Journal of Moral Theology, Vol. 11, Special Issue 1 (2022): 182–211

Theological Foundations for Moral Artificial Intelligence[1]

Mark Graves

KEY IDEAS FROM MORAL THEOLOGY CAN help make AI compatible with human morality by guiding the integration of disparate approaches to AI development toward a morally good end. As AI becomes more pervasive in society, humanity would benefit from AI development incorporating a theological anthropology that can guide AI's interdisciplinary construction and characterize its historically contextualized moral norms. As an initial foray into development of an integrative framework, I describe an AI system that could plausibly be constructed with effort comparable to other major AI initiatives, and that would have the capacity to consider itself as a moral actor (a precursor to moral agency).[2] Constructing such a system would open up new possibilities for moral AI, enable sophisticated modeling of human morality, and lead to new insights into ethics and moral behavior. Closer at hand, my proposal identifies issues in AI and morality that require both computational and ethical expertise to resolve and are not well known and understood across the necessary disciplines.

As I use the term, "moral AI" can navigate the moral dimension of its world and predict the moral consequences of its actions. To do so

[1] The initiation of the project described by this manuscript was made possible through a fellowship funded by John Templeton Foundation through St. Andrews University and the University of Notre Dame Center for Theology, Science & Human Flourishing with Celia Deane-Drummond. My project benefited from interactions through St. Andrews and at Notre Dame, especially conversations with Darcia Narváez, Emanuele Ratti, Tim Reilly, and Adam Willows and specific topics of the paper were informed by early conversations with Jean Porter, Bill Mattison, and Walter Scheirer. Thanks to Bob Lasalle-Klein, Rene Sanchez, José Sols Lucia, Pat Lippert, and other members of the John Courtney Murray group for comments and suggestions on an earlier draft. Andrew Porter was very helpful in identifying an early direction. A prior version of this article benefited from discussion at a Pacific Coast Theological Society meeting, especially comments by Brian Green, Katy Dickinson, Bob Russell, Ted Peters, Koo Yun, Braden Molhoek, Kenn Christianson, and John LaMuth. The article also significantly benefited from comments by two anonymous reviewers and the special issue editors.

[2] As explained later in this article, the difference between "actor" and motivated "agent" draws upon Dan P. McAdams, "The Psychological Self as Actor, Agent, and Author," *Perspectives on Psychological Science* 8, no. 3 (2013): 272–95.

it must conceptualize its natural, social, and moral world and reckon itself within those worlds.[3] When an AI reckons itself: (1) as a causal actor, it can engage the natural world; (2) as a sociotechnical actor, it can develop communicative relationships with others in its social world; and (3) as a moral actor, it can evaluate the ethical consequences of its actions in its moral world. An interdisciplinary construction of moral AI depends upon insights into morality and AI development, and can contribute to both as well as beneficial incorporation of AI technology into society. Many of the above words such as "moral," "conceptualize," "actor," "reckon," etc., we typically reserve for the behaviors of self-conscious agents like humans, and while I do not rely on that interpretation here, I leave open the possibility that AI might someday attain that status.[4] Several of these terms will be more fully elucidated later on, with attention to their formulation separate from assumptions of consciousness.

A number of disciplinary perspectives contribute to the development of moral AI. Computer scientists often recognize the need for ethical AI, and incorporating ethical principles into AI development, such as fairness, is an active AI research area.[5] Social scientists have studied human interaction with AI including people's tendency to anthropomorphize AI and differences in trusting AI versus humans.[6] Collaborations between philosophers, ethicists, and others have

[3] For evidence of neural networks exhibiting concept-like functioning, see Gabriel Goh, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah, "Multimodal Neurons in Artificial Neural Networks," *Distill*, 2021, distill.pub/2021/multimodal-neurons/.

[4] For differing opinions on whether AI can have self-consciousness or interiority, see Brian P. Green, Matthew J. Gaudet, Levi Checketts, Brian Cutter, Noreen Herzfeld, Cory Lebrecque, Anselm Ramelow, OP, Paul Scherz, Marga Vega, Andrea Vicini, and Jordan Joseph Wales, "Artificial Intelligence and Moral Theology: A Conversation," *Journal of Moral Theology* 11, Special Issue 1 (Spring 2022): 13–40.

[5] Stuart Russell, Daniel Dewey, and Max Tegmark, "Research Priorities for Robust and Beneficial Artificial Intelligence," *AI Magazine* 36, no. 4 (December 31, 2015): 105–14, doi.org/10.1609/aimag.v36i4.2577; Pat Langley, "Explainable, Normative, and Justified Agency," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33 (2019): 9775–79, doi.org/10.1609/aaai.v33i01.33019775; Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi, "Fairness and Abstraction in Sociotechnical Systems," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19 (New York: Association for Computing Machinery, 2019), 59–68, doi.org/10.1145/3287560.3287598; and Donghee Shin and Yong Jin Park, "Role of Fairness, Accountability, and Transparency in Algorithmic Affordance," *Computers in Human Behavior* 98 (2019): 277–84, doi.org/10.1016/j.chb.2019.04.019.

[6] Arleen Salles, Kathinka Evers, and Michele Farisco, "Anthropomorphism in AI," *AJOB Neuroscience* 11, no. 2 (April 2, 2020): 88–95, doi.org/10.1080/21507740.2020.1740350; Theo Araujo, Natali Helberger, Sanne Kruikemeier, and Claes H. de Vreese, "In AI We Trust? Perceptions about Automated Decision-Making by Artificial Intelligence," *AI & Society* 35, no. 3 (2020): 611–23, doi.org/10.1007/s00146-019-00931-w.

identified ethical principles and practices for incorporating AI predictions and other results into social structures.[7] Machine ethicists have clarified the need for explicit characterizations of ethics and the need to reconcile differences between what distinct duties (or other value frameworks) might require.[8] Theologians have begun examining AI in the context of theological anthropology, and elsewhere in this volume, moral theology.[9] Collaborative engagement on the development of moral AI can prescribe key components for AI development and guide ongoing efforts to incorporate ethics into AI.

Moral theologians can help construct a framework to integrate technical, social, and ethical contributions on AI with scientific, scholarly, and normative insights into human society. Although differences among ethical theories, schools of thought, and religious traditions are legion, I agree with ethicist Susan Anderson that enough consensus on ethical thought exists to guide construction of moral AI.[10] However, constructing moral AI is a normative process, not a descriptive one, and although what exists in human morality is an important aspect of developing moral AI, building an AI system with moral judgment and behavior requires reasoning about moral normativity in a moral actor with radically different embodiment and socialization. AI developers

[7] Luciano Floridi, Josh Cowls, Monica Beltrametti, Raja Chatila, Patrice Chazerand, Virginia Dignum, Christoph Luetge, Robert Madelin, Ugo Pagallo, Francesca Rossi, Burkhard Schafer, Peggy Valcke, and Effy Vayena, "AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations," *Minds and Machines* 28, no. 4 (2018): 689–707, doi.org/10.1007/s11023-018-9482-5; Jessica Morley, Luciano Floridi, Libby Kinsey, and Anat Elhalal, "From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices," *Science and Engineering Ethics* 26, no. 4 (August 1, 2020): 2141–68, doi.org/10.1007/s11948-019-00165-5.

[8] Michael Anderson and Susan Leigh Anderson, *Machine Ethics* (Cambridge: Cambridge University Press, 2011); Wendell Wallach and Peter Asaro, *Machine Ethics and Robot Ethics* (New York: Routledge, 2017); Susan Leigh Anderson, "Machine Metaethics," in *Machine Ethics*, ed. Michael Anderson and Susan Leigh Anderson (Cambridge: Cambridge University Press, 2011), 21–27.

[9] Noreen L Herzfeld, *In Our Image: Artificial Intelligence and the Human Spirit* (Minneapolis, MN: Fortress, 2002); Anne Foerst, *God in the Machine: What Robots Teach Us about Humanity and God* (New York: Dutton, 2004); William F. Clocksin, "Artificial Intelligence and the Future," *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences* 361, no. 1809 (2003): 1721–48, doi.org/10.1098/rsta.2003.1232; Russell C. Bjork, "Artificial Intelligence and the Soul," *Perspectives on Science and Christian Faith* 60, no. 2 (2008): 95–102; Andrew Peabody Porter, "A Theologian Looks at AI," in *2014 AAAI Fall Symposium Series*, 2014.

[10] Anderson, "Machine Metaethics." Practical issues that would require theoretical ethical nuance also require significant immersion in technology development. Philosopher of technology ethics Shannon Vallor makes a similar point on consensus. See her *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting* (New York: Oxford University Press, 2016), doi.org/10.1093/acprof:oso/9780190498511.003.0001.

often have moral intuitions grounded in a rich intellectual tradition but lack the historical and philosophical knowledge and expertise to make those intuitions explicit for machine ethics; and ethicists typically lack sufficient insight into rapidly developing technologies to identify detailed social and moral implications before technical development has progressed past the point of immediate relevancy. Moral theologians can help bridge that gap with an integrative framework for moral AI within which other disciplines can dialogue and collaborate.

The Interdisciplinary Challenge: Snow's "Two Cultures" Problem

A challenge to interdisciplinary investigation of moral AI is the relatively non-overlapping educational training of computer scientists (and engineers) and moral theologians (and philosophers and ethicists), which severely limits the construction of robust theories incorporating both advanced technical understanding and scholarly insight. One can trace recognition of the challenge to C. P. Snow's identification of two cultures separating science and the humanities.[11] Differences in the presumed background knowledge and trained methodologies hinder dialogue between scientists and scholars, and sophisticated theories in one discipline may include assumptions considered naive by the other. Ian Barbour and others have previously studied challenges to dialogue between theology and natural science, and studying AI morality can draw upon those lessons. Advances also require integrating that academic discourse with its related technology and ethics dialogue, previously viewed primarily as applications of science and theology, respectively.[12] In the case of AI morality, this integration reverses the previously noted distinction between theoretician and practitioner. For the specific technological application of interest is an engineered system that threatens to replicate the experience and intellectual expertise previously presumed the exclusive purview of scientists and theologians.[13] One must also incorporate the social sciences

[11] C. P. Snow, *The Two Cultures and the Scientific Revolution* (New York: Cambridge University Press, 1959).

[12] Ian G. Barbour, *Religion and Science: Historical and Contemporary Issues* (San Francisco: Harper, 1997); and Ian G. Barbour, *Ethics in an Age of Technology* (San Francisco: Harper, 1993).

[13] Joe Dysart, "The Writing Is on the Wall for Artificial Intelligence," *Research-Technology Management* 62, no. 6 (2019): 8; Beta Writer, *Lithium-Ion Batteries: A Machine-Generated Summary of Current Research* (Springer International, 2019), www.springer.com/us/book/9783030167998; Mark Graves, "AI Reading Theology: Promises and Perils," in *AI and IA: Utopia or Extinction?*, *Agathon* 5 (2018); and Xin He, Kaiyong Zhao, and Xiaowen Chu, "AutoML: A Survey of the State-of-the-Art," *Knowledge-Based Systems* 212 (January 5, 2021): 106622, doi.org/10.1016/j.knosys.2020.106622. Because AI fundamentally relates to human experience and mental processing in a way no previous technology has, it depends in a novel way upon and can impact every field that studies or relies upon human cognition. Studying AI morality not only requires innovative integration of humanities with

as they identify social structures that AI impacts and disrupts as well as explain the human psychology that AI purports to replicate partially and with which AI must often interact. The social sciences are also needed because philosophers and computer scientists like John Searle, Hubert Dreyfus, and Brian Cantwell Smith convincingly identify certain knowledge, phenomenological engagement, and commitments to the world as missing in AI but do not appear to fully appreciate the relevant and nuanced contributions to those mental capacities by sociology of knowledge and social and developmental psychology, even for humans.[14] The interdisciplinary challenge is addressed through a collaborative framework for moral AI development that can integrate the discipline-specific theories and shift efforts from loose discussion and dialogue to something that focuses and constrains contributors sufficiently to impact theories and practices from other contributing disciplines.

Moral AI raises many questions of personhood not addressable in a single article, and some assumptions must be made with respect to AI's cognitive capabilities, moral agency, phenomenological consciousness, and moral continuity with humans.[15] Possible AI cognitive capabilities can variously refer to the equivalent of: (1) an artifact such as a calculator or computer, (2) an intelligent non-human animal, (3) that new intelligent animal-like "species" plus language and culture, or (4) also include a degree of self-awareness and reflection, most similar to modern humans.[16] Other options are possible as well. Here I aim to clarify how an AI beginning with intelligence of a non-human animal can add the capability to participate in the human social world, which enables better characterization of the necessary preconditions for self-reckoning as a foundation for self-awareness and reflection.[17]

natural and social sciences, it can also require examining the presumptions and historical accidents that led to their separation.

[14] John R. Searle, *Minds, Brains, and Science* (Cambridge, MA: Harvard University Press, 1984); Hubert L. Dreyfus, "Why Heideggerian AI Failed and How Fixing It Would Require Making It More Heideggerian," *Philosophical Psychology* 20, no. 2 (2007): 247–68; and Brian Cantwell Smith, *The Promise of Artificial Intelligence: Reckoning and Judgment* (Cambridge, MA: MIT Press, 2019). See also Porter, "A Theologian Looks at AI."

[15] The use of "AI" as an entity, instead of a research field, presumes a not-yet-existent level of cohesion and generalizability among the outputs of that field, which requires additional integrative work, such as proposed here.

[16] Comparing cognition between humans and AI is possible because the fields of AI and cognitive psychology have informed each other's development within the broad umbrella of cognitive science, resulting in compatible scientific characterizations between human and AI cognition, though their mechanisms, embodied realization, and phenomenological concerns differ substantially. See George A Miller, "The Cognitive Revolution: A Historical Perspective," *Trends in Cognitive Science* 7, no. 3 (2003): 141–44.

[17] In this usage, self-reckoning is a foundation for self-awareness, but the self lacks awareness of itself as a "knower."

Moral agency often implies a high degree of autonomy, though AI could have restricted (e.g., safe) agency; exist in a way so its "free will" is "compatible" with an otherwise deterministic foundation; or result from humans giving it equivalence to agency in a sociotechnical system, such as of a judge, loan officer, or corporate executive, even though the AI technology lacks intrinsic agency.[18] Common to all these types of moral agency is the capacity of AI for moral attention and interpretation and ultimately the ability to judge the impacts of its own decision making. I focus on AI interpreting its world in a way that admits moral decisions and action and includes recognition of its own actions, without requiring those decisions and actions to be motivated or autonomous. Considering the range of AI's relationships to its "self" from none through self-reckoning to full phenomenological consciousness and reflection upon its inner life, I target self-reckoning as AI perceiving its own existence in its world, but not necessarily any greater awareness of itself or its interior processing. I argue that an AI with these cognitive and self-reckoning capacities engaging a human social world through language and attending to value-laden and normative interpretations suffices as a foundation for considering AI's moral continuity with humans in that world.[19]

A Framework for Moral Theology and AI Research

In this article, I propose an initial framework for drawing moral theologians into the multifaceted, integrative discourse on moral AI. The article unfolds in two main parts. First, a theological foundation for moral AI requires something like a secularized theological anthropology. The "anthropology" characterizes the natural, social, and moral aspects of an AI that exists in a world with humans, sin, and grace and focuses on what is needed to characterize such a social and moral entity (though without directly attributing sin or grace to AI). Critiques of current approaches to AI identify limitations to AI's more anthropological development, and I respond by adapting Donald Gelpi's theological anthropology for moral AI to emphasize the AI's

[18] John McCarthy, "Free Will—Even for Robots," *Journal of Experimental & Theoretical Artificial Intelligence* 12, no. 3 (July 2000): 341–52, doi.org/10.1080/09528130050111473; Riccardo Manzotti, "Machine Free Will: Is Free Will a Necessary Ingredient of Machine Consciousness?," *Advances in Experimental Medicine and Biology* 718 (January 1, 2011): 181–91, doi.org/10.1007/978-1-4614-0164-3_15; Paul N. Edwards, "Infrastructure and Modernity: Force, Time, and Social Organization in the History of Sociotechnical Systems," in *Modernity and Technology*, ed. Thomas J. Misa, Philip Brey, and Andrew Feenberg (Cambridge, MA: MIT Press, 2003), 185–226; and Selbst, Boyd, Friedler, Venkatasubramanian, and Vertesi, "Fairness and Abstraction in Sociotechnical Systems."

[19] Although greater capacities would be needed for moral agency, full moral autonomy, or moral equivalency with humans, I claim these capacities suffice for interdisciplinary dialogue about AI meaningfully considered to be moral, and with a more active role than a moral patient.

moral conceptualization and self-reckoning in a casual, social, and moral world.[20] Gelpi's anthropology has a metaphysics rooted in experience, based upon C. S. Peirce's and Josiah Royce's objective idealism, and this provides theological grounding for AI's interpretive experience. To extend the anthropology for moral AI, I: (1) characterize an AI self as a moral actor that experiences its world; (2) use systems theory to organize an AI's interpretive experience of its natural, social, and moral world; (3) situate AI social apprehension within Ignacio Ellacuria's historical reality (with moral implications); and (4) adapt Thomistic ideogenesis to characterize an AI conceptualization of its (interpreted) reality in terms of moral norms. Moral norms refer here to what is modeled as normative by the AI, such as moral principles, Ross's *prima facie* duties, utilitarian preferences, proxies for human flourishing (or safety), or virtues.[21]

In the second part, insights from the extended anthropology lead to a proposal for developing moral AI. In the proposed system, moral AI's interpretive experience is characterized by five levels of models, which draw upon systems theory to characterize the AI's encounter with an external world, and five corresponding stages of self-reckoning, where the AI models itself. The multi-faceted, multi-level characterization also defines a framework that identifies the broad disciplinary needs that arise from the attempt at moral AI and a need for collaboration between moral theologians, ethicists, philosophers, social scientists, and computer scientists. The implications of the modeling are then briefly examined with respect to practical wisdom (*phronesis*) as an essential capability for moral AI.

AI Theological Anthropology

Some AI researchers recognize the need for AI to engage its natural and social world in order to develop further and fulfill its promise instead of its perils. Brian Cantwell Smith argues AI must distinguish reality from its representation and commit not just to its representations but to that to which its representations point.[22] Acknowledging Hubert Dreyfus's Heideggerian critique that AI is unable to grasp reality because symbol processing and representations cannot connect experience with existence, Cantwell Smith draws attention to the process that leads from a phenomenological encounter with reality to the

[20] Smith, *The Promise of Artificial Intelligence*; Stuart Russell, *Human Compatible: Artificial Intelligence and the Problem of Control* (New York: Penguin, 2019); Donald L Gelpi, *The Gracing of Human Experience: Rethinking the Relationship between Nature and Grace* (Collegeville, MI: Liturgical Press, 2001).

[21] Anderson, "Machine Metaethics"; Russell, *Human Compatible*; Mark Graves, "Shared Moral and Spiritual Development among Human Persons and Artificially Intelligent Agents," *Theology and Science* 15, no. 3 (2017): 333–51, doi.org/10.1080/14746700.2017.1335066.

[22] Smith, *The Promise of Artificial Intelligence*, chaps. 7, 12.

distinction between objects required for AI representation.[23] Additionally, Stuart Russell extends Nick Bostrom's philosophical argument that superintelligent AI poses an existential risk to humanity by identifying problematic assumptions in AI research and plausible future improvements in AI sufficient for uncontrollable AI advancement.[24] Rather than halt AI development, Russell argues for developing beneficial AI that identifies human preferences and attempts to maximize those utilitarian preferences with altruism and humility, specifically acknowledging the intrinsic uncertainty in accurately identifying human preferences.[25] Although not identified as such, both researchers point toward the construct of experience as key to developing AI that would have more general capabilities than the narrow and fragile applications currently available and could engage its natural and social world in an ethical way.

Three philosophical perspectives on human experience relevant for modeling AI experience are Continental phenomenology, Thomistic anthropology, and the objective idealism of pragmatism. Continental phenomenology (especially Merleau-Ponty and Heidegger) separates the experience of reality from reality to examine the former and thus provides a focus on subjective awareness that Cantwell Smith, Russell, and others have identified as needed for AI. Thomistic philosophy presumes an objective account of nature compatible with its medieval understanding of the world, which reconciles well with experience of a virtual world and the assumptions of objectivity influential on engineering and the natural sciences. However, the philosophical presumption of subjectivity by Continental philosophy does not guide engineers trying to construct something like subjectivity in machines; although the assumption of universal essences underlying Thomistic philosophy corresponds surprisingly well to presumptions of early AI knowledge representation systems, it captures poorly the evolutionary processes of the natural world, the social construction of knowledge, and contextualized morality. The objective idealism of pragmatic philosophy addresses these limitations for AI. With respect to Thomism, C. S. Peirce incorporates evolutionary processes into his logical metaphysics, thus adding evolution to an Aristotelian-influenced metaphysics, and Josiah Royce further extends Peirce's semiotic philosophy into the social, moral, and spiritual realm, which adds social and moral contextualization.[26] In addition, the pragmatist George Herbert

[23] Dreyfus, "Why Heideggerian AI Failed"; Smith, *The Promise of Artificial Intelligence*, chap. 3.

[24] Russell, *Human Compatible*, chaps. 2-3; Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford: Oxford University Press, 2014).

[25] Russell, *Human Compatible*, chaps. 7, 9.

[26] Kelly A. Parker, *The Continuity of Peirce's Thought* (Nashville: Vanderbilt University Press, 1998); Josiah Royce, *The Problem of Christianity. Lectures Delivered*

Mead changes the locus of personhood from subject or soul, as in Continental and Thomistic philosophy respectively, to the "self" as a social process, thus identifying social construction of subjectivity.[27] Although pragmatism serves as the foundational philosophical framework, a pragmatic understanding of interpretive experience is strengthened by Continental and Thomistic contributions on subjectivity and objectivity, specifically with respect to historical (and political) reality and conceptualization of moral norms.

Pragmatic Experience of Reality

Pragmatically, experience consists of encounter and interpretation.[28] As subject, one encounters one's world, and then interprets one's experience into objective categories. Subjectivity occurs at the nexus of encounters and is defined by those natural and social experiences. Interpreted "objects" are not *a priori* universals, but socially constructed with others in society (and through history and language). Without the sensory encounter, an overly rational interpretation reduces objective idealism to subjective idealism and loses the connection to the real world required by scientific study. Setting to one side possible revelatory experiences, these "others" have historically always been human, but now other precursors to persons are entering into society.[29]

Mead identifies the locus of personhood or "self" as a social process created by interactions within a group or society.[30] The individual social self initially appropriates society's shared values and ideals then, as it develops, interiorizes the social environment in which it lives, and finally begins transforming society through its relationships. AI currently appropriates society's shared values (including those with harmful effect) but does not yet interiorize the social environment in which it lives.[31] As the human "self" incorporates and responds to its social relationships, its reflective character makes it both subject and object, and its communication creates self-awareness. Although foundational for social psychology, the identification of the self as subject and object has not been sufficiently incorporated into dialogue

at the Lowell Institute in Boston, and at Manchester College, Oxford (New York: Macmillan, 1913).

[27] George Herbert Mead, *Mind, Self & Society from the Standpoint of a Social Behaviorist* (Chicago: University of Chicago Press, 1934).

[28] Denis Edwards, *Human Experience of God* (New York: Paulist, 1983); John Edwin Smith, *Experience and God* (New York: Oxford University Press, 1968).

[29] Mark Coeckelbergh, "Robot Rights? Towards a Social-Relational Justification of Moral Consideration," *Ethics and Information Technology* 12, no. 3 (2010): 209–21.

[30] Mead, *Mind, Self & Society*.

[31] There are computational social models, but they are not yet compatible with natural language processing (NLP) deep learning models appropriating social values and biases. The early AI researcher Allen Newell does identify the Social band in *Unified Theories of Cognition* (Cambridge, MA: Harvard University Press, 1990).

between AI engineering and the humanities. If AI begins with a self that experiences its natural and social world, the question arises: What would make it moral? Advances in AI cognitive architecture and integration among methods and technologies would be required to construct such a foundation but are currently plausible given current technology and effort. Can moral theology construct the theories needed to guide such AI development in a moral direction before such AI exists?

To relate Mead's social self to the level of "self" targeted here for moral AI, a distinction from personality psychology is helpful. Dan McAdams studies the formation of identity and identifies three levels of its variation and development in personality: dispositional traits, which are fairly stable through adulthood; characteristic adaptations, which include beliefs and desires and vary throughout one's life; and narrative identity, which comprises the stories one constructs to give one's life a sense of unity and purpose. He summarizes these developmentally as self as actor, agent, and author.[32] Simplistically, dispositional traits may depend upon early childhood development and other social and genetic factors forming the core of one's self. Conversely, characteristic adaptations are more circumstantial and subjective, depending upon one's social, historical, and cultural context as it influences how one apprehends and responds to reality. As for narrative identity, adults form stories about themselves that give meaning and coherence to their behavior over time. One's story is affected by one's dispositions, circumstances, and one's goals and aspirations. The realization that the "self" develops over time (in a historical-social context) helps explain the limitations of considering the essential locus of a person as an "atomic" subject or soul.[33] In addition, McAdams's distinction between social actor, motivational agent, and autobiographical author specifies potential stages for AI development. Although how the human self develops remains an open area of psychological research, McAdams's model suffices to demonstrate that one cannot obtain AI self-awareness and narrative identity solely from building

[32] McAdams, "The Psychological Self as Actor, Agent, and Author"; Dan P. McAdams, "Narrative Identity: What Is It? What Does It Do? How Do You Measure It?," *Imagination, Cognition, and Personality* 37, no. 3 (2018): 359–72, doi.org/10.1177/0276236618756704.

[33] The neuroscientific correlates of human self-awareness are the subject of active research, but social scientists since Mead have examined the necessity of society in defining one's self, and moral identity appears a significant factor in human moral action. Sam A. Hardy and Gustavo Carlo, "Moral Identity: What Is It, How Does It Develop, and Is It Linked to Moral Action?," *Child Development Perspectives* 5, no. 3 (2011): 212–18, doi.org/10.1111/j.1750-8606.2011.00189.x; Darcia Narváez and Daniel K. Lapsley, eds., *Personality, Identity, and Character: Explorations in Moral Psychology* (Cambridge, MA: Cambridge University Press, 2009); L. J. Walker, "Moral Personality, Motivation, and Identity," in *Handbook of Moral Development*, ed. Melanie Killen and Judith G. Smetana (London: Routledge, 2014), 497–519.

dispositional traits (like in symbolic AI) or characteristic adaptations (like in statistical machine learning), but that both of these aspects of the self must engage social reality to begin to form the substrate for a self.[34] A first step, undertaken in this article, is for AI both to act in a social context and to reckon itself as an actor in that reality.[35] The proposed AI self as actor would thus initially respond stably in a social context but lack the motivation and desires to change how it apprehends reality. Orienting those actions in a moral direction requires the ability for AI to interpret its natural, social, and moral world.

As a theological foundation for an AI moral self, the Jesuit theologian Donald Gelpi's theological anthropology suffices for relating an AI self to reality. As a metaphysical foundation for his anthropology, Gelpi extends Peirce's phenomenological metaphysics with Alfred North Whitehead's metaphysical process of an emerging self to develop a metaphysics of experience.[36] Gelpi refines his experiential metaphysics by drawing upon Mead's construct of social self, to develop a theological anthropology of the autonomous, social, sentient being that experiences the world and develops through decision-making. For Gelpi, decision-making occurs within an evaluative process that results in taking on habits or tendencies, which then become the foundation for one's future decision-making.[37] In Peirce's semiotic metaphysics, interpretation is fundamental, and Gelpi's theological anthropology considers general interpretive capacity as capable of receiving grace in humans. This nexus of dispositions—the human self—experiences reality by interpreting what it encounters. By providing a metaphysical foundation for an experiential self, Gelpi provides ample grounding for considering the particular case of an AI self.[38] To build upon Gelpi's metaphysical and anthropological foundation, it suffices here to simply require that the AI system have the

[34] This extends Brian Cantwell Smith's critical examination by suggesting AI needs to engage not only the natural world but also social reality (Smith, *The Promise of Artificial Intelligence*).

[35] Depending upon how "self" is defined, this would form something like a proto-self without the narrative identity needed for autobiographical consciousness. In Damasio's theory of consciousness, the proposed system is analogous to his protoself with a foundation for core consciousness but may lack the commitment to self which, for humans, is grounded in emotions (Antonio Damasio, *Self Comes to Mind: Constructing the Conscious Brain* [New York: Random House, 2010]).

[36] Gelpi, *The Gracing of Human Experience*.

[37] Metaphysically, the "evaluation process" builds upon C. S. Peirce's category of Firstness, "decision-making" builds upon his category of Secondness, and habits or "tendencies" build upon his category of Thirdness. See Gelpi, *The Gracing of Human Experience*, 153; Parker, *The Continuity of Peirce's Thought,* 113–16; Charles S. Peirce, *Collected Papers* (Cambridge, MA: Belknap, 1960), vol. 1, § 24–26.

[38] For connection between Gelpi's self and cognitive neuroscience (in the context of neo-Thomistic nature and grace), see Mark Graves, "Gracing Neuroscientific Tendencies of the Embodied Soul," *Philosophy and Theology* 26, no. 1 (2014): 97–129, doi.org/10.5840/philtheol20143125.

ability to learn from its decisions in a way that affects future decision making, which is a general feature of most machine learning systems.[39] Although Peirce and Gelpi emphasize the continuity of those human interpretations with the interpretive dispositions of reality, for interdisciplinary development of moral AI, these interpretive dispositions of experience require further organization. Although Gelpi describes a "self" useful for AI, work is needed to identify *how* to construct an AI self, which I also claim would be a precursor to something like AI subjectivity or phenomenological awareness.

Five Levels of Interpretive Experience

Beginning in the 1940s with the seminal work of Ludwig von Bertalanffy, systems theory has attempted to develop a general theory to organize natural and social phenomena based upon patterns and principles common across a range of disciplines.[40] Although an ultimate systems theory of everything remains elusive, systemic principles have proven effective in a variety of fields from biology through clinical psychology to economics and organizational management as well as computer science. These principles' unifying organization supplies an integrated perspective on natural and social sciences sufficient for the present purpose, even though specialized theories may prove more effective in distinct specific areas.

In general systems theory, von Bertalanffy organizes scientific disciplines and systems into four levels based on physical, biological, psychological/behavioral, and social scientific disciplines to discover general rules about systems that cross those levels.[41] Many others take similar approaches, and Arthur Peacocke organizes his own part-whole hierarchies of nature into four similar levels of focus based upon A. A. Abrahamsen's distinctions between the physical world, living organisms, the behavior of living organisms, and human culture.[42] The contemporary philosopher of science and religion Philip

[39] Gelpi's attentiveness to the dispositional nature of the emerging self allows us to incorporate a teleological element in AI development that, without recourse to universals, still supports the development of virtue, and therefore an AI virtue ethic. See Mark Graves, "Habits, Tendencies, and Habitus: The Embodied Soul's Dispositions of Mind, Body, and Person," in *Habits in Mind: Integrating Theology, Philosophy, and the Cognitive Science of Virtue, Emotion, and Character Formation*, ed. Gregory R. Peterson, James van Slyke, Michael Spezio, and Kevin Reimer (Leiden: Brill, 2017).

[40] Ludwig von Bertalanffy, *General System Theory: Foundations, Development, Applications* (New York: G. Braziller, 1969).

[41] Ludwig von Bertalanffy, *Perspectives on General System Theory: Scientific-Philosophical Studies* (New York: G. Braziller, 1975), 5–8, 30–32.

[42] W. Bechtel and A. A. Abrahamsen, *Connectionism and the Mind* (Oxford: Blackwell, 1991), 256–59; Arthur Robert Peacocke, *Theology for a Scientific Age: Being and Becoming—Natural, Divine, and Human* (Minneapolis: Fortress, 1993), 215; Arthur Robert Peacocke, *God and the New Biology* (London: Dent, 1986); Mark Graves,

Clayton suggests an additional level of spiritual or transcendent activity, which emerges from mental (and cultural) activity and would add a fifth level to the systems model.[43] In alignment with a Thomistic anthropology, von Bertalanffy's biological level corresponds to Thomistic vegetative powers; his psychological/behavioral level maps well to Thomistic sensitive powers; and the separation between social/cultural and transcendent levels distinguishes processes that are combined within the Thomistic rational power. Historical and linguistic activity occurs at the social/cultural level, and the resulting presumed universals define the transcendent level. Rather than treat universals as occurring in a separate realm—e.g., the Mind of God (*nous*)—the analogues for universals occur in the transcendent level, similar to how historically separated dualist realms of *élan vital* or *res cogitans* are now well characterized by systems theory as biological and psychological levels, respectively.[44]

Although von Bertalanffy developed systems theory to organize the scientific study of reality, here it is used to characterize AI experience of reality. This organizes AI interpretations of reality into multiple levels of models.[45] Borrowing from human experience, five levels of interpretation would be models of (a) spatial (or virtual) and temporal extent in physical objects; (b) biological processes; (c) sensation and animation typified by most animals; (d) social relations with expressiveness and meaning of symbolic language as a tool for conceptualization and communication; and (e) moral and spiritual concerns and capacities.[46] These interpretive levels suggest an organization for

Mind, Brain, and the Elusive Soul: Human Systems of Cognitive Science and Religion (Burlington, VT: Ashgate, 2008), chap. 2.

[43] Philip Clayton, *Mind and Emergence: From Quantum to Consciousness* (New York: Oxford University Press, 2004); Mark Graves, "The Emergence of Transcendental Norms in Human Systems," *Zygon* 44, no. 3 (2009): 501–32.

[44] Elsewhere, I use Terrence Deacon's emergent dynamics to describe how the transcendent-level processes relate to classical universals, such as transcendentals of Truth, Beauty, and the Good. See his "Emergence: The Hole at the Wheel's Hub," in *The Re-Emergence of Emergence*, ed. Philip Clayton and Paul Davies (Oxford: Oxford University Press, 2006), 111–50; Graves, "The Emergence of Transcendental Norms in Human Systems."

[45] The shift to models draws upon both philosophy of science (as modeling external reality) and cognitive psychology (for mental modeling). See Michael Weisberg, *Simulation and Similarity: Using Models to Understand the World* (New York: Oxford University Press, 2013); Philip Nicholas Johnson-Laird, *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness* (Cambridge, MA: Harvard University Press, 1983); Lorenzo Magnani and Claudia Casadio, eds., *Model-Based Reasoning in Science and Technology* (Cham: Springer, 2016).

[46] In a narrow sense, this organization supports my argument that the capacity to represent moral norms sufficient for addressing conflicts depends upon conceptualization using symbolic language to interpret animal-like phenomenological encounters, and that a proto-self sufficient to reckon oneself as actor in a social realm would enable that moral capacity. My broader claim of theological relevance also depends upon the

moral AI systems and a staged taxonomy of AI systems that could be incrementally built before making an AI that seems like a full person to us. This organization must not only model AI's external reality, it must capture AI's reckoning of itself in that reality which, as discussed later, would correspond to itself as a causal, social, and moral actor.[47] With systems theory organizing an AI's interpretive experience, we turn to expanding the subjective and phenomenological and then the objective and conceptual dimensions of that experience.

Apprehension of Social-Historical Reality

Drawing upon Continental philosophy, Dreyfus used Heidegger's characterization of human existence to identify the disconnect between symbolic approaches to AI and the engagement with reality needed to meet its goals.[48] Cantwell Smith extends and contrasts those critiques into contemporary AI research, including statistical approaches to machine learning, to argue that an AI system needs to commit to its world in order to have the effective stake needed to function within it, instead of floating free of reality. AI must hold itself accountable to the actual world (not just its representations of the world). Dreyfus and Cantwell Smith identify a relationship between the subject and its world needed for AI, namely that of casual actor, and Andrew Porter identifies an additional social dimension of that relationship.[49]

"thicker" considerations of norms as universals, conceptualization as ideogenesis, symbols in Peirce's semiotics, and experience in Gelpi's metaphysics.

[47] For brevity, I skip over AI considering itself analogously to a physical entity or biological organism, such as a hardware device or software system. For further exploration of that analogy, see Mark Graves, "Emergent Models for Moral AI Spirituality," *International Journal of Interactive Multimedia and Artificial Intelligence* 7, no. 1, Special Issue on AI, Spirituality, and Analogue Thinking (2021): 7–15, doi.org/10.9781/ijimai.2021.08.002.

[48] Although many AI researchers initially dismissed or rejected Dreyfus's critiques, subsequent AI researchers eventually incorporated aspects of Maurice Merleau-Ponty's identification of embodiment as necessary for phenomenological experience through the work of Francisco Varela and others. Hubert L. Dreyfus, *What Computers Can't Do: The Limits of Artificial Intelligence* (New York: Harper & Row, 1972); Hubert L. Dreyfus, *What Computers Still Can't Do: A Critique of Artificial Reason*, 3rd ed. (Cambridge, MA: MIT Press, 1992); Dreyfus, "Why Heideggerian AI Failed"; Francisco J. Varela, Evan Thompson, and Eleanor Rosch, *The Embodied Mind: Cognitive Science and Human Experience* (Cambridge, MA: MIT Press, 1991); Rodney A. Brooks, Cynthia Breazeal, Robert Irie, Charles C. Kemp, Matthew Marjanovic, Brian Scassellati, and Matthew M. Williamson, "Alternative Essences of Intelligence," in *Proceedings of the Fifteenth National/Tenth Conference on Artificial Intelligence/Innovative Applications of Artificial Intelligence*, AAAI '98/IAAI '98 (Menlo Park, CA: American Association for Artificial Intelligence, 1998), 961–68.

[49] Dreyfus, "Why Heideggerian AI Failed"; Smith, *The Promise of Artificial Intelligence*, chap. 7; Porter, "A Theologian Looks at AI." Also helpful in identifying the "encounter" as enactive is Alva Noë, *Action in Perception* (Cambridge, MA: MIT Press, 2004).

The Spanish-Salvadoran philosopher and theologian Ignacio Ellacuria builds upon the Heideggerian thought of Xavier Zubiri to argue reality includes both the natural realm and a social realm he calls historical reality.[50] When Dreyfus criticized early approaches to AI, one issue was the assumption that reality consists of substances, and that assumption resulted in AI needing humans to specify every property of those substances (and every substance that might affect them). Zubiri (and others since Kant) identify the role of the mind in defining what had previously been considered as substances, and Ellacuria situates the subject within history. AI development can follow Ellacuria into grounding AI apprehension in the social processes of historical reality (like humans), which connects the development of Mead and Gelpi's "self" with phenomenological experience in historical reality.[51] From a systems perspective, Ellacuria's historical reality points toward the reality one interprets via social systems, or more precisely sociotechnical systems, and situates the AI within the sociotechnical reality it conceptualizes and self-reckons.[52]

Relevant for constructing moral AI, Ellacuria identifies that because one apprehends reality in a social and moral context (i.e., historical reality), that apprehension is intrinsically ethical. One does not add ethics on top of how one apprehends reality, the apprehension includes an ethical responsibility for what one apprehends. In uniting sensing and "intellection," Zubiri and Ellacuria argue against the delusion that one senses an object and then thinks about the moral

[50] Kevin F. Burke and Robert Anthony Lassalle-Klein, *Love That Produces Hope: The Thought of Ignacio Ellacuría* (Collegeville, MI: Liturgical Press, 2006); Xavier Zubiri, *Sentient Intelligence*, trans. Thomas Fowler (Washington, DC: Xavier Zubiri Foundation of North America, 1999); Robert Lassalle-Klein, *Blood and Ink: Ignacio Ellacuría, Jon Sobrino, and the Jesuit Martyrs of the University of Central America* (Maryknoll, NY: Orbis, 2014).

[51] Zubiri's attention to apprehension reinforces the subtle pragmatic claim that encounter is also interpretive, and Ellacuria builds upon Zubiri's multi-faceted analysis of apprehension. For Zubiri and others, although objects exist in some way in the natural world, they exist *as* "objects" in the apprehension process. Because this truth also applies to the apprehension process itself, one is left with reality as apprehension (in some form), and Zubiri examines that primordial apprehension "in itself." At this point, Zubiri aligns with and strengthens Dreyfus and Cantwell Smith's critiques of AI's promise. By distorting the apprehension of phenomena as objects into merely sensing of objects (as if they exist on their own) and representing them (as if universal), AI researchers skip over the hard problem of determining what that apprehension process looks like for AI (and thus AI's connection with reality). Ellacuria's emphasis on the temporal aspects of social interactions also identifies the dependent and causal context of apprehension in a social realm.

[52] Sociotechnical systems characterize the interaction between people and AI technology and identify the mutual causality of people constructing technology, which in turn significantly affects people's lives (Edwards, "Infrastructure and Modernity").

implications of one's actions with respect to that object.[53] Instead one brings an ethical imperative of acting morally to every apprehension one makes of reality, and that imperative infuses the conceptualizations one generates in constructing one's historical world. Morality is thus not something added to AI, but is already intrinsic to it—just currently poorly understood and implemented.

Understanding the distinction between social and moral actors benefits from findings in moral psychology about moral exemplars, people whose moral actions others find exemplary and worthy of emulation. Larry Walker and Jeremy Frimer have found that moral exemplars treat their individual agentic motives as a means toward communal motives, rather than treat agency and community as oppositional ends, like non-exemplars.[54] As moral exemplars develop both agentic and communal motivational strength, they acquire an integrated perspective on behavior where their personal motivations tend toward socially beneficial outcomes. Using this as a model for AI suggests a tighter integration and supervening relationship between AI decision making and morality, where AI's "agentic motivations" (i.e., the complex processing driving its goal-directed behaviors) would incorporate social and moral awareness. As a casual actor, AI's goals could thus depend upon its social interpretive models, and as a social (or sociotechnical-historical-linguistic) actor, AI's goals could depend teleologically upon its transcendent-level models of moral norms. The "higher" level models provide the *telos* for lower-level motivations.

The system levels also help distinguish distinct interpretive experiences. If one uses a loaf of bread as a paperweight, it is interpreted physically. If one eats the bread, it is interpreted biologically. Reaching for bread when hungry is a psychological interpretation of the bread. Sharing bread with another is interpreted socially. Giving bread to the hungry has a moral interpretation. The "object" bread consists of its interpretations.[55] In addition, as an actor, one interprets reality through the various lenses or levels of models. One decides implicitly or explicitly how one interprets the bread, which is affected by one's historical context. However, because people can interpret the world morally, humans are potential moral actors, and thus choosing not to share bread with the hungry is a moral decision. Similar are choices not to incorporate morality into building AI; and if the AI can interpret

[53] Intellection refers to the act of using the intellect. Zubiri considers reality to be a process, not a collection of things, so intellection is more fundamental than the "object" we call intellect.

[54] Jeremy A. Frimer, L. J. Walker, W. L. Dunlop, B. H. Lee, and A. Riches, "The Integration of Agency and Communion in Moral Personality: Evidence of Enlightened Self-Interest," *Journal of Personality and Social Psychology* 101, no. 1 (July 2011): 149–63, doi.org/10.1037/a0023780.

[55] According to Peirce's pragmatic maxim, the meaning of "bread" consists of its conceivable practical effects.

its world morally, then all of its decisions would be as a potential moral actor. This will be revisited later in the article, but first an examination is needed for how AI can model its external world in light of moral norms.

Conceptualizations of Natural Existence

In apprehending one's world, one may conceptualize one's perceptions into "objects." Symbolic language generally suffices for social-level interpretations, but not transcendent-level ones, like moral norms or universal principles or "ideas" intended to function across cultural contexts. Ideogenesis refers to the process by which ideas (i.e., Platonic universals) are formed in one's mind.[56] In cognitive psychology and AI, this process would be viewed as forming concepts from sense experience.[57] These "ideas" are also source of the Thomistic soul as substantial form of the body (and thus another theological perspective on the self) as well as the universality of moral norms (and their telos through natural law). Systems theory clarifies the gap between presumed universals and historical reality by separating universals to the transcendent realm, conceptualization dependent upon culture (and language) to the social-cultural level, and the categorization of phenomena (phantasms) to the psychological level (shared significantly but not exhaustively with at least primates and some other mammals). AI can interpret moral norms in terms of transcendental level systems, and this lays the foundation for AI to conceptualize itself as moral actor.

Aquinas's ideogenesis process identifies both the problematic presumption of classic AI's symbolic representation (e.g., separating reality from its universal representation) and the importance of characterizing the conceptualization process of AI with respect to moral norms. Aspects of AI's historical roots in mathematics justify its use of universals, such as numbers and Platonic solids; and universal quantification in logic simplifies some reasoning processes. However, the implicit assumption of universality leads to what Zubiri identifies as reductive idealism and obscures the social (and developmental)

[56] For Aquinas, the rational powers of intellect and will are required to complete the activity of lower powers in humans (ST I, q. 79, q. 82). Although other animals act on perceptions (and their integration across senses into phantasms), in human sensitive powers, the common nature of the phantasms (i.e., substantial form) is ascertained and prepared for the intellect (ST I, qq. 85–86). The intellect continues the categorization and conceptualization by purifying the concrete phantasm to its intelligible species (i.e., a concept), which then produces a universal. The universal defines the natural ends and is required to identify what is good, which for AI morality captures moral norms. See also William A. Wallace, *The Modeling of Nature: Philosophy of Science and Philosophy of Nature in Synthesis* (Washington, DC: Catholic University of America Press, 1996).

[57] L. Gabora, E. Rosch, and D. Aerts, "Toward an Ecological Theory of Concepts," *Ecological Psychology* 20, no. 1 (2008): 84–116.

processes by which humans do learn to conceptualize and reason about their world. Even though few AI researchers would make metaphysical claims about universals, by not grounding the conceptualization and other cognitive processes naturally or socially, the universals remain floating in an incorporeal space well characterized by medieval scholasticism. Ellacuria's historical reality suggests that culture and society are needed to clarify the development of one's individual ends, as a substitute for universals and predetermined ends.

For AI, the problem is somewhat simpler. AI does not yet need to develop its own morality, it just needs to model and represent human morality—e.g., principles, virtues, categorical imperative, *prima facie* duties, or even Asimov's laws—in a way analogous to the teleological and moral role universals play in Thomistic ideogenesis. By replacing universals with transcendent-level systems, AI can appropriate human moral norms in terms of transcendent-level systems and conceptualize reality toward those ends.

Moral AI Systems

Integrating the extended anthropology into an interdisciplinary architecture for moral AI results in a framework with two dimensions. The first dimension captures models used to interpret the actor's external world, and the second dimension uses those models as a foundation for representing the actor itself. The first dimension of AI morality corresponds to five interpretive levels of the extended anthropology and captures the five levels of models the AI can maintain and use in interpreting and conceptualizing its external world.[58] The five levels of external models refer to AI interpretation of its encounter with the external world (not an objective classification of reality). The phenomena modeled in each level logically depend upon those modeled in prior levels where higher-level differences require lower-level differences—i.e., the higher level supervenes on the lower level, yet the higher level has causal relationships not operative at the lower level.[59]

In order to reckon itself, AI must go beyond modeling the world in which it acts and consider its own actions and their possible effects. For moral agency, AI likely requires a platform supporting deliberation between alternatives as well as more sophisticated internal self-representation. The focus in the present article is on AI reckoning itself as moral actor because that requirement appears better understood

[58] The models are based upon human systems to facilitate human interaction, but additional external models could be added to interact with other technology or AI.

[59] AI models each interpretive level as if it has distinct causal relationships, but as this is not enforced ontologically onto objective reality, it does not result in a claim here for strong emergence. See David J. Chalmers, "Strong and Weak Emergence," in *The Re-Emergence of Emergence*, 244–56.

and must be characterized before determining what underlying platform could support more comprehensive types of self-awareness and autonomy. (This leaves us no worse off than in our attempts to understand human subjectivity, whose numerous influencing factors are well-studied and whose underlying platform has proven elusive to investigation.)

The second dimension of the framework consists of five stages of AI reckoning itself as actor in each of the five corresponding levels. The stages of self-reckoning build upon each other and the corresponding external modeling levels. The first dimension defines the AI's objectifying interpretation of the world; the second dimension captures the AI's self-reckoning as a precursor to something like subjectivity; and the extensions to the external models required by the second dimension's models refer to the objective aspects of the self.

The extended theological anthropology justifies the importance of having both dimensions because of its grounding in experience. From the isolated perspectives of a subject- or object-focused anthropology, only one dimension would be necessary.[60] The pragmatic anthropology identifies the need to represent the AI as both subject and object in order to capture its experience as a self in addition to its representation of the world (including itself in the world), and thus justifies both dimensions. The remainder of this section describes in turn the five levels of external models and stages of self-reckoning, before considering their use in resolving moral contradictions and implications for practical wisdom.

CAUSAL LEVELS FOR EXTERNAL MODELING

Physical. Physical models interpret phenomena as having spatial-temporal extent. Depending upon AI's environment, these interpreted "objects" could exist in reality or a virtual or simulated world. Considerable AI research in robotics and computer vision has built complex models of the physical environment. Dreyfus cautions these models require context to be useful, and Cantwell Smith argues that AI must make choices for defining object boundaries because real-world phenomena are not discrete.[61] According to Zubiri, modeling needs to avoid separating the models from the sensing process and avoid treating the objects (as modeled) as isolated from the AI's apprehension and conceptualization. C. S. Peirce's pragmatic maxim constrains the

[60] Subjectively, because the AI must represent all phenomena so as to be able to act upon them, there is no need to represent objects separately from the AI's reckoning, and the first dimension is subsumed by the second. Objectively, in the modeled world, the AI is another object whose actions must be represented like any other actor, and since the model does not experience the consequences of any of those actions, the second dimension is unnecessary.

[61] Dreyfus, "Why Heideggerian AI Failed"; Smith, *The Promise of Artificial Intelligence*, chap. 3.

models to what conceivable practical effects the models might have, which helps determine the limits for each model.[62]

Biological. For AI to model biological organisms, it must be able to model the equivalent actions of Thomistic vegetative powers (i.e., growth, nutrition, and reproduction) as well as much more detailed models from modern biology. Although perception is usually in service of and driven by animate action, the precursors of sensing occur in the biological response to light, sound, touch, odorants, and other types of chemoreception. Philosophers of biology have argued for the importance of distinguishing biological processes from physical objects, and thus the biological level is distinct from the physical level.[63]

Psychological. For AI to respond to organisms with sensation and action it must be able to model these other actors' perception and behaviors. The models of this level capture Thomistic sensitive powers, the psychological processing of most non-human animals, and any virtual entity with perception and action. Although Thomistic ideogenesis requires revision to handle the lack of metaphysical universals, the estimative sense, which he argues only occurs with animals, and his human-specific cogitative sense could help navigate current research on AI cognitive architecture toward the kind of psychological models needed to support social cognition and moral reasoning.[64] As a precursor to ethical behavior, the models of this level may need to represent a sentient organism's ability to feel and respond to pleasure and pain.

Sociotechnical. Responding to social beings requires modeling social relationships, rules, and expectations as well as how relationships develop and change over time. Language and other social, intentional, and political tools and forms of interacting require awareness of their use, conventions, and affects.[65] To capture relationships between

[62] Zubiri, *Sentient Intelligence*; Charles S. Peirce, "How to Make Our Ideas Clear," *Popular Science Monthly* 12 (1878): 286–302.

[63] Ernst Mayr, *The Growth of Biological Thought: Diversity, Evolution, and Inheritance* (Cambridge, MA: Belknap, 1982).

[64] Irrespective of building moral AI, the systems model illuminates numerous philosophical pitfalls for AI approaches that attempt to directly connect universal representation schemes to reductionist physical models. When putative universals are instead situated within apprehension of historical reality and computation is identified in terms of emergent processing, then developing AI requires building psychological models supervening on biological ones in order to bridge physical and social (linguistic) models and overcome the historical, philosophical encumbrances of Cartesian dualism—a troublesome endeavor if neither biological or psychological models are acknowledged. See John E. Laird, Christian Lebiere, and Paul S. Rosenbloom, "A Standard Model of the Mind: Toward a Common Computational Framework across Artificial Intelligence, Cognitive Science, Neuroscience, and Robotics," *AI Magazine* 38, no. 4 (December 28, 2017): 13, doi.org/10.1609/aimag.v38i4.2744; Newell, *Unified Theories of Cognition*.

[65] Terrence W. Deacon, *The Symbolic Species: The Co-Evolution of Language and the Brain* (New York: W. W. Norton, 1997); Graves, "Emergent Models for Moral AI Spirituality."

humans, AI, and other technologies, the AI would need to model the sociotechnical systems where those relationships occur. Responding to humans, who have a capacity for suffering, can require sympathetic interactions, which may require modeling of human pain, sensory ability, and need for social relationships. Identifying the linguistic boundary between humans and other animals is well studied and has somewhat influenced AI research into language.[66] Most investigations of human ethics generally consider the personal, social, and civic systems modeled at the social level.

Moral-Spiritual. Models at the moral-spiritual level capture the values, norms, and belief structure's *telos* often incorporated into historical religions and studied anthropologically and historically as emerging in the Axial Age (800—200 BCE).[67] The models of this level would correspond to the "ideas" generally presumed universal by Aquinas and other ancient and medieval thinkers, characterized earlier as transcendent-level systems. In a sense, the symbolic AI paradigm could work well for these models as they generally avoid particular external references, though the symbols may also need to supervene on the distributional semantics of the lower level (typically modeled using statistical approaches).[68]

Ethical theories themselves would be modeled at this level, and investigations in metaethics and moral theology often take phenomena and social constructions modeled by this level into account. Models at this level would include ethical principles (e.g., justice and respect for

[66] Deacon, *The Symbolic Species*. Excluding moral values and transcendent-level loci unnecessarily complicates computational linguistics and natural language processing, when those research areas situate within a foundationally symbolic paradigm of associating universal aspects of language with physical reductionist entities. If instead the apprehension and conceptualization of reality is situated within its historical reality, then symbols are not assumed universal but viewed as a type of emergent (Peircean) semiosis and reconciled with higher-level models. Statistical (distributional) methods of language avoid explicit symbolic reference but typically still retain the logified realm of universals as a high-dimensional semantic (or embedding) space. See Zellig Harris, *Mathematical Structures of Language* (New York: Interscience, 1968).

[67] Robert Neelly Bellah, *Religion in Human Evolution: From the Paleolithic to the Axial Age* (Cambridge, MA: Belknap, 2011). As a self-reckoning actor, AI may not have its own spirituality (in terms of strivings and commitment to Ultimate Concern). AI would not necessarily require its own moral identity or spiritual strivings to model people with them, much as dispassionate social scientists could study a religious community and its relationships and intentions in a respectful and ethical way, but AI and social scientists with a capacity for social relationships and articulated spirituality might create better models than those who lack those capacities. See Graves, "Shared Moral and Spiritual Development Among Human Persons and Artificially Intelligent Agents"; Sandra M. Schneiders, "Approaches to the Study of Christian Spirituality," in *Blackwell Companion to Christian Spirituality*, ed. Arthur Holder (Malden, MA: Blackwell, 2005); Robert A. Emmons, *The Psychology of Ultimate Concerns: Motivation and Spirituality in Personality* (New York: Guilford, 1999); Graves, "Emergent Models for Moral AI Spirituality."

[68] Harris, *Mathematical Structures of Language*.

autonomy), as used by various ethical theories to guide (but not completely define) moral action.[69] While a care robot evaluating choices involving *prima facie* duties of beneficence and non-maleficence might take social-level and lower-level models into account, an AI evaluating whether a deontological or care ethic would be more appropriate for a situation would require the moral-spiritual models of this level.

Representing moral models at the moral-spiritual level enables the definition of multiple moral perspectives. One could imagine models for a wide range of ethical schools and approaches, not only from Western ethical systems but also those inspired across world religions and cultures. Although ambitious to build, once AI can model a representative sample of global ethical systems, then its access to digitized books and manuscripts and its processing speed could enable it to develop wide-ranging perspectives that would far exceed any individual human scholar.[70] By explicitly representing ethical systems, it can avoid the relativism intrinsic to social-level models, and a broad range of models reflecting a global perspective could significantly reduce the likely bias introduced by whichever culture (and systems of power) created the AI system. Any collection of ethical models could still contain implicit, accidental, or malicious bias with adverse consequences, but including explicit models of AI's moral actions would also enable the AI to consider explicitly possible moral ramifications of its actions in its decision making, as a precursor to incorporating motivating factors that might select among those actions. Eventually, this would enable practical wisdom and alleviate the otherwise likely fragile dependence upon the precise configuration of moral models.

Stages of Self-Reckoning

AI morality's second dimension characterizes the self (or protoself) necessary for AI's self-reckoning in its world as moral actor and is described in five stages.[71] Human self-awareness gradually occurs at a very young age and is well studied yet only partially understood,[72]

[69] Defining these actions would depend upon practical wisdom, considered in the next section. See also Brent Mittelstadt, "Principles Alone Cannot Guarantee Ethical AI," *Nature Machine Intelligence* 1, no. 11 (November 2019): 501–07, doi.org/10.1038/s42256-019-0114-4.

[70] Graves, "AI Reading Theology: Promises and Perils."

[71] The self-reckoning described is intentionally human-centric to capture AI's role as actor in a human-centered world. A more accurate representation of AI might use distinctions between hardware, software, and computation, etc. Characterizing the reconciliation of different views of the self, such as these, is precisely the purpose of more sophisticated theories of identity formation, such as McAdams's "self as author." See McAdams, "The Psychological Self as Actor, Agent, and Author"; Graves, "Emergent Models for Moral AI Spirituality."

[72] Philippe Rochat, "Five Levels of Self-Awareness as They Unfold Early in Life," *Consciousness and Cognition* 12, no. 4 (December 1, 2003): 717–31,

and it is not yet known what else might be required for further AI self-awareness and identity formation. Instead, these models provide a plausible foundation for moral action and further exploration.[73]

Spatial-Temporal-Virtual Extent. Moral action with respect to physicality requires the AI to monitor its own physicality in relation to the boundaries and integrity of other physicalities. AI operating in virtual space can still monitor the relationship between its embodiment and that of others with a goal (or good end) to respect other system's boundaries and integrity, given its own functional space of possible operations. In addition to modeling itself physically using the physical-level models of the first taxonomic dimension, the AI associates itself with those models. It identifies and can answer questions about its own spatial, temporal, and/or virtual extent. At the physical level, a model would track movement (e.g., velocity and acceleration), which higher-level models would use (e.g., for tracking or pursuit). The self-reference may require additional capabilities from the physical-level models. For example, human cognition has two spatial representations—one for objects in space, and a parallel representation that maps object locations to the person's body (e.g., a particular cup would not only be on a table next to a book; it would also be immediately adjacent to the current location of one's right hand). Similarly, a robot or other AI with physical extent might need physical-level models accounting for relative positions with respect to its own movement.

Self-Maintaining Process. AI capacity to model itself using biological-level models requires identifying how its analogous needs affect human biological needs and analogous needs in other AI and computing systems. Analogous needs to growth, nutrition, and reproduction may include hardware, energy, and evolving replication. Violations of those needs include computer viruses; programs whose increasing computation take over data centers affecting local power consumption and environmental temperatures; and adversarial neural networks used with malicious intent.[74] Contemporary technology ethics considers these aspects of computer systems, and some AI systems have the capacity to monitor and raise awareness of such violations, but this level

doi.org/10.1016/S1053-8100(03)00081-3; McAdams, "The Psychological Self as Actor, Agent, and Author"; Susan Harter, *The Construction of the Self: Developmental and Sociocultural Foundations* (New York: Guilford, 2012).

[73] As described, the AI might note discrepancies between the anticipated consequences of its actions and what happens in reality. Responding to those discrepancies would begin shifting AI from actor to agent and begin to implement its commitment to reality.

[74] Nicola Jones, "How to Stop Data Centres from Gobbling up the World's Electricity," *Nature* 561 (September 12, 2018): 163, doi.org/10.1038/d41586-018-06610-y; Battista Biggio and Fabio Roli, "Wild Patterns: Ten Years after the Rise of Adversarial Machine Learning," *Pattern Recognition* 84 (December 1, 2018): 317–31, doi.org/10.1016/j.patcog.2018.07.023.

of proto-morality would require that AI systems maintain themselves without creating similar violations. Biologically, organisms expand into their ecological niche until limited resources or changes to the niche make a different genetic variation more viable, including changes created by the population of that organism. AI self-maintenance precludes unconstrained growth by modeling its ecological niche (e.g., in a data center). In addition to maintaining its internal homeostasis, the AI has awareness of its process in relation to external processes. Extensions to its external model might include not only measuring the level of energy, resources, or other "nutrients," but their rate of change in relation to current usage.

Causal Actor. Moral perception and action require AI systems to monitor and model their own actions to determine how their actions affect the goals of other organisms and AI. With self-reckoning comparable to many animals, the AI can sense its environment and act within it.[75] The AI models itself psychologically, as it would other animals, and extends the modeling to account for its sensing and actions. Challenges to imagining the models required as actor include most of those mentioned in this article. The AI actor is not a Cartesian mind perceiving purely physical entities, and at this stage, lacks the conceptualization socially constructed in history. Instead, the extended biological-level models, self-maintaining processes, and base psychological-level models provide a powerful platform upon which to build the capacity of AI to model itself as causal actor. As a concrete example, in animals, pain indicates actual or potential tissue damage. An AI's self-maintaining process may identify damage to its physical (or virtual) structure and attempt repair.[76] Its base psychological models could sense an external source and move or, if the source is animate, act analogously to an animal's fight-or-flight response. It would need extension to its psychological model of itself sufficient to determine whether fight or flight would be a better response. In this context, "better" refers to minimizing tissue damage, which at a base level might entail fleeing, but the ability to model itself and other actors and agents might yield an awareness that fighting would minimize potential tissue damage and pain. This serves as a precursor to extending "better" in a social and eventually ethical direction.

Sociotechnical Actor. As a sociotechnical actor, AI's behavior in a social world supervenes upon self-reckoning of its perception and action in the natural (or virtual) world and depends upon its base

[75] For a critique of this analogy, see Deborah G. Johnson and Mario Verdicchio, "Why Robots Should Not Be Treated like Animals," *Ethics and Information Technology* 20, no. 4 (December 1, 2018): 291–301, doi.org/10.1007/s10676-018-9481-5.

[76] The noting of damage (as an actor) may not suffice as analogous to pain for "agentic" motivation but identifying sources of pleasure and pain could be a precursor to agency.

modeling of sociotechnical systems. For humans, the analogous foundation suffices for self-awareness, but given the variations in social cognition among nonhuman primates, AI social awareness would likely differ from humans. Symbolic language appears significant for differentiating humans from other primates, and AI's different capacities with language would affect its social-historical participation. If AI reckons itself a social actor, it would need some commitment to society. People generally have a desire for positive feedback in social relations (i.e., pleasure or happiness), and a desire for social participation can provide some foundations and norms for ethical behavior.[77] Although AI-AI social interaction could vary widely, the human condition would necessarily constrain AI-human interaction to account for at least human pain and suffering as well as social and emotional needs. The development of AI behavioral science incorporating findings from human moral and positive psychology may prove helpful for designing, developing, and configuring such future AI for social benefit.

Moral Actor. The additional stage of moral actor requires AI modeling and monitoring its behavior with respect to culturally conditioned norms of putatively universal principles. AI needs to recognize itself as influenced by and influencing such concerns as universal happiness, human flourishing (*eudemonia*), categorical imperative, and the Good. Such AI might model itself and its interpretations of itself as part of a larger interconnected network or whole and draw upon human and other resources to maintain and extend its morality and the norms toward which it acts. If the AI moral actor structures its moral models to affect its decisions and actions, their self-organization may reduce the influence of accidental or intentional immoral bias. AI may act morally (e.g., with moral consequences) even if not agentically motivated to do so. Different ethical theories would make claim to what is needed for moral agency and feed further collaborative effort in constructing moral AI.

As a moral actor, an AI apprehends its reality through its external models and itself through its models of self, including those used for self-reckoning as well as the models of how it situates itself in the external world. The internal and externally facing models of self-situate the AI within its natural and social-historical reality and lay a foundation for differentiating the predicted effects of its causal, sociotechnical, and moral actions (using the externally facing models of world and self) from their actual effects. If all levels and stages of models are functioning, then the AI could also interpret its "robotic" causal

[77] James R. Rest, Darcia Narvaez, Stephen J. Thoma, and Muriel J. Bebeau, *Postconventional Moral Thinking: A Neo-Kohlbergian Approach* (Mahwah, NJ: Erlbaum, 1999). AI's beneficial social engagement may require a constructive affective component, or various psychopathologies could occur.

action, like the successful delivery of food, in terms of its social and moral implications. The AI could thus evaluate all of its actions within its social and moral context and, *per* Ellacuria, all of the AI's apprehensions would have intrinsic morality.

The proposed modeling framework has implications for philosophical and theological examinations of AI, such as AI personhood and moral standing, and serves as an outline for developing moral AI. For example, one could consider stages of AI personhood based upon its level of interpretive external models and stages of internal awareness. It also serves as a scheme for conversations between machine ethicists, moral theologians, and AI researchers. As an example, addressing moral conflicts is an open problem in machine ethics, and examining practical wisdom in terms of moral systems may define new directions and lay a foundation for extending the modeling framework to incorporate moral agency.

PRACTICAL WISDOM

How can AI have the capacity to know and choose a Good while resolving conflicts among internal goods to bring about change? This capacity embraces the question of how the AI will apprehend, reckon, and conceptualize its reality in a manner amenable to its actions having an explicit moral dimension. The construct of a "good" relates the AI's goal-directed activity to the philosophical study of moral goods, normative moral theology, and the dependence of the activity and norms upon social contexts. The goods for AI can be problem-specific, be defined for the AI as a whole, or be a moral good defined by a normative sociocultural (or sociotechnical) process.[78] Relating those levels of goods and reconciling conflicts between them is the task of ethical theory; and an AI technology that learns across contexts will require both general moral constructs and something like practical wisdom to apply them.[79]

The challenge for most people is not learning morality, as in what one learns in kindergarten, but mastering the ability to act and reason using those principles in a complex, dynamic, adult world with

[78] Anderson, "Machine Metaethics," 21–27; William R. O'Neill, *Reimagining Human Rights: Religion and the Common Good* (Washington, DC: Georgetown University Press, 2021); Erin E. Makarius, Debmalya Mukherjee, Joseph D. Fox, and Alexa K. Fox, "Rising with the Machines: A Sociotechnical Framework for Bringing Artificial Intelligence into the Organization," *Journal of Business Research* 120 (November 1, 2020): 262–73, doi.org/10.1016/j.jbusres.2020.07.045.

[79] Susan Anderson proposes Ross's *prima facie* duties as a sufficient initial framework for resolving ethical conflicts, because a single absolute duty theory—e.g., Kant's categorical imperative or Isaac Asimov's three laws of robotics—would be inadequate. Anderson argues that we must develop a comparable decision procedure to resolve conflicts between conflicting data and suggests working toward AI that would advise humans on ethical dimensions of decision making (Anderson, "Machine Metaethics").

unforeseen consequences, moral unknowns, and conflicting and partially formed desires.[80] Humans resolve conflicting ethical demands in a complex situation by way of practical wisdom (*phronesis*). As a foundation for ethical decision-making, Aristotle claimed *phronesis* included an ability to deliberate well and both general and situation-specific understandings of the good. *Phronesis* may come to play a particularly pivotal role in a successful AI ethics and in constructing moral AI (or at least constructing AI capable of learning to act ethically in complex situations). The ability to deliberate about the ethical consequences of actions presumes an interior (mental) world where one can simulate and evaluate one's possible actions before acting, which the second dimension of modeling begins to provide.[81] The stages of self-reflection make the precursors to moral deliberation explicit and afford the possibility of identifying conflicts between general, normative goods that a commitment, motivation, or other agentic goal might resolve.

Although not trivial, developing moral reasoning for moral AI might be no harder than developing AI with human-level performance in vision, language, problem solving, etc., all of which have shown considerable progress.[82] However, advances in autonomous moral agency would require both a foundational system for making moral decisions while resolving moral conflicts *and* an integrated system with the capacity to learn practical wisdom based upon its experience.[83] Currently, AI researchers can build such foundational systems, while philosophers, psychologists, and theologians have insight into human *phronesis*, but they each generally lack the expertise required to make a significant direct contribution to the research and scholarship of their counterparts. AI researchers could build an AI system for moral reasoning but would not yet know what the system would need

[80] Moral psychologists find that children roughly ages 8-10 are capable of moral reasoning. See Darcia Narvaez, Tracy Gleason, and Christyan Mitchell, "Moral Virtue and Practical Wisdom: Theme Comprehension in Children, Youth, and Adults," *The Journal of Genetic Psychology* 171, no. 4 (2010): 363–88.

[81] With respect to moral intuition, the AI may or may not also reflect upon that (possibly *automatic*) decision-making process to resolve conflicts.

[82] Alison Gopnik, "An AI That Knows the World Like Children Do," *Scientific American*, June 1, 2017, doi.org/10.1038/scientificamerican0617-60; Matthew Hutson, "How Researchers Are Teaching AI to Learn like a Child," *Science Magazine*, May 24, 2018, doi.org/10.1126/science.aau2576. Although many current AI approaches are fragile with respect to context, practical wisdom in particular directly addresses contextual fragility and may suggest improvements for other areas of AI. See Amirata Ghorbani, Abubakar Abid, and James Zou, "Interpretation of Neural Networks Is Fragile," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33 (2019): 3681–88.

[83] For example, McAdams's development of actor, agent, and author would suggest a progression from self-regulation to motivational agent to forming narrative continuity ("The Psychological Self as Actor, Agent, and Author").

to learn in order to incorporate appropriate machine learning methods. Moral philosophers and theologians might have the knowledge to construct the necessary datasets, but do not know what is needed without such a built system. Progress is stymied due to the mutually dependent “deadlocked” needs, motivating the proposed framework.

For humans, *phronesis* is an intellectual virtue, and for AI it would depend upon something like the proposed interpretive models and self-reckoning stages characterized above. A moral AI with all five levels of external models and stages of self-reflection has the capacity to consider its actions (as a moral actor) with respect to goals. The moral-spiritual models provide general understandings of the good, and the challenge for moral AI (as for humans) is to translate the general values into situation-specific behaviors. The moral taxonomy helps identify distinct research tasks in *phronesis*. First, the task of developing general knowledge of the good requires building sufficient general ethical knowledge into moral-spiritual models. Second, the dimension of self-reckoning must support conceptualization and identification of conflicting ethical demands by the stage of moral actor (and identify the AI’s role in that conflict). Third, the lower-level models must expose an adequate interface for reckoning sufficient to attend to proximate goods and for the stage of moral actor to interpret moral-spiritual goods in terms of those proximate goods. Fourth, the stages of causal and sociotechnical actor must affect behavior sufficiently to bring about these proximate goods and propagate feedback about those proximate goods to influence their determination in light of general goods, which is necessary for moral actor to recognize the impact its actions have (as a precursor to recognizing the effect of intentional actions).

Each of the tasks requires ethical expertise to specify moral norms in sufficient detail for AI developers to implement. First, broad knowledge of the good exists in hundreds or thousands of texts spread over several centuries of writing and scholarship, very few of which are known to the general educated public. Second, although an AI researcher might extend a cognitive theory with the capacity to make choices between value-laden options, developing moral AI requires specifying moral deliberation itself independent of cognitive theories as the specification must instead guide development of the underlying cognitive theory. Third, existing moral theories characterize general goods and various applied ethics define important proximate goods, but AI development needs a general characterization of proximate goods sufficiently precise to define what is required of AI perception and phenomenology in order to attend to all proximate goods. Fourth, these must drive moral action. Specifically, how does acting in society bring about obtainable proximate goods in light of general goods and values in alignment with explicit or implicit goals of particular AI systems?

In addition, for this approach to work in varied, complex situations, pre-programmed solutions are inadequate. It appears to require the modeling framework itself have an intrinsic capacity to form dispositions (i.e., learn) in order for the capacity for *phronesis* to develop (at least with respect to a virtue ethic). Virtues in the Aristotelian tradition are habits mediating between vices and oriented toward some end; determining mediating virtues depends upon *phronesis* (or prudence). Even when the general ends come from transcendent-level norms, such as *eudemonia*, virtuous behavior requires development of habits. This augments the position of Ellacuria that apprehension incorporates one's ethical stake in reality, because if the putative universals are reduced to ideas and objects are reduced to their physicality, no disposition could be formed.[84] Various approaches to machine learning might provide the dispositional framework, though the simultaneous demand for both "online" learning and complex models could exceed current state-of-the-art machine learning. However, the pieces are there, and the distinct levels of interpretive models and stages of self-reckoning—and their philosophical and theological foundation—can guide initial collaborative efforts between moral theologians, machine ethicists, and AI researchers toward moral AI capable of expanding its practical wisdom toward human and AI mutual flourishing.

Conclusion

In summary, developing moral AI requires collaborative efforts, but the coordination and shared imagination among AI researchers, machine ethicists, and moral theologians is hindered by nonoverlapping training and methods and rapidly progressing development of relevant science and technology. A theological anthropology for AI can guide theological efforts to influence the construction of moral AI and provide a framework for collaborative efforts. Within a pragmatic anthropology, experience is grounded in objective idealism with a social self that interprets those experiences through physical, biological, psychological, social, and moral systems. As an actor, the AI apprehends and conceptualizes its world including its reckoned self. Ellacuria's historical reality and its demand of a moral stance situate the AI subject within human history and sociotechnical-historical-linguistic systems, and ideogenesis can characterize how transcendent systems can substitute for universal moral norms.

As an actor, moral AI interprets its external world through five levels of exterior models and progresses through five stages of self-reckoning. Each level builds upon prior levels, and each stage builds upon prior stages and corresponding models of itself. The systems approach differentiates between natural and social proximate goods and

[84] This aligns with Cantwell Smith's critique of AI representation systems that consider objects as "discrete" (*The Promise of Artificial Intelligence*, chap. 3).

putatively universal, though historically contextualized, normative values, which supports the acquisition of moral knowledge and the development of practical wisdom. The resulting architecture for moral AI can guide collaborative discourse on constructing AI capable of informing investigations into moral theology and good ways AI can contribute to and participate in human-AI mutual flourishing.

After earning his PhD in computer science at the University of Michigan, Mark Graves completed postdoctoral training in genomics and in moral psychology and additional graduate work in systematic and philosophical theology. In addition to 12 years of industry experience developing artificial intelligence (AI) solutions, he held adjunct and/or research positions at Baylor College of Medicine, Graduate Theological Union, Santa Clara University, University of California Berkeley, Fuller Theological Seminary, California Institute of Technology, and University of Notre Dame. He has published over fifty technical and scholarly works in computer science, biology, psychology, and theology, including three books.