Theological Foundations for Moral Artificial Intelligence

Mark Graves

mgraves@nd.edu

Abstract

A theological anthropology for artificial intelligence (AI) can improve the increasing integration of AI within society by clarifying uncertainty about AI in relation to human nature. To help coordinate the different, underlying philosophical assumptions made by scholars, engineers, and social scientists involved in AI development or study, three theological anthropologies are adapted for AI drawing upon Continental (Heideggerian), Thomistic, and pragmatic philosophy to focus on AI subjectivity, soul, and self, respectively. Within that multi-faceted anthropology, reconciling Xavier Zubiri's apprehension of reality with Thomas Aquinas's ideogenesis addresses AI's dualist and reductionist barriers to meaningful conceptualization and interpretation of experience. A theological anthropology for moral AI integrates Ignacio Ellacuria's ethical stake in the apprehension of historical reality into a systems framework capable of modeling AI's external reality and internal self-reflection at multiple levels of awareness. Modeling AI's interpretive experience and internal awareness of its causal and moral agency can help moral AI resolve conflicts between its normative values (e.g., *prima facie* duties) and develop the practical wisdom (*phronesis*) needed to apply its general moral models. [173 of 150-200 words]

Keywords: artificial intelligence, machine ethics, phronesis, systems theory, theological anthropology

Note to Reader

This paper is planned for a special issue of *Journal of Moral Theology* on Artificial Intelligence and Machine Learning. I'd appreciate any comments, suggestions, or other thoughts you have on this draft.

11,800 words [of 9,000–14,000 words]

Introduction

How can moral theologians improve the increasing integration of artificial intelligence (AI) within society? Advances in AI technology raise fears and hopes as the emerging, person-like characteristics of AI call into question long-held religious and secular assumptions about human nature and increase uncertainty about how AI will affect humanity's future. Moral theologians can clarify those fears and hopes, address underlying questions about human nature in relation to AI, and contribute to the development of moral AI by defining appropriate functional norms for the rapidly expanding technology. Investigating constructive paths for AI's continued, realistic contribution to society requires both technical knowledge and moral insight. In the theologian and computer scientist Noreen Herzfeld's examination of AI and human spirituality, she identifies a central role for theological anthropology and focuses on the parallels between the Christian doctrine of *imago Dei* and human desire to create AI in our image, and others have also begun examining the relationship between AI and the human person.[1] In moral theology, a clear understanding of plausible AI personhood can help address social concerns about AI technology and contribute plausible frameworks for normative moral behavior to AI developers who would otherwise consider developing an ethical system for AI behavior to be daunting and morally hazardous. Collaborative

[1] Noreen L Herzfeld, *In Our Image : Artificial Intelligence and the Human Spirit* (Minneapolis, MN: Fortress Press, 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, https://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.

engagement on the development of moral AI can prescribe key components for AI development and guide ongoing efforts in incorporating ethics into AI.

The present article unfolds in three main parts. First, a foundational issue in attempting such an investigation is considered, examined, and addressed. Any individual's imaginative contribution is necessarily limited when the object of study can redefine what it means to reason or make evaluative judgments and is changing too fast for anyone to completely keep up with its advances. As one's hopes and fears can affect one's approach to unknown unknowns, examining optimistic and pessimistic perspectives on human nature and AI technology in society can guide the social imagination for moral theologians into a direction for collaborative development.

Second, a theological foundation for moral AI requires a theological anthropology, and three plausible (albeit partial) theological anthropologies for AI are proposed based upon existing theological anthropologies for humans with a focus on subjectivity, soul, and self as needed for moral conceptualization and action. The first anthropology draws upon Continental philosophy and extends Heideggerian AI with insights from Xavier Zubiri and Ignacio Ellacuria to examine subjective awareness of AI situated in an historical context. The second anthropology reinterprets Thomistic understanding of soul, as form of the body, and its vegetative, sensitive, and rational powers using natural sciences and systems theory to examine how moral AI can conceptualize its world. These theological anthropologies provide important, though partial, perspectives on AI and are synthesized into a third anthropology that reconciles their respective subjective and objective assumptions to characterize AI interpretive experience by drawing upon social scientific understanding of the self and pragmatic

philosophy. The pragmatic and semiotic construct of interpretive experience integrates subjective awareness and objective conceptualization sufficient to study AI that interprets its experience of its natural, virtual, and social world and evaluates possible actions through a moral lens.

In the third part, the insights from considering what is needed to adapt the three human theological anthropologies for AI leads to a collaborative framework for developing moral AI. Interpretive experience in moral AI is characterized by five levels of models, which characterize AI's encounter with an external world, and five corresponding stages of internal awareness, where AI models itself. The multi-faceted, multi-level framework characterizes and relates broad disciplinary needs for moral AI. The implications of the models are then briefly examined with respect to practical wisdom (*phronesis*) as essential to moral AI.

Imagining Moral AI

Two foundational challenges

There are at least two challenges to attempt to imagine moral AI for interdisciplinary investigation. The first challenge to scholarly investigation of moral AI is that the relatively non-overlapping educational training of (and scholarly venues for) engineers, neuroscientists, and moral philosophers and theologians severely limits the construction of robust theories incorporating both advanced technical understanding and scholarly insight. The second challenge is that the rapidly progressing developments in both AI and cognitive neuroscience repeatedly and quickly alter the capabilities of AI and a scientific understanding of human nature making it essentially impossible for individuals to imagine likely, realistic, comprehensive configurations for moral AI

technology. Together, the challenges paradoxically require and hinder collaborative, integrative efforts, but their closer examination suggests a possible path forward.

One can trace recognition of the first challenge to C.P. Snow's 1959 identification of two cultures separating science and the humanities.[2] 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 from another discipline. Ian Barbour and others have previously studied challenges to dialogue between theology and natural science, and studying AI morality also requires integrating that discourse with its related technology and ethics dialogue, previously viewed primarily as applications of science and theology, respectively.[3] However, that integration, for the case of AI morality, reverses the previously noted distinction between scholar-scientist and practitioner by defining the specific application area of technology to be an engineered system that threatens to replicate the experience and intellectual expertise previously presumed the exclusive purview of scientists and theologians.[4] In addition, the social sciences must also be incorporated as they play an important role in identifying the social structures that AI impacts and disrupts as well as explaining the human psychology that AI partially purports to replicate and with which AI often must

[2] C P Snow, *The Two Cultures and the Scientific Revolution* (New York,: Cambridge University Press, 1959).
[3] Ian G Barbour, *Religion and Science: Historical and Contemporary Issues* (San Francisco: HarperSanFrancisco, 1997); Ian G. Barbour, *Ethics in an Age of Technology* (San Francisco: HarperSanFrancisco, 1993).
[4] 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 Publishing, 2019), https://www.springer.com/us/book/9783030167998; Mark Graves, "AI Reading Theology: Promises and Perils," in *AI and IA: Utopia or Extinction?*, vol. 5, Agathon Journal (ATF Press, 2018).

interact. Because AI fundamentally relates to human reason and experience in a way no previous technology has, it novelly depends upon and can impact every field that studies or depends upon human reason or experience. Studying AI morality not only requires novel integration of humanities and natural and social sciences, it can also require examining the presumptions and historical accidents that led to their separation.

The second challenge is that the complex focus of study identified by the first challenge is morphing too fast for any human person to keep up, and thus makes it logically impossible for any individual to imagine the possible impact of AI on society. This demands a collaborative and socially imaginative approach, as even identifying the specifics of the hindrance requires an interdisciplinary perspective. From a computational perspective, one's science fiction-fueled optimism for AI capabilities is quickly dashed by the mundane primitive operations available using modern computers, and it becomes easy to assume that the silicon machines cannot receive the Cartesian ghost comprising the presumed human mind. However, even for the technologically sophisticated engineer, the human imagination lacks the psychological capacity to predict what those simple operations can compute at speeds of a billion billion (10^{18}) operations per second on a billion billion bytes of data (1 exabyte). Cognitive neuroscience determines constraints on human imagination, identifies the limits that AI may try to meet and exceed, and explains how relatively simple neurobiological function supports the range of human reason and experience. Neuroscientifically, even though human cognition appears biologically limited by the almost 100 billion neurons in the human brain, the theoretically possible number of interconnections between those neurons exceeds the number of atoms in the known universe. The constrained synaptic connections between neurons give rise to the

fundamental precursors of the human mind and self, and although human evolution took over a billion years to find a neuroanatomical platform to support human imagination and existence, nothing precludes constructing a similar platform with similar functionality in silicon. Although many important aspects of human existence cannot be reduced to embodied neurobiological activity, they generally also depend upon a person's social, historical, linguistic, and cultural context. AI isolated from those contexts would certainly lack those emergent, human-like aspects, but then so would a completely isolated human. Philosophically, it would be difficult to develop robust theories of cognitive function that machines cannot perform without inadvertently creating an absurd argument that the human brain cannot perform that function either, and additionally, any precise characterization of mental process not easily ascribed to brain function immediately becomes an open research question in the rapidly advancing neuroscience and AI research programs. Theologically, although hard to imagine how AI might develop the capacity for moral reflection, the human ability to form moral, civic, political, and religious systems using a brain comprised of complex relationships between relatively simple components appears good evidence for AI's ability to do likewise.

The aim of the present article is to propose an initial framework for engaging moral theologians in the multifaceted, integrative discourse on moral AI. Progress is made toward the second challenge by identifying four perspectives on human-AI relations that distinguish previously imagined scenarios and suggest directions for fruitful collaboration. Even though outcomes for moral AI remain unknown, some directions are more likely constructive than others. Towards these ends, a theological anthropology for moral AI is proposed along with a systems framework for organizing its collaborative

construction. The first challenge is addressed by identifying specific areas where developers of AI technology and moral theologians can focus collaborative efforts and by supporting those areas philosophically from the perspective of humanities scholars, natural scientists, and social scientists. Some of that discussion occurs in the context of machine ethics, which is examined next, but for a deeper theological foundation, three perspectives on theological anthropology are developed to examine the subjectivity, soul, and self of AI.

Moral Theology for Machine Ethics

Machine ethics (or computational ethics) examines the development of ethical reasoning and behavior in computers. It generally contrasts with ethics of technology (or technology ethics), which examines the moral issues arising from the development of technology, including AI, and its relationship to and impact on human values and flourishing. This article focuses primarily on machine ethics and the construction of moral AI, though given the potential consequences of such development, reasons for doing so are also examined.

In one of the earliest papers on machine ethics, the philosopher James Moor distinguishes between (i) ethical impact of AI, such as removing humans from danger; (ii) implicit ethics, such as pharmacy safety software designed to prevent harm from adverse pharmaceutical interactions; (iii) explicit ethics, such as more recently, Arkin's ethical governor for military Rules of Engagement; and (iv) full moral agency, which

does not yet exist.[5] The philosopher Colin Allen and others argue that machine ethics must be made explicit (and not just remain implicit) because of the unpredictable new ways AI technology can initiate decision making.[6] Most efforts in machine ethics (and the present article) generally focus on making ethics explicit with an eye toward some level of moral agency.

Although machine ethics began as a somewhat speculative area of applied ethics, advances in AI technology have facilitated its interdisciplinary expansion within the discipline of computer science and the field of AI.[7] In AI's historical development alongside cognitive science, one can distinguish research programs primarily focused on using AI to understand human cognition and those focused on building intelligent machines. Similarly, as machine ethics expands, researchers begin to focus on using computer simulations to better understand ethical theories or using ethical insights to build fair, accountable, and transparent computational systems.[8] These two foci of machine ethics also draw upon moral psychology, which studies how and why people

[5] James H. Moor, "What Is Computer Ethics?," *Metaphilosophy* 16, no. 4 (1985): 266–75, https://doi.org/10.1111/j.1467-9973.1985.tb00173.x; Ronald C Arkin, *Governing Lethal Behavior in Autonomous Robots* (Boca Raton: CRC Press, 2009).
[6] Colin Allen, Wendell Wallach, and Iva Smit, "Why Machine Ethics?," *IEEE Intelligent Systems*, no. July/August (2006): 12–17.
[7] Michael Anderson and Susan Leigh Anderson, *Machine Ethics* (Cambridge University Press, 2011); Stuart J Russell and Peter Norvig, *Artificial Intelligence : A Modern Approach* (Upper Saddle River, NJ: Prentice Hall, 2010).
[8] Don Howard and Ioan Muntean, "A Minimalist Model of the Artificial Autonomous Moral Agent (AAMA)" (2016 AAAI Spring Symposium on Ethical and Moral Considerations in Non-Human Agents, Stanford University: AAAI Publications, 2016); Andrew D. Selbst et al., "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, https://doi.org/10.1145/3287560.3287598.

develop, behave, and reason morally.[9] As computational models become more sophisticated, those models may provide additional tools for psychologists and neuroscientists studying human morality. At the intersection of the engineering, humanist, and scientific approaches is creating models of morality, which could then be tailored either for ethical machines or a better understanding of human morality.

The ethicist Susan Anderson argues well that a good initial step toward artificial autonomous moral agents, i.e. AI with full moral agency, is to work toward AI that would advise humans on ethical dimensions of decision making.[10] Although acknowledging disagreements between potential ethical frameworks, she claims sufficient agreement exists on specific ethical decisions to begin developing such an ethical advisor and proposes Ross's *prima facie* duties as a sufficient initial framework. In arguing for the inadequacy of a single absolute duty theory, such as Kant's categorical imperative or Isaac Asimov's three laws of robotics, Anderson identifies the need for and lack of a comparable decision procedure to resolve conflicts between conflicting data.

In humans, the resolution of conflicting ethical demands depends upon practical wisdom (*phronesis*). Phronesis may play a particularly pivotal role in machine ethics and developing moral AI. A general assumption among computer scientists is that ethics is harder than other AI tasks, if not impossible to implement; but moral psychologists find that children roughly ages 8-10 are capable of moral reasoning.[11] The challenge for most

---

[9] Darcia Narváez and Daniel K. Lapsley, eds., *Personality, Identity, and Character : Explorations in Moral Psychology* (Cambridge: Cambridge University Press, 2009).
[10] Susan Leigh Anderson, "Machine Metaethics," in *Machine Ethics*, ed. Michael Anderson and Susan Leigh Anderson (Cambridge University Press, 2011), 21–27.
[11] 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.

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 unforeseen consequences and moral hazards. 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, which have all shown considerable progress.[12] However, advances in autonomous moral agency would require both a foundational system to make moral decisions while resolving moral conflicts *and* an integrated system with the capacity to learn practical wisdom based upon its experience. Currently, AI researchers have the skills to build a foundational system, and philosophers, psychologists, and theologians have insight into human phronesis, but they each generally lack the level of expertise required to make significant direct contribution to research and scholarship of the other. AI researchers could build an AI system for moral reasoning but would not yet know what the system would need 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. Thus, progress is stymied due to the mutually dependent "deadlocked" needs.

As an initial foray into the impasse, I describe an AI system that could plausibly be constructed, with effort comparable to other major AI initiatives, and has the apparent capacity to resolve moral conflicts, by explicitly modeling causal actions and the resolution processes. Constructing such a system would significantly improve the impact

[12] Alison Gopnik, "An AI That Knows the World Like Children Do," *Scientific American*, June 1, 2017, https://doi.org/10.1038/scientificamerican0617-60; Matthew Hutson, "How Researchers Are Teaching AI to Learn like a Child," *Science Magazine*, May 24, 2018, https://doi.org/10.1126/science.aau2576.

of AI in society, enable sophisticated modeling of human morality, and lead to new insights into ethics and moral behavior. More feasibly, the proposed modeling system 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.

Because moral theologians frequently engage in dialogue across humanities and natural and social sciences, moral theology can help provide a broad integrative framework. In addition, moral theology incorporates at least three specific areas of productive expertise. First, constructing machine ethics 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 behavior requires reasoning about moral normativity outside a particular person's context, which moral theologians can expound. Although differences among ethical theories, schools of thought, and religious traditions are legion, I agree with Susan Anderson that enough consensus on ethical thought exists to guide construction of moral AI.[13] Although some engineers, psychologists, and philosophers might have concerns that moral theologians would impose a particular moral theory on AI, moral theology is instead needed to define a normative process appropriate for moral AI that is not an idiosyncratic, culture-specific, and incompletely considered one.

Second, Christian moral theologians typically know well the normative historical and philosophical theories that would have outsized influence on the development of AI

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

outside of Asia and especially in the US and Europe. Because of the long, intertwined development of Christian theology and Western intellectual thought, and its subsequent secularization, many AI developers 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, much less to consider them globally or to fully appreciate the implications for AI ethical autonomy. In the present article, I initiate the development of three theological anthropologies for AI by adapting established human theological anthropologies already beneficial to moral theology. The theological anthropologies bridge moral theology with the specifications needed to construct moral AI.

Third, given the wide spectrum of hopes and fears for AI and the technological, social, and psychological factors that limit imagining a plausible technological and social trajectory for AI, a preliminary step toward integrative discourse is to investigate the perspectives on human nature and the social impact of AI that may influence one's hopes and fears. Moral theologians have the expertise to examine and influence the related public discourse and technological trajectory, and an initial analysis is discussed next.

Human Nature and Technology

Using moral theology to examine moral AI depends upon theological anthropology and one's perspective on the relationship between technology and society. Whether one has optimistic or pessimistic perspectives on human nature and on the relationship between AI and society colors one's hopes and fears about AI and is worth investigating. After considering those alternatives, I argue moral theologians can best influence the development of AI by taking optimistic perspectives on both human nature

and the impact of AI on society, which would produce a positive social influence on AI development.

Examinations of possible AI-human relations are influenced by responses to two questions about human nature and AI's increasing participation in society, which are considered here:

- Does one have an optimistic or pessimistic perspective on human nature?
- Does one have an optimistic or pessimistic perspective on the social impact of AI?

For concreteness, one can consider Augustine to have a pessimistic view on human nature, where one lacks the natural capacity to choose the Good; Irenaeus to have an optimistic view on human nature, where one has the capacity to choose Good but it requires development; the computer scientist Ray Kurzweil to have an optimistic view on what AI would accomplish with advanced research contributions and increased human wealth; and philosopher Nick Bostrom to have a pessimistic view on powerful AI as fundamental existential risk.[14]

The two responses to the two questions lead to four types of scenarios on the future of AI in human society. (i) A pessimistic perspective on both human nature and AI potential is well characterized by Isaac Asimov's classic science fiction in which tragically flawed humans amplify those flaws through misguided constructions of AI (and its ethics) with unforeseen and tragic effect. (ii) A pessimistic view on human nature and optimistic view on AI influence includes transhumanists and others who expect AI to

[14] Ray Kurzweil, *The Singularity Is near: When Humans Transcend Biology* (New York: Viking, 2005); Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford: Oxford University Press, 2014).

help humans overcome our limitations.[15] (iii) Conversely, an optimistic perspective on human nature and pessimistic view on AI includes ethicists and others arguing the need for AI safety where skilled people construct safeguards to prevent AI from causing harm.[16] (iv) The optimistic perspective on both human nature and AI includes machine ethicists who argue for embedding moral decision making in AI and who presume human-derived morality and its AI implementation suffice for autonomous AI.

Under various scenarios, each position has plausible claim to superiority, but I argue that currently in most theological contexts, the optimistic perspective on both human nature and AI is most effective and valuable to public discourse. It is important for thought leaders in business and science to raise public awareness of dangers and social consequences possible with AI technology development if left unchecked, but theologians arguing for human natural inadequacy or dangers of technology are unlikely to make constructive cultural contribution in a contemporary context. Theologians can readily identify the secular hubris and historical improbability of AI and other technology as salvation, but without participation in constructive alternatives, there is little justification for those warnings to be headed. AI safety may appeal in some contexts, such as control of autonomous weapons, but AI safety appears to set humans up for failure with ever increasing AI capabilities overtaking human-created safeguards as well as presumes those safeguards will only be used for the common good. AI safety also overly constrains beneficial applications due to risk; prevents development of appropriate

[15] Ted Peters, "Theologians Testing Transhumanism," *Theology and Science* 13, no. 2 (2015): 130–49, https://doi.org/10.1080/14746700.2015.1023524; Jeanine Thweatt-Bates, *Cyborg Selves: A Theological Anthropology of the Posthuman* (New York: Routledge, 2012).

[16] Dario Amodei et al., "Concrete Problems in AI Safety," *ArXiv:1606.06565 [Cs]*, July 25, 2016, http://arxiv.org/abs/1606.06565.

AI responsibility; and would lead to human oppression of AI, should AI gain attributes otherwise worthy of dignity. Alternatively, an optimistic view of humanity and AI technology supports a mutually beneficial relationship, situates emerging technology within a tradition of moral and spiritual formation, and enables historical access and moral guidance to those developing socially transformative technology.[17] A theological perspective that emerging AI should be treated with dignity is most likely to lead to a society where that stance is appropriate, and in addition, cultivates beneficial virtues for humans in technology-permeated societies.[18]

Although there is an additional opposing case to be made for moral theologians to balance both optimistic and pessimistic perspectives, developing moral AI cannot be an individual exercise. A balanced approach would result in continued deployment of AI technology without sufficient ethical development. In the current social context there exist sufficient, naturally balancing desires and impediments for AI technologists attempting to incorporate ethical insights and enough risks and incentives for social incorporation of AI that moral theologians can enthusiastically embrace the development of moral AI without immediate concern for human nature rejecting the Good or society prematurely embracing a dangerous technology. Undoubtedly both will occur, but the only way for moral theologians and ethicists to affect the resulting collective moral hazards is to participate in the process sufficiently to recognize, identify, and combat them in their particular, complex, and rapidly developing contexts. The contemporary

[17] As clarification, an optimistic perspective on human nature in moral action with respect to AI does not necessarily preclude a pessimistic, Augustinian anthropology with respect to Christian salvation.
[18] Vallor, *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting*.

social context for AI adoption has sufficient conflicting forces that moral theologians must fully engage the development of AI morality in order to create a stable place of balance within society for both fear-based and hope-driven efforts. The pace of developing AI technology is growing too fast for those knowledgeable about ethical frameworks to withhold that knowledge out of concern it would accelerate AI development into moral hazards. AI development is already immersed in moral hazards, and it now needs illumination of those currently existing ones.

AI Theological Anthropology

One way to develop an optimistic AI anthropology would be to identify how AI can have the capacity to know and choose a Good and to resolve conflicts among those internal goods to bring about change. The construct of a "good" relates the goal-directed activity common to AI with the philosophical study of moral goods, the normative aspects of moral theology, and its dependence upon social contexts. The goods for AI can be problem-specific, defined for the AI as a whole, or a moral good, such as Justice prescribed as a *prima facie* duty. Relating those levels of goods and reconciling conflicts between them engages ethical theory and technical development, and constructing AI that learns across contexts requires both general moral constructs and something like phronesis to apply them.

Toward that end two anthropologies are first suggested for AI, drawing upon existing human theological anthropologies with philosophical grounding in Continental or Thomistic philosophy. Each approach identifies issues particularly relevant to current and near future AI development and engages humanities and natural sciences to initiate an AI anthropology with a focus on subjectivity and objectivity, respectively. The first

anthropology draws upon Continental philosophy and extends Heideggerian AI with insights from Xavier Zubiri and Ignacio Ellacuria to examine subjective awareness of AI situated in an historical context. The second anthropology reinterprets Thomistic understanding of soul, as form of the body, and its vegetative, sensitive, and rational powers using natural sciences and systems theory to examine how moral AI can conceptualize its world. Then after identifying some strengths and limitations, a third anthropology is developed using a pragmatist approach to the social sciences to integrate aspects of both initial anthropologies. Constructing subjective- and objective-focused anthropologies clarifies the presumptions of moral AI from the relatively nonoverlapping perspectives of the humanities or science/engineering and identifies the foundation needed for a synthesized anthropology based upon the self's interpretive experience.

Phenomenological Awareness

From Kant's "turn to the subject" to Edmund Husserl's phenomenology and Maurice Blondel's subjective science, one can distill the roots of Continental philosophy's grounding in a rational thinker to focus on AI's potential subjectivity. Heidegger partially psychologized Husserl's transcendental phenomenology with his focus on human (personal) existence, its precursors (*Dasein*) and tools ready-to-hand (*Zuhanden*). Hubert Dreyfus critiques AI from an Heideggerian perspective as never able to grasp reality because symbol processing and representations lack the precursors to personhood.[19] Although Dreyfus's critiques of AI and its assumptions are often

[19] Hubert 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, Mass.: MIT Press, 1992); H L Dreyfus, "Why Heideggerian AI Failed and How Fixing It Would Require Making It More Heideggerian," *Philosophical Psychology* 20, no. 2 (2007): 247–68.

warranted, the philosophical presumptions of subjectivity do not guide engineers trying to construct something like subjectivity in machines. The Continental approach to AI morality nevertheless remains significant because of its intertwined development with moral philosophy and sophisticated efforts to incorporate scientific findings into an ultimately subjective realm, which can be an important corrective to naive and unsupported objective presuppositions about scientific knowledge. Philosophically sophisticated interpretations into human subjectivity (e.g., qualia) might indirectly yield a framework sufficient for AI subjectivity, and the studies that characterize human subjectivity as a phenomenological process have informed AI research. Although many AI researchers generally dismissed or rejected Dreyfus' critiques, some have incorporated aspects of Maurice Merleau-Ponty identification of embodiment as necessary for phenomenological experience through the work of Francisco Varela and others.[20] The present anthropology attempts to extend a phenomenological approach to AI by incorporating additional aspects of Heideggerian thought as developed by Xavier Zubiri and Ignacio Ellacuria.[21]

The work of Spanish philosopher Xavier Zubiri is relevant for AI as he incorporates scientific findings into a Continental framework with a phenomenological

[20] Francisco J Varela, Evan Thompson, and Eleanor Rosch, *The Embodied Mind : Cognitive Science and Human Experience* (Cambridge, Mass.: MIT Press, 1991); Rodney A. Brooks et al., "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, USA: American Association for Artificial Intelligence, 1998), 961–968.

[21] Efforts to directly relate Heideggerian AI to theological anthropology include Porter, "A Theologian Looks at AI."

focus.[22] As a student of Husserl and Heidegger, Zubiri engaged personally and intellectually in early twentieth-century scientific activity to develop his construct of sentient intelligence, a way of apprehending reality, which he claimed is entirely compatible with modern science. After characterizing intellection as the active verb for using the intellect, Zubiri argued that an artificial dichotomization splitting sensation from intellection resulted in a logification of intelligence and entification of reality, i.e., reducing intelligence to *logos* and reality to entities. Zubiri critiques the "reductive idealism" resulting from logification and argues sensing is an aspect of intellection, not a preconceptual precursor to it—a point that resonates with claims for embodied cognition.[23] In addition, Zubiri clarifies reality is a process, not a collection of things, in arguing broadly against physical reductionism. Simply, the Cartesian mind-body split creates a mind too Platonic and a body presumed too much of an entity (*ente*)—both of which distort reality. Zubiri's solution reclaimed reality as primary and identified sensing and intellection as two dimensions of sentient intelligence, which I argue is also a plausible foundation for artificial intelligence or at least an important corrective to existing assumptions. Classic approaches to AI, or "Good Old-Fashioned AI" (GOFAI), depend upon symbol manipulation and were generally understood through a lens of logical positivism, which strictly separated the sensing of objects in the world from the

[22] Xavier Zubiri, *Sentient Intelligence*, trans. Thomas Fowler (Washington, DC: The Xavier Zubiri Foundation of North America, 1999). See Robert Lassalle-Klein, *Blood and Ink: Ignacio Ellacuria, Jon Sobrino, and the Jesuit Martyrs of the University of Central America* (Maryknoll, New York: Orbis Books, 2014)., chap 5 for a deeper examination of Zubiri's arguments as summarized here.
[23] Varela, Thompson, and Rosch, *The Embodied Mind*.

symbols representing those objects. Because these symbols lack real-world grounding, they exist similarly to how medieval scholars considered the realm of universals.[24]

For our purposes, Zubiri's arguments clarify there are no object specifications intrinsic to reality that one must map to putative universals in one's mind. Instead, one's apprehension of reality is what defines the "objects" as objects (and that in the context of how one might use those objects). Zubiri then identified three aspects of intellection: (i) a primordial apprehension "in itself"; (ii) what is real with respect to other real things; and (iii) apprehension vis-à-vis already apprehended realities. Although the first aspect may eventually become relevant for AI philosophy, the other two aspects more directly impact AI anthropology given the current stage of AI development. The second aspect of Zubiri's intellection characterizes that the act of conceptualization occurs within apprehension rather than as a separate logifying process, and the third aspect explains how those apprehensive acts become seen (in and by humans) as a worldly reality, specifically as an act of reason. The conceptualization process for AI is important for morality because the conceptualizing must also carry moral weight, which lacks grounding when the concepts are divorced from reality.

As described further in the objective-focused anthropology below, Zubiri's second aspect reinforces AI development away from apprehending symbols (and sub-symbolic constructs) as universals and guides the synthesized anthropology toward identifying the apprehension itself as fundamental to reality. Although the third aspect does not appear strictly necessary for current development of AI morality, it can clarify

[24] Although newer "deep learning" approaches to AI lack explicit symbols, they are often interpreted similarly, with symbolic representation distributed across the neural net.

why a presumed Cartesian ghost is not necessary for humans or machines. The apparent reality of mind is a consequence of one's apprehension of reality, not a precursor to it.

In addition to Zubiri's direct contribution to AI as sentient intelligence, moral AI can also benefit from Zubiri's major theological interpreter, the Spanish-Salvadoran philosopher Ignacio Ellacuria.[25] Ellacuria acknowledges sentient intelligence's role in apprehending reality and also situates reality as including both the natural realm and historical reality, which incorporates society's trajectory into the reality which, and in which, we apprehend.[26] When Dreyfus criticized early approaches to AI, one issue was the assumption that reality consists of substances, and that assumption resulted in AI striving to learn properties of those substances (e.g., the frame problem). Zubiri (and others since Kant) identify the role of the mind in defining what had previously been considered as substances, and Ellacuria follows through with Zubiri's de-logification by situating the subject within history.

Part of Ellacuria's motivation was that the substantial categories created by Western European philosophy proved inadequate to address the brute reality of his Latin American situation. As AI has an even more radically different physical, historical, and embodied context, Ellacuria's multicultural perspective can also expand Continental philosophy's contribution to AI. Ellacuria argued that sentient intelligence is historical—occurring and defined within the context of human history—and a similar argument holds for artificial intelligence.

---

[25] Kevin F. Burke and Robert Anthony Lassalle-Klein, *Love That Produces Hope: The Thought of Ignacio Ellacuría* (Collegeville, Minn.: Liturgical Press, 2006).
[26] Lassalle-Klein, *Blood and Ink*, 221.

Ellacuria identifies the political implications of separating sensibility from intellection and characterizes three dimensions of facing real things as real: (i) becoming aware of what is at stake in reality; (ii) an ethical demand to "pick up" or assume responsibility for reality; and (iii) a praxis-related demand to change or take charge of reality. Ellacuria's first dimension identifies the movement for becoming aware of one's distinction from reality and his second dimension clarifies the ethical stake in how one apprehends reality. Relevant for constructing moral AI, Ellacuria identifies that how one apprehends reality occurs in a social context, he calls historical reality, and that the 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 clarifying the distortion between sensing and intellection, Zubiri and Ellacuria illuminate the delusion that one senses an object and then thinks about the moral implications of one's actions with respect to that object. Instead one brings an ethical imperative of acting morally to every apprehension one makes of reality, and that infuses the conceptualizations one generates in constructing one's historical world. Conversely, until AI can assume ethical responsibility for its reality, then humans will not recognize its apprehension as intelligent. This constrains the forthcoming synthesized anthropology to tightly integrate conceptualization and moral considerations as an aspect of apprehension.

If one incorporates Ellacuria's insight, then AI must incorporate the ethical implications of its actions with respect to what it apprehends as part of the apprehension (and conceptualization) process itself. This aligns with current practice in AI data ethics, which has realized that one cannot remove bias from data modeling, and thus AI model

descriptions and result explanations must be explicit about those biases.[27] Morality is thus not something added to AI, but is already intrinsic to it—just currently, poorly understood and implemented. AI's conceptualizations cannot exist as symbols in a universal realm if they need to carry moral weight, i.e., identify a good. In addition, Ellacuria identifies that moral AI should recognize that its apprehension occurs in a historical context. These points will be revisited in the synthesized anthropology and form the basis for suggesting two dimensions of moral AI architecture, modeling both its apprehension and the AI itself as (historical) agent.

Natural Existence

An alternative anthropology for AI draws upon natural sciences and a systemic perspective that emphasizes objectivity. Classically, studies of theological anthropology are often grounded in characterizations of the human soul, and a Thomistic account of the soul as form of the body is illustrative and particularly helpful. Thomistic anthropology has had extensive impact on theology and the history of Western (European) intellectual thought, including moral theology, moral philosophy, and virtue ethics. In addition, a Thomistic anthropology provides at least four aspects of what is needed to develop an AI anthropology integrated with a systems approach to science.

First, a monistic soul overcomes misleading or incomplete reductionist and dualist assumptions of physicality (physical reductionism), the mind (*res cogitans*), and

[27] Osonde Osoba and William IV Welser, "An Intelligence in Our Image: The Risks of Bias and Errors in Artificial Intelligence," *Rand Corporation*, 2017; David Danks and Alex John London, "Algorithmic Bias in Autonomous Systems," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence* (Twenty-Sixth International Joint Conference on Artificial Intelligence, Melbourne, Australia: International Joint Conferences on Artificial Intelligence Organization, 2017), 4691–97, https://doi.org/10.24963/ijcai.2017/654.

spirituality (e.g., Neoplatonic), which would be difficult for AI development to overcome. Although philosophically sophisticated adherents of both reductionist and dualist positions generally nuance their arguments to overcome powerful scientific and philosophical critiques of either extreme, a growing number of scholars seriously engaging both science and theology argue for a mediating position for human nature, such as nonreductive physicalism or emergence.[28] These mediating positions acknowledge both physical reality and the need for additional explanatory or causal types of reality that are neither separate from nor reducible to purely physical existence. These nuanced mediating positions are needed both for the valid appropriation of human cognition for AI and for theories guiding AI development. Thomistic monism directly addresses Neoplatonic (spiritual) dualism, and provides a historically relevant precursor framework to reconsider Cartesian (mental) dualism. Form and formal cause are effective in countering physical reductionism and provide constructs to reexamine what was historically interpreted as substance (what Zubiri called entification). The synthesized anthropology aims for the emergence of AI morality in a way that aligns with human morality, though does not depend upon human morality actually being emergent.

Second, scientist-theologian Arthur Peacocke and others have argued for interpreting information within the ancient category of form.[29] This supports using

---

[28] Nancey Murphy, "Physicalism Without Reductionism: Toward a Scientifically, Philosophically, and Theologically Sound Portrait of Human Nature," *Zygon* 34, no. 4 (1999): 551–71, https://doi.org/10.1111/0591-2385.00236; Philip Clayton, *Mind and Emergence : From Quantum to Consciousness* (New York: Oxford University Press, 2004); Philip Clayton and Paul Davies, eds., *The Re-Emergence of Emergence : The Emergentist Hypothesis from Science to Religion* (Oxford: Oxford University Press, 2006).

[29] Arthur Robert Peacocke, *Theology for a Scientific Age: Being and Becoming-- Natural, Divine, and Human* (Minneapolis: Fortress Press, 1993); Niels Henrik Gregersen, "God,

Aquinas' understanding of soul as form of the body to examine the form of AI and thus characterize AI soul in terms of its information processing. Although challenging to imagine the connection between a metaphysical interpretation of form as soul and Claude Shannon's definitive formalization of information in terms of communication capacity (i.e., bits), form and information have closer connection in biology, which Peacocke considers. The reconciliation of form and information is needed to substantiate both theological investigation of computational processes (AI soul) and AI's processing of its perceived world with moral ends (AI phenomenological apprehension). As another theological reconciliation, Pannenberg's multifaceted understanding of information as a complex generative process with Trinitarian (and Neoplatonic) roots aligns with Terrence Deacon's reformulation of information theory and what he identifies as Darwinian information in his emergent selection dynamics.[30] A careful consideration of monistic soul as form of the body characterized through the lens of contemporary science can provide a more powerful understanding of information and the unity of an AI person.

Third, Thomistic vegetative, sensitive, and rational powers of the human soul map reasonably well to both AI cognitive architecture and the human sciences. In particular, the early AI researcher Allen Newell distinguished between cognitive and rational levels for AI architecture.[31] Subsequent research has identified key aspects of the cognitive

---

Information, and Complexity: From Descriptive to Explorative Metaphysics," *Theology & Science* 11, no. 4 (2013): 394–423, https://doi.org/10.1080/14746700.2013.866475; Mark Graves, *Mind, Brain, and the Elusive Soul : Human Systems of Cognitive Science and Religion* (Aldershot, Hants, England ; Burlington, VT: Ashgate, 2008), chap. 2.

[30] Mark Graves, "Places of Information Generation: Bridging Pannenberg's Logos and Deacon's Emerging Semiosis," *Theology and Science* 14, no. 3 (2016): 305–24, https://doi.org/10.1080/14746700.2016.1191880.

[31] Allen Newell, *Unified Theories of Cognition* (Cambridge, Mass: Harvard University Press, 1990).

level, such as memory, learning, and attention that loosely overlap with Thomistic sensitive powers.[32] For Aquinas, the rational powers of intellect and will are required to complete the activity of lower powers in humans. 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. This occurs in the sensitive powers for Aquinas, but goes beyond what is generally considered an aspect of a cognitive architecture for AI. The intellect continues the categorization and conceptualization by purifying the concrete phantasm to its intelligible species, or a concept, which can guide understanding of conceptualization in AI. For Aquinas, the ideogenesis process continues to produce a universal from the intelligible species. The universal defines the natural ends and is required to identify what is good, thus ideogenesis is significant for AI morality.

Fourth, Aquinas's ideogenesis process of ascertaining the substantial form, sensible species, and universal essence from phenomena identifies both the problematic presumption of classic AI's symbolic representation (e.g., assuming universal referents) and the importance of characterizing the conceptualization process of AI. Aspects of AI's historical roots in mathematics provides some justification for universals, such as numbers and Platonic solids, and universal quantification in logic simplifies some reasoning processes. However, the implicit assumption of universals reinforces the logification and entification tendencies that Zubiri identifies and obscures the social (and developmental) processes by which humans do learn to conceptualize and reason about

[32] 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, https://doi.org/10.1609/aimag.v38i4.2744.

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.

These four aspects of Thomistic anthropology identify potential areas of collaboration for theologians and AI researchers to better understand their complementary goals and presuppositions. Together they suggest Thomistic powers of the soul can guide development of AI architecture. Without requiring the body to be human, the vegetative, sensitive, and rational powers characterize a possible form for AI. If one predefined a collection of universals, perhaps for a video game or other online world, then an AI avatar architected with Thomistic powers of the soul would most likely behave intelligently. The challenge in the real world is that those universals are not predefined, so Thomistic rational powers must not depend upon universals (analogously to Zubiri's delogification of intellection). Although universals are not explicitly required until late in ideogenesis, that end influences the entire ideogenesis process, in part due to Aquinas's presumption of natural law. Reconciling a Thomistic anthropology with modern science could lead in a variety of directions, but since the goal here is to synthesize with the subject-focused anthropology, 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.[33] Systems theory can incorporate the architectural requirements for Thomistic powers of the soul and social sciences sufficient to reconcile the role of universals.

[33] Newell also acknowledges the Social band in Newell, *Unified Theories of Cognition*.

Systems theory, beginning in 1940s with the seminal work of Ludwig von Bertalanffy, attempted to develop a general theory to organize natural and social phenomena based upon patterns and principles common across a range of disciplines.[34] 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, and that unifying organization suffices for characterizing an integrated perspective on natural and social sciences, even though specialized theories may prove more effective in distinct specific areas.

In general system 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.[35] Arthur Peacocke argues similarly in dialogue between theology and science and organizes his part-whole hierarchies of nature into four similar levels of focus based upon A. A. Abrahamsen distinctions between the physical world, living organisms, the behavior of living organisms, and human culture.[36] The contemporary philosopher of science and religion Philip Clayton suggests an additional emergent level of spiritual or transcendent activity, which emerges from mental (and cultural) activity, which in a

---

[34] Ludwig von Bertalanffy, *General System Theory: Foundations, Development, Applications* (New York: G. Braziller, 1969).
[35] Ludwig von Bertalanffy, *Perspectives on General System Theory : Scientific-Philosophical Studies* (New York: G. Braziller, 1975), 5–8, 30–32.
[36] W Bechtel and A Abrahamsen, *Connectionism and the Mind* (Oxford and Cambridge, Mass.: Blackwell, 1991), 256–59; Peacocke, *Theology for a Scientific Age: Being and Becoming-- Natural, Divine, and Human*, 215; Arthur Robert Peacocke, *God and the New Biology* (London: Dent, 1986).

systems model would capture activity at a fifth spiritual or transcendent level.[37] 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 illuminates processes conflated within Thomistic rational powers. Historical and linguistic activity occurs at the social/cultural level, and the resulting presumed universals define the transcendent level. Systems theory clarifies the levels within which emergence may occur, and systems can characterize the information processing Thomistically ascribed to form. 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 historical dualist realms of *elan vital* or *res cogitans* are now well characterized by systems theory as biological and psychological levels, respectively. As analogues to universals, the theologian David Tracy defines the symbols that reveal permanent possibilities of meaning or truth as "classics," and Terrence Deacon's emergent dynamics can be used to describe how the transcendent-level processes relate to classical universals, such as transcendentals of Truth, Beauty, and the Good.[38]

In parallel to the claim that emergent systems suffice for what Peacocke calls a revised natural theology, the present article shows how such a systems theory can extend the apprehension and conceptualization of AI to include morality. Systems theory

---

[37] Clayton, *Mind and Emergence*; Mark Graves, "The Emergence of Transcendental Norms in Human Systems," *Zygon* 44, no. 3 (2009): 501–32.

[38] David Tracy, *The Analogical Imagination : Christian Theology and the Culture of Pluralism* (New York: Crossroad, 1981), 8; Terrence W Deacon, "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."

clarifies ideogenesis by separating universals to the transcendent realm, conceptualization dependent upon culture (and language) to the social-cultural level, and the categorization of phantasms to the psychological level (shared significantly but not exhaustively with at least primates and some other mammals). For a human anthropology, many nontrivial steps would be required to explain the emergence, nature, and top-down influence of the analogues to universals with respect to social processes. However 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., virtues, categorical imperative, *prima facie* duties, or even Asimov's laws—in a way analogous to universals. Instead of replicating ideogenesis as it results in universals, AI can appropriate human moral norms in terms of transcendent-level systems and conceptualize reality toward those ends. Although one could treat Asimov's laws as universals, resolving conflicts between *prima facie* duties, such as beneficence, nonmalfeasance, and justice requires situating those duties in a social context, which transcendent-level systems would better support. Both anthropologies align with apprehension/perception grounding the conceptualization of reality occurring in a social context guided/directed by the contextualized ethical norms. In order for AI categorization and conceptualization to own its moral stake, sufficient for resolving conflicts between ethical norms or duties, AI must situate its cognitive processes within a social/historical process, and that requires synthesis between the subjective-focused and object-focused anthropology and incorporation of social sciences.

Interpretive Experience

Although each of the two anthropological approaches has its strengths and weaknesses as standalone bases for AI anthropology, they best serve as complementary

lenses to examine AI morality. The subjective-focused anthropology addresses the importance of identifying the locus of personhood and describes how that locus apprehends reality in a social context but only provides a descriptive, not constructive, characterization of how the subject comes to be. The objective-focused anthropology characterizes reality, with emergence a plausible, but not strictly required, explanation for its interrelationships (as perceived by humans and studied by science), but despite explaining social relations, it does not explain sufficiently how categories and conceptualizations are formed individually and scientifically. Two gaps in synthesizing subjective- and objective-focused anthropologies are (i) the nature of the subjective locus, and (ii) how that relates back to the remainder of reality. These two gaps are the focus of the social sciences and pragmatic philosophy, respectively.

Social psychology as founded by George Herbert Mead identifies the locus of personhood, or "self," as a social process created by interactions within a group or society.[39] The individual social self initially appropriates the society's shared values and ideals, then as it emerges, interiorizes the social environment in which it lives, and finally begins transforming society through its relationships. As the 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 between AI engineering and the humanities.

To relate Mead's social self back to the previous anthropologies, some distinctions from personality and social psychology are helpful. The psychologist Dan

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

McAdams identifies three levels or 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 are the stories one constructs to give one's life a sense of unity and purpose. Simplistically, dispositional traits may depend upon genetic predispositions, early childhood development, and other factors forming a core to one's self that would align with individual variations historically attributed to one's soul. Conversely, characteristic adaptations are more circumstantial, and subjective, depending upon one's social, historical, and cultural context as it influences how one apprehends reality and responds to the reality one finds at hand. In the context of one's unfolding life in relation to others, one forms an identity that gives meaning and coherence to one's behavior over time. The story one tells about oneself is affected by one's dispositions and circumstances and by one's goals and aspirations. One's story may align well with reality (for those humble and self-actualized) or diverge radically (in delusion), but it creates a unity the other anthropologies presumed ontologically prior to the self. One's subjective awareness is not an abstract locus (which Zubiri warned had been logoified) or a substantial form, and it does not reduce to one's natural existence and social circumstances (as it is a story one creates about one's self). The realization the "self" develops over time (in a historical-social context) identifies the limitations of considering the essential locus of a person as an "atomic" subject or soul. It remains to be seen whether AI would develop such a self, given similar social and cognitive capacities as humans, or whether it depends upon peculiarities of human memory and other cognitive functions. However, nothing appears to preclude development of at least similar precursors for AI, which can apprehend and conceptualize

reality in a way analogous to humans, and if so constituted could then apprehend itself as social creature, even if its "self", or proto-self, differed significantly from humans.

The second anthropological gap requires explaining how the self relates back to the remainder of reality, which pragmatism addresses. The Jesuit theologian Donald Gelpi extends Mead's construct of social self in a metaphysical direction, incorporating Alfred North Whitehead's metaphysical process of an emerging self into C.S. Peirce's phenomenological metaphysics to develop a metaphysics of experience.[40] As an aspect of Gelpi's experiential metaphysics, he develops a theological anthropology of an autonomous, social, sentient being that experiences the world and develops through decision-making. For Gelpi, the decision-making occurs within an evaluative process that results in taking on of habits or tendencies, which then become the foundation for one's future decision-making. Although Gelpi extends his metaphysical development by incorporating Peirce's synechism, and its continuity across physical and mental dispositions, it suffices here to simply require that the AI system have the ability to learn from its decisions in a way that affects future decision making, which is a general feature of most machine learning systems and an explicit characteristic of reinforcement deep learning. The dispositional nature of Gelpi's emerging self incorporates the teleological requirement of AI development distinct from universals in a way amenable to the development of virtue, which supports development of an AI virtue ethic.[41] In addition, Joseph Bracken, somewhat conversely from Gelpi, borrows from Peirce in Bracken's

[40] Donald L Gelpi, *The Gracing of Human Experience: Rethinking the Relationship between Nature and Grace* (Collegeville, Minn.: Liturgical Press, 2001).
[41] Mark Graves, "Habits, Tendencies, and Habitus: The Embodied Soul's Dispositions of Mind, Body, and Person," ed. Gregory R Peterson et al., *Habits in Mind: Integrating Theology, Philosophy, and the Cognitive Science of Virtue, Emotion, and Character Formation* (Brill, 2017).

refinement of Whitehead's process thought to replace Whitehead's eternal objects with systems theory to characterize the intersubjective nature of the self, which might support well a deontological ethic of care similar to what Susan Anderson's *prima facie* duties might require.[42]

The construct of experience is key to reconciling subjective- and objective-focused anthropologies. As subject, one encounters one's world, and then interprets one's experience into objective categories. Being explicit about the encounter and interpretation enables experience to serve as a pragmatic foundation for the synthesized anthropology. Subjectivity occurs at the nexus of encounters and is defined by those natural and social experiences.[43] The "objective" categories of interpretation are not *a priori* universals, but socially constructed with others in society (and through history). Previously these "others" have always been human (setting aside possible revelatory experiences), and now other precursors to persons are entering into that society. Ellucuria recognizes that the precursor to personhood (*Dasein*) is not an aspect of objective reality but occurs within historical reality—a reality in which AI is currently emerging. Within the objective-focused anthropology, a dualistic mind or soul accessing universals is unnecessary, instead one must incorporate the equivalent of historical reality into monistic experience.

The correspondence between von Bertalanffy's systems theory, as extended by Peacocke and Clayton, and Zubiri's sentient intelligence suggests organizing interpretations as multiple levels of models that AI can use to interpret its reality.

---

[42] Joseph A Bracken, *Subjectivity, Objectivity, & Intersubjectivity : A New Paradigm for Religion and Science* (West Conshohocken, Pa.: Templeton Foundation Press, 2009).
[43] John Edwin Smith, *Experience and God* (New York: Oxford University Press, 1968); Denis Edwards, *Human Experience of God* (New York: Paulist Press, 1983).

Borrowing from human experience, five levels of interpretation would be (a) models of spatial (or virtual) and temporal extent in physical objects; (b) biological processes, including classic vegetative powers of nourishment, growth, and reproduction; (c) sensation and animation typified by most animals; (d) expressiveness and meaning of symbolic language as a tool for conceptualization and communication; and (e) moral and spiritual concerns and capacities. These interpretive levels suggest an organization for moral AI systems and a staged taxonomy of precursor AI systems.

Moral AI Systems

The taxonomy of AI morality has two dimensions. The first dimension of AI morality captures a categorization of five levels of models the AI can maintain and use in deliberation among possible actions. The phenomena modeled in each level logically depend upon the prior levels where higher-level differences require lower-level differences, i.e., the higher levels supervenes on the lower level, yet the higher level has causal relationships not operative at the lower level.

In addition to modeling the world in which AI acts, in order to deliberate, AI must also consider its own actions and their possible effects. The neuroscientific correlates of human self-awareness is an open and active research area, 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.[44] For moral autonomy, AI likely

[44] 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, https://doi.org/10.1111/j.1750-8606.2011.00189.x; Narváez and Lapsley, *Personality, Identity, and Character : Explorations in Moral Psychology*; 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,

requires a platform supporting deliberation as well as internal representations of its self. The subsequent focus in the present article is on AI self modeling because its requirements appear better understood than those of the underlying platform, it may prove necessary to characterize an AI self prior to building the platform requirements for it, and it would align with the current understanding of human subjectivity, whose numerous influencing factors are well-studied and whose underlying platform has proven elusive to investigation.

The five levels of external models and five stages of internal awareness correspond to five interpretive levels of the synthesized anthropology. The interpretive lens clarifies that the proposed five levels of external models refer to AI interpretation of its encounter with its external world, not an objective classification of reality as the natural scientific anthropology emphasizes nor reified phenomena as Zubiri argues against. The interpretive lens also clarifies that the internal awareness is historically situated in the AI's experience and not logified as Zubiri cautions. The stages of internal awareness build upon each other and the corresponding external modeling levels. The five levels of external models and stages of internal awareness are described in turn, before considering their use in resolving moral contradictions and implications for practical wisdom.

Causal Levels for External Modeling

Physical. An awareness of spatial and temporal extent would suggest AI respects boundaries, preserves integrity of systems and objects, and for example as a robot, does

---

https://doi.org/10.1177/0276236618756704; L J Walker, "Moral Personality, Motivation, and Identity," *Handbook of Moral Development*, 2014, 497–519.

not break anything. Classical ethical considerations of this level include recognition of public space and private property. AI in virtual space can also be aware of and account for other AI and human embodiment in that virtual space. Physical-level models were developed among some of the earliest AI systems. Dreyfus critiques their use as context-free symbols and drew upon Heidegger to emphasize the dependence of these models upon how they would be used, which was picked up by later robotics researchers. The modeling framework needs to avoid logifying the models as separate from the sensing process and avoid treating the objects (as modeled) isolated from the AI's apprehension, which also circumvents traps of physical reductionism and dualist presuppositions. C.S. Peirce's pragmatic maxim constrains the models to what conceivable practical effects the models (i.e., one's conception) might have, which aligns with enactive cognition.[45]

Biological. The ability for AI to respond to biological organisms would require modeling their trajectories for growth, nutrition, and reproduction, and for AI to incorporate into its decision-making whether it is assisting or hindering those ends. Significant ethical considerations of the biological level include topics studied in ecological (or environmental) ethics. The biological level models the actions of Thomistic vegetative powers. 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. Ernst Mayr and other philosophers of biology have argued for the importance of distinguishing biological processes from physical objects, and most models of "computation" as a process would require biological-level models. In virtual space, a static web page (url) might only require a

[45] Charles S Peirce, "How to Make Our Ideas Clear," *Popular Science Monthly* 12 (1878): 286–302; Varela, Thompson, and Rosch, *The Embodied Mind*.

physical-level model, but a webservice (e.g., microservice) or even a web form would be better modeled analogously to biological organisms.

Psychological. Modeling and responding to organisms with sensation and action would include the ability to model the other agents explicit and implicit goals and evaluate its decisions that help, hinder, or remain neutral to those goals. Implicit goals could include the avoidance of pain, i.e., threat of possible tissue damage, and sentient organisms' felt response to pain. Ethical considerations of creatures modeled at this level occur in animal ethics. Physical-level models are required to capture animation and autonomy. The sensing precursors may fully develop into sentience, and the level captures Thomistic sensitive powers. 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. Irrespective of building moral AI, the systems model illuminates numerous philosophical pitfalls for AI approaches that attempt to directly connect universals 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.

Social. 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 an awareness of their use, conventions, and affects.[46] 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. Most investigations of human ethics generally consider the personal, social, and civic systems modeled at the social level. Identifying the linguistic boundary between humans and other animals is well studied and has somewhat influenced AI research into language.[47] 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.[48]

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 as emerging in the Axial Age (c. 800-200 BCE).[49] Ethical theories themselves would be modeled at this level and investigations in metaethics and moral

[46] Terrence W Deacon, *The Symbolic Species : The Co-Evolution of Language and the Brain* (New York: W.W. Norton, 1997).
[47] Deacon.
[48] Zellig Harris, *Mathematical Structures of Language* (New York: Interscience, 1968).
[49] Robert Neelly Bellah, *Religion in Human Evolution : From the Paleolithic to the Axial Age* (Cambridge, Mass.: Belknap Press of Harvard University Press, 2011).

theology often take phenomena and social constructions modeled by this level into account. While a care robot evaluating choices involving *prima facie* duties of beneficence and nonmalfeasance might take social-level and lower-level models into account, an AI evaluating whether a deontological or care ethic would be more appropriate would require the moral-spiritual models of this level.[50]

Stages of Internal Awareness

AI morality's second dimension captures AI's capacity to model itself as moral agent and is described as five stages. The first dimension captures models used to interpret the agent's external world, and the second dimension uses those models as a foundation for representing the agent itself. Human self-awareness gradually occurs at a very young age and is well studied yet only partially understood.[51] The second dimension characterizes models of the self necessary for internal awareness, though it is not yet known what else might be required for AI self awareness and identity formation. Instead these models provide a plausible foundation for moral behavior and further exploration.

Spatial-Temporal-Virtual Extent. Moral agency 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

---

[50] 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. However, both AI and social scientists with a capacity for some social relationships and articulated spirituality might create better models than those who lack those capacities.

[51] Philippe Rochat, "Five Levels of Self-Awareness as They Unfold Early in Life," *Consciousness and Cognition*, Self and Action, 12, no. 4 (December 1, 2003): 717–31, https://doi.org/10.1016/S1053-8100(03)00081-3; Dan P. McAdams, "The Psychological Self as Actor, Agent, and Author," *Perspectives on Psychological Science* 8, no. 3 (2013): 272–95; Susan Harter, *The Construction of the Self: Developmental and Sociocultural Foundations* (New York: Guilford Press, 2012).

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, but not the choices necessary for animation. However, 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 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.[52]

[52] Nicola Jones, "How to Stop Data Centres from Gobbling up the World's Electricity," *Nature* 561 (September 12, 2018): 163, https://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, https://doi.org/10.1016/j.patcog.2018.07.023.

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 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. 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.

Sensing-Animate Agent. Moral agency requires AI systems to monitor and model their own actions to determine how their actions affect the goals of other organisms and AI. With an internal awareness comparable to many animals, the AI can sense its environment and act within it. The AI models itself psychologically, as it would other animals, and extends the modeling to account for its sensing and actions. Challenges to imagine the models required for agency include most of those mentioned in this article. The AI agent is not a logoified mind perceiving reified 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 agent. 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. 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 agents might yield an awareness that fighting would minimize potential tissue damage and pain. At this stage, AI lacks the social awareness to, for example from humans, consider one's offspring as particularly vulnerable extensions to one's body and thus worth defending despite severe actual tissue damage. But the precursors to extending 'better' in a socially and eventually ethical direction exist at the sensing-animate agent stage.

Social-Historical Participant. As a social-historical participant, AI's internal awareness supervenes upon its internal awareness of its causal agency and depends upon its base social modeling. 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 models itself as a social being and has a desire for positive feedback in social relations, i.e., pleasure or happiness, then that desire for social participation can provide some norms for ethical behavior. 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.

Moral Agent. An additional level of AI morality would require AI modeling and monitoring its behavior with respect to culturally conditioned norms of putatively universal principles. AI would need 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, the norms to which it aspires, and the further ultimate concerns toward which it strives.

The optimistic-vs-pessimistic perspective on human nature and technology may influence whether AI should or could model ethical theories and norms. If one is optimistic about human AI society cooperatively developing meaningful values and norms, then AI modeling itself as a moral agent is important. But if one is pessimistic about those outcomes or capacities, then the levels of models and stages of development can help with that analysis. Eliminating the stage of moral agency would still enable social participation but preclude the AI viewing itself as participating in and influencing moral and spiritual systems and phenomena. If one were optimistic about human nature, and pessimistic about technology in society, then one might prefer socially participatory AI who lack self monitoring and reference to putatively universal norms, such as universal rights and justice; though this would require optimistic views of human nature, or an unscrupulous human could socially manipulate AI for nefarious ends. Conversely, a pessimistic view of human nature and optimistic view of technology might recommend minimizing human input on the moral and spiritual level of modeling beyond what is required for AI moral agency. Other configurations are possible with omission of certain

social models and awareness possibly analogous to various forms of human abnormal psychology. The development of AI behavioral science incorporating findings from human moral and positive psychology may prove helpful for designing, developing, and configuring such future moral AI.

The proposed modeling framework has implications for philosophical examinations of AI, such as AI personhood; and 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. In particular, as described earlier in the article, addressing moral conflicts is an open problem in machine ethics for which practical wisdom appears required.

Practical Wisdom

As a foundation for ethical decision-making, Aristotle claimed practical wisdom (*phronesis*) included an ability to deliberate well and both general and situation-specific understandings of the good. The ability to deliberate presumes an interior (mental) world where one can simulate and evaluate one's possible actions before acting, which the second dimension of modeling can provide. The models of internal awareness make moral deliberation explicit and affords the possibility of resolving conflicts between general, normative goods.

A moral AI with all five levels of external models and stages of internal awareness has the capacity to consider its actions (as a causal agent) 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 internal awareness must support deliberation and resolution of conflicting ethical demands by the stage of moral agency. Third, the lower-level models must expose an adequate interface for internal awareness sufficient to attend to proximate goods and for the stage of moral agency to interpret moral-spiritual goods in terms of those proximate goods. Fourth, the stages of causal agency and social participation 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 agency to form intentions.

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, how do these connect into moral action? Specifically, how does causal agency in

society bring about obtainable proximate goods in light of general goods and values toward which one strives?

As an intellectual virtue, phronesis depends upon the interpretive models and internal awareness characterized above. Virtues in the Aristotelian tradition are habits mediating between vices and oriented toward some end, and 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 logified universal and entified object were separated, no disposition could be formed. In addition, 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).

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 internal awareness—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, rapidly progressing

development of relevant science and technology, and disparate perspectives on human nature and the impact of technology on society. These factors limit the imaginative and incremental construction of AI capable of moral reason, deliberation, and action, especially in complex realistic situations with apparent conflicts between moral goods. AI theological anthropology can guide theological efforts to influence the construction of moral AI, and an initial step is to work out the image of AI in comparison to humans.

One can examine the image from subjective, objective, and experiential perspectives building upon Continental, Thomistic, and pragmatic philosophy to characterize that image as subject, soul, and self, respectively. Continental philosophy supports reasoning about subjective phenomenological experience and engages AI through embodied and enactive cognition and the Heideggerian challenges of Dreyfus. Zubiri identifies the logified intelligence and entified reality that affects human apprehension, and the similarly adjusted AI apprehension addresses hindrances to AI research. Ellacuria's historical reality and its demand of a moral stance situates an AI subject within human history and social and linguistic context. Conversely, Thomistic philosophy presumes an objective reality with universals identifying unambiguous Goods for a teleological oriented soul. Revising with contemporary science and systems theory structures the powers of an AI soul such that putative universals can be socially and historically contextualized for the conceptualization needed by AI moral reasoning. Synthesizing the subjective and objective-focused anthropology into a pragmatic anthropology focuses AI on the interpretation of experience and can draw upon social sciences to define an AI self.

Moral AI interprets its external world, through five levels of models, and progresses through stages of internal awareness, which build upon similarly organized models of itself and prior stages of initial awareness. Although unknown what else might be required for self awareness, the models and stages appear sufficient for explicitly capturing the AI's causal activity and resolution of conflicting moral goods. The systems approach differentiates between natural and social proximate goods and 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.