. By contrast, in light of the archetype-ectypal thinking, this ontological connection emphasises the *partial* agency of the surrogate AMA such that the responsibility gap is closed.[44] In this vein, human pastoral caregivers cannot escape pastoral responsibility. We will return to this subject anon. Human agents cannot escape moral responsibility by deploying computational artefacts in circumstances to make ethical decisions. It is always the HMA who designs and uses computational artefacts to address ethical questions.

This triple moral implication, derived from the above two syllogisms, shows that the idea that the AMA has partial and ectypal moral agency opens up a theological way to deal with moral questions related to the extensive applications of computational artefacts in human daily lives. One notable application is about AI-powered caregiving practices. In what follows, I shall use Christian pastoral care as a case to illustrate how the AMA involves morally in human life.

IV. AMAs and Pastoral Care

The progress in the development of computational artefacts, especially AI, has drawn much attention to the deployment of these artefacts into religious practice. A recent example of this type is the discussion over AI in religious pastoral care. By religious pastoral care, I refer to the care that clergy, religious communities, and laity offer from a religious perspective to those who are undergoing suffering and troubles. William Young reminds us that although AI-driven carebots are currently unable to provide religious pastoral care, religious communities should be ready to deal with the questions on the reception of 'automating relationships in ministry' because of the rapid progress in technology.[45] What underlies this readiness should include the view of the moral status of carebots. To be sure, caring relationships do not equate to moral relationships. It is nevertheless true that caregiving practices are embedded with moral goods and, consequently, intertwined with moral agency. Coupled with existing observations on AI and carebots in caregiving practices and healthcare, the idea of ectypal and partial AMA can offer three guiding principles for coping with moral issues in relation to the deployment of AI in religious pastoral care. I proceed to focus on carebots in Christian pastoral care. The three principles of AI-powered

[44] My stance is not geared toward an optimistic attitude toward emerging technology. Daniel Tigard observes that techno-optimists would like to bridge the responsibility gap since they 'would prefer to harness the newfound benefits of technology and proceed with its deployment.' Daniel W. Tigard, "There Is No Techno-Responsibility Gap," *Philosophy & Technology* 34 (2021): p. 590; doi: https://doi.org/10.1007/s13347-020-00414-7.

[45] William Young, "Virtual Pastor: Virtualization, AI, and Pastoral Care," *Theology and Science* 20.1 (2022): pp. 6–22.

Christian pastoral care are, respectively, raised in light of the three observations made earlier on partial artificial moral agency.

The first principle is that Christian communities need be ready to deploy AI-powered carebots into Christian pastoral care since they extend the HMA's agency in pastoral caregiving practices through human ectypal creativity in designing pastoral carebots. In other words, pastoral carebots as the ectype of human pastoral caregivers extend human agency in pastoral care. To be sure, AI-driven systems can liberate ministers from some routine work of pastoral care. For example, some Christian believers may expect ministers to send out Bible verses every day so that they can be strengthened to go through and endure occasional troubles and difficulties. We can imagine that an AI-driven automated system is capable of sending daily Bible verse that responds to one's troubles or to topical events which are likely to trouble us (e.g., the COVID-19 pandemic). In this way, ministers can focus more attention on others' critical needs of pastoral care, say, pastoral care at the end of life.

This principle carries a crucial implication that since the AMA is the extension of the HMA, pastoral care provided by AI-driven systems must have impacts on human ministers, more so because AI-powered pastoral carebots are designed to perform and augment caregiving practices after the model of human pastoral caregivers. In examining caregiving practices of carebots, Shannon Vallor reminds us that caregiving is not merely important for care-receivers but also has ethical significance for caregivers precisely because caregiving practices are embedded with moral goods.[46] A case in point is reciprocity in caregiving practices. Vallor maintains that we should consider reciprocity a virtue 'for understanding how to reciprocate well, in the right ways, at the right times, and as appropriate to particular circumstances and people, is part of what it means to become a good person.'[47] In this light, reciprocity as a virtue means that the caregiver's morality is being shaped through caregiving practices. So is in Christian pastoral care. The debates over whether or not AI-driven systems can be deployed into pastoral care often revolve around care-receivers. Yet, it is worth noting that pastoral caregivers themselves are being morally shaped in the course of caregiving practices. Paul the apostle writes, 'Rejoice with those who rejoice, weep with those who weep' (Rom. 12:15; NRSV). Christian pastoral care emphasises more 'rejoice and weep *with*' than 'rejoice and weep' itself. In pastoral caregiving practice, caregivers and care-receivers are united. Seen from this perspective, human pastoral caregivers cannot be completely

[46] Shannon Vallor, "Carebots and Caregivers: Sustaining the Ethical Ideal of Care in the Twenty-First Century," *Philosophy & Technology* 24.3 (2011): pp. 251–256.
[47] Vallor, "Carebots and Caregivers," p. 257.

replaced while deploying AI systems in Christian pastoral care. It is always the HMA as a pastoral caregiver who performs her pastoral and moral actions toward care-receivers.

The second principle concerning AI in pastoral care is to recognise the limitations of AI-driven carebots in dealing with personal dilemmas, showing the limited and partial nature of AMA. As noted earlier, computational artefacts as the ectypal AMA fail to cope with all moral questions because of the personal nature of moral dilemmas. Christian pastoral care is a similar case. It is a fata morgana to design a standardised pastoral carebot that lives up to the pastoral demands of all Christian communities and individuals.

The emphasis on the specificity of caregiving practices is characteristic of the works by Aimee Van Wynsberghe, a leading expert in the field of AI ethics. Van Wynsberghe argues that ethics concerning carebots must attend to 'the specific context of use, the unique needs of users, the tasks for which the robot will be used, as well as the technical capabilities of the robot.'[48] Given this, she suggests that ethics should be integrated into the design process of carebots. By doing so, a framework—which includes fundamental care values—can be created to build the specific relationship between the specificity of caregiving practices and technical capabilities. Taken from care ethics, these care values are attentiveness, responsibility, competence, and reciprocity.[49] Van Wynsberghe contests that in conjunction with the specific context and individual characteristics of care-receivers, these care values can guide the design process of carebots in a specific way and for particular caregiving practices.[50] Following this, we should compare the caregiving practices with the addition of carebots with those performed in a traditional way without carebots. By doing so, carebots can be evaluated morally and according to specific contexts and practices.[51]

Van Wynsberghe's methodology is directed at carebots in healthcare, but it is a heuristic for considering the deployment of carebots in Christian pastoral care. We can add some *theological values* to these four care values in the design process of pastoral carebots. For example, the Christian notion of hope can be a value that is used to evaluate pastoral carebots. We could ask whether care-receivers in a specific pastoral context hold the hope for God's faithfulness and deliverance more firmly with the addition of carebots to pastoral care. In short, the second principle shows that the deployment of carebots into Christian pastoral care is complicated and

---

[48] Aimee Van Wynsberghe, "Designing Robots for Care: Care Centred Value-Sensitive Design," *Science and Engineering Ethics* 19.2 (2013): p. 408.
[49] Van Wynsberghe, "Designing Robots for Care," p. 411.
[50] Van Wynsberghe, "Designing Robots for Care," pp. 415–416.
[51] Van Wynsberghe, "Designing Robots for Care," p. 424.

intertwined with what particular care-receivers really need in specific contexts from a perspective of Christian faith.

The third principle of deploying pastoral carebots into Christian pastoral care is that human ministers cannot escape responsibility insofar as carebots are ectypal caregivers and ontologically connected with human caregivers. This pastoral responsibility and ontological connection, which rest with human ectypal creative work, are splendidly manifest in the uniqueness of human responses to care-receivers. In exploring carebots for aged care, Robert Sparrow and Linda Sparrow note that human physical bodies are indispensable for human caregiving practices. This is so because we cannot understand the suffering of care-receivers without our physical bodies.

> Moreover, entities which do not understand the facts about human experience and mortality that make tears appropriate will be unable to fulfil this caring role. Sometimes the only appropriate response to another's suffering is the acknowledgement that we too share these frailties, as for instance, when our friend's suffering moves us to tears. Entities which do not share these frailties are therefore incapable of responding appropriately to them.[52]

Sparrow and Sparrow do not deny the possibility of human-level carebots altogether but leave this question open. However, I would be less convinced that carebots are capable of sharing human mortality and frailties based on algorithms and silicon-based systems.[53]

This unique bodily feature of human caregiving practices brings to light the partiality of the carebot as a caregiver and an AMA, laying emphasis on the responsibility that human caregivers should take on in pastoral care. There is no responsibility gap in pastoral care. In this respect, Amy Michelle DeBaets reminds us that the Christian idea of love—which underscores the mutuality in love—helps us conceive of a carebot not as the sole caregiver. Rather, carebots should be designed to keep human-and-human relationships in healthcare and to maintain the mutual love between caregivers and care-receivers.[54] Viewed in this light, the carebot's agency in pastoral caregiving is partial precisely because it is the mutual love between the caregiver and the care-receiver that needs to be nurtured through pastoral care. Whilst considering the role of carebots in Christian pastoral

---

[52] Robert Sparrow and Linda Sparrow, "In the Hands of Machines? The Future of Aged Care," *Minds and Machines* 16 (2006): p. 154.

[53] Jobst Landgrebe and Barry Smith's latest study provides one of the most cogent arguments against human-level AI, showing the essential distinction between humans and AI as well as computational artefacts; Jobst Landgrebe and Barry Smith, *Why Machines Will Never Rule the World: Artificial Intelligence without Fear* (New York: Routledge, 2023).

[54] Amy Michelle DeBaets, "The Robot Will See You Now: Reflections on Technologies in Healthcare," in Scott A. Midson (ed.), *Love, Technology and Theology* (London: T&T Clark, 2020), pp. 93–108.

care, we should ponder how the so-called responsibility gap is closed by such mutual love and how the HMA should not escape but rather take on the responsibility to provide pastoral care for others.

V. Conclusion

What is the moral status of computational artefacts? This article has articulated a theological account of AMA. It turns down an optimistic position that classifies the AMA and the HMA into the same category. In the meanwhile, the dismissal of AMA is declined.

The theology of archetype-ectype offers an ontological lens through which to get hold of the moral connection between the AMA and the HMA. That is, the AMA is the ectype of and ontologically connected with the HMA, and so artificial moral agency is partial and human moral values are mediated and extended through the computational artefacts. This opens up a vista for further discussions over the role of computational artefacts in human moral life. In particular, I use Christian pastoral care to illustrate that human pastoral care is extended through the AMA's limited pastoral caregiving practices and that the HMA is always responsible for pastoral care. Needless to say, further steps need to be taken to explore the deployment of AMAs into Christian pastoral care. It should be recognised, however, that the idea of ectypal and partial artificial moral agency offers some guiding principles for the deployment of computational artefacts into pastoral care as well as other spheres of human life.