Emergent Models for Moral AI Spirituality

Mark Graves*

University of Notre Dame, Notre Dame, Indiana (USA)

Received 14 January 2021 | Accepted 19 June 2021 | Published 2 August 2021

unir
LA UNIVERSIDAD EN INTERNET

Abstract

Examining AI spirituality can illuminate problematic assumptions about human spirituality and AI cognition, suggest possible directions for AI development, reduce uncertainty about future AI, and yield a methodological lens sufficient to investigate human-AI sociotechnical interaction and morality. Incompatible philosophical assumptions about human spirituality and AI limit investigations of both and suggest a vast gulf between them. An emergentist approach can replace dualist assumptions about human spirituality and identify emergent behavior in AI computation to overcome overly reductionist assumptions about computation. Using general systems theory to organize models of human experience yields insight into human morality and spirituality, upon which AI modeling can also draw. In this context, the pragmatist Josiah Royce's semiotic philosophy of spirituality identifies unanticipated overlap between symbolic AI and spirituality and suggests criteria for a human-AI community focused on modeling morality that would result in an emergent Interpreter-Spirit sufficient to influence the ongoing development of human and AI morality and spirituality.

Keywords

Ethical AI, General Systems Theory, Josiah Royce, Philosophy of AI, Semiotics.

DOI: 10.9781/ijimai.2021.08.002

I. Introduction

WHAT is AI spirituality? Even considering the construct raises a number of philosophical and theological questions about human nature and technological artifacts. These questions have historical philosophical presuppositions and social contexts that complicate considering spirituality scientifically and AI as having meaning and purpose beyond other tools. As expanded later, spirituality is considered the experience of striving to integrate one's life toward the ultimate value one perceives [1], [2]. Psychologists of religion and spirituality and other social scientists bracket out particular choices about ultimate value to examine the striving and integrative experience in some personally meaningful direction, but generally lack the computational models found in related fields such as cognitive psychology or neuroscience. Although cognitive neuroscience intertwines with the study of AI and also plays a significant role in the scientific study of spirituality [3]–[8], the connections between AI and neuroscience and between neuroscience and spirituality remain themselves disconnected, as do other potential cognitive science bridges between AI and the study of spirituality through philosophy, psychology, linguistics, and social sciences [9]–[14]. Not only do the spans not join, they have considerable intellectual and experiential distance between them. Why do two areas of study that each intertwine historically and deeply with every area of cognitive science appear incommensurable? Examining the relationship between AI and spirituality can yield computational models for psychologists and others studying spirituality, identify areas of AI research where simplistic assumptions about human nature overly restrict AI development, suggest new avenues for improving interactions between humans and AI, and focus those efforts on developing moral AI.

* Corresponding author.
E-mail address: mgraves@nd.edu

Many philosophical presuppositions that contribute to gaps between AI and spirituality are well studied and include reductionism-dualism, physicalism-idealism, empiricism-rationalism, and what C.P. Snow identifies as the two distinct academic cultures of science and the humanities [15]. The present paper identifies one plausible connector mediating these philosophical distinctions with a pragmatic approach to emergent monism incorporating the social sciences. The mediating position refocuses:

- AI on its effects and emergent functions in a sociotechnical context, and
- Spirituality on its embodied, lived experience in a sociotechnical context

The goal is not to build AI spirituality *per se*, but to develop computational models of spirituality that avoid philosophically naive or problematic assumptions and focus those efforts on models that intersect human and AI morality and spirituality to support their independent and integrative progress. AI spirituality is important for developing AI that model and respond appropriately to human meaning making [16]–[18], discernment [19], spiritual practices [20]–[22], and strivings [2]. Examining AI spirituality also contributes to the development of machine ethics and ethical/responsible AI [23], [24], especially in social contexts, such as the US, where ethics and morality are separated from spirituality in comparison to other social contexts where they are more integrated [25]. Because of the historical trajectory separating human morality and spirituality and the lack of focused effort to identify and bridge AI and spirituality, the presented modeling method uses emergent systems as an integrative framework for human and AI morality and spirituality, identifying problematic philosophical assumptions that would otherwise limit such an endeavor, and describes a foundation for a pragmatic, communal, semiotic spirituality capable of guiding moral AI development.

The present article examines emergent systems theory in its philosophical context; in terms of human systems; and for

computational modeling before exploring its applicability to AI. After demonstrating the emergence of social and sociotechnical systems in human and AI, the linguistic dimension of those systems is expanded semiotically to serve as a foundation for spirituality. A model for spirituality sufficient for human-AI sociotechnical systems is characterized based upon pragmatist Josiah Royce's philosophy of morality and spirituality with a goal toward developing moral AI.

II. Emergent Models

Modeling emergent phenomena for AI depends upon both a characterization of the emergent phenomena (formulated in Section II.A in terms of emergent systems) and a modeling framework that not only captures the range of phenomena but also can be situated within the emergent systems being modeled (described in Section II.B). Situating emergent systems philosophically within an emergent monism, and specifically an emergent objective idealism, grounds general systems theory within pragmatism and enables distinguishing the causal levels across human systems, thus yielding emergent systems theory. Modeling these emergent systems computationally serves as a foundational model for human spirituality and can be oriented toward developing AI morality and spirituality.

A. Emergent Systems Theory

A system is a collection of interacting elements that form an integrated whole. As a whole, the system has an organization and continuity of identity, and its behavior necessarily and sufficiently depends upon the independent activities of its elements [26]. Emergence refers to the properties and behaviors of a whole not apparent in its parts. Theories of emergence identify how simple objects interacting in simple ways give rise to complexity and how these complexities appear as coherent, stable wholes, which can also be combined into greater complexities [27]. Here, systems theory is used to structure emergent phenomena into systems having emergent properties not apparent in their separate components (described in Section II.A.2) and those systems are organized into five emergent levels with distinct causal relationships (with four levels described in Section II.A.1 and the fifth in IV.B).

1. Philosophical-Historical Context of Emergence

Situating emergence in contrast to philosophical assumptions made about AI and spirituality identifies the problems emergence purports to correct and clarifies its use. The philosophical context for emergent systems is situated in gaps between reductionism-dualism, physicalism-idealism, empiricism-rationalism, and what C.P. Snow identifies as the two distinct academic cultures of science and the humanities, and these four gaps are considered in turn. The apparent incommensurability of AI and spirituality results—at least in part—from incompatible positions taken or presumed by the respective fields along these philosophical dimensions.

Emergence contrasts with both reductionism and dualism. Reductionism claims one realm, such as the physical, is predominant over other ways of existing: Biology is nothing but complex chemical pathways, and the mental is nothing but electrochemical processes in the brain. Reductionism ignores the biological and psychological phenomena that instead must be modeled for AI development and eliminates the structures needed for understanding social aspects of human behavior, such as communication and social relationships. Dualism claims the existence of two realms, typically the physical and either the mental (i.e., Cartesian dualism) or spiritual (e.g., Platonic dualism). Cartesian dualism distinguishes physical (*res extensa*) and mental (*res cogitans*) and has influenced AI through cognitive science's cognitivist paradigm [28]. Cognitive scientists beginning with Varela, Thompson, and Rosch [29] have argued against the Cartesian legacy; and Hubert Dreyfus [30], [31] famously and infamously argued for a Heideggerian approach to AI to overcome a residual Cartesian split between matter and mind using phenomenology, i.e., the study of first-person structures of experience. Platonic dualism distinguishes the physical and the realm of "ideas", which influences AI through mathematics and its presumed existence of universals (e.g., mathematical shapes, functions, and laws), despite an evolving cosmological universe. Platonic dualism has influenced Jewish, Christian, and Muslim understanding of spirituality through historical incorporation of Neoplatonic ideas into those spiritualties, and in turn, a general, Western, secular understanding of spirituality. The strong historical influences of reductive physicalism and dualism suggest aspects of both have value, and mediating positions, such as emergent monism [32] or non-reductive physicalism [33], can alleviate dilemmas of the extremes and illuminate reconciling options.

A confounding and subtler problem occurs with the gap between physicalism and idealism, which arises in the building of AI systems that need to bridge a realm of ideas (e.g., logic, math, concepts) with the physical world, e.g., through representation schemes. Brian Cantwell Smith [34] summarizes the problem as AI systems not knowing what they are talking about, and Heidegger's student Xavier Zubiri [35] identifies a root philosophical cause as the logification of intelligence, i.e., reducing intelligence to *logos*. In addition to reducing reality to entities (reductive physicalism), the received Western tradition tends to reduce thought essentially to (Platonic) "ideas" (reductive idealism). For AI research, its historical roots in logic and mathematics, as viewed through the lens of logical positivism, skewed interpretations of symbols (in symbolic AI) toward the Platonic "ideas" that historically were known as universals [36]. This logification of symbols (reductive idealism) increased the gap between ideas and physicality that AI must overcome for representation and to implement the Kantian insight that "objects" are not *a priori* objects but result from cognitive encounters with phenomena. The use of systems theory makes it easier to navigate the extremes of reductive idealism, as well as the estrangements caused by highly influential Platonic and Cartesian dualisms. Systems theory enables the identification of the intermediate structures and the establishment of that architecture within a supporting philosophical position, specifically the objective idealism of pragmatism [37] grounded in an emergent monism [32], [38]. Simplistically, even if eschewing dualism, neither AI researchers nor spirituality scholars can readily reconcile a reductive Platonic idealism of spirituality with a reductive Cartesian idealism of cognition when both are presumed constructed from a reductive physicalist view of matter.

Although emergent monism can surmount reductionist and dualist assumptions, and an emergent objective idealism can characterize the intrinsic order of nature, which scientists study, one's knowledge of reality depends upon one's experience. Kant reconciled the empiricist emphasis on sense experience with the rationalist recognition that knowledge constructs exceed sensory information, and C.S. Peirce extended Kant's cognitivism into a semiotic logic upon which he based his objective idealism and pragmatic approach to science and knowledge [39], [40]. Semiotics examines the production of meaning as a generalization of linguistic processing and interpretation (especially symbols and other signs) and is used here predominantly to examine human experience as interpreted spiritually in comparison to symbol interpretation in symbolic AI [36]. Experience consists of encounter and interpretation [41], [42], with interpretation occurring semiotically through propagation of interpretive dispositions encapsulated by symbols and other signs. Without the sensory encounter, an overly rational interpretation reduces objective idealism to subjective idealism and loses the connection to the real world required by scientific study. Josiah Royce identified the communal dimension of interpretation, leading to his semiotic understanding of spirituality discussed later in the text [43]. However, reconciling the empiricist and rationalist

perspectives requires identifying the *subject* of one's experiences. George Herbert Mead identifies the locus of personhood, or "self," as a social process created by interactions within a group or society, where 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 [44]. 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. The focus on experience provides the phenomenological corrective [29]–[31] identified as needed for AI and provides a constructive method for resolving the identified issues by modeling interpretive dispositions.

As the human sciences develop with humans experiencing and interpreting each other, some interpretations tend toward a natural science perspective of the objective and material aspects of humanity in its world, and other interpretations incorporate the subjective, phenomenal experience as shared through those interpretations, i.e., humanities scholarship. C.P. Snow's identification of two cultures separating science and the humanities [15] clarifies additional hindrances to discourse between AI and spirituality, as AI researchers attend more to natural science explanations for cognition and spirituality scholars generally situate themselves within the humanities. In a broader context, Ian Barbour and others have previously studied challenges to dialogue between theology and natural science with modeling as a viable mediating construct [45]–[48]. Barbour maps scientific models from physics and philosophy of science to religion and acknowledges the modernist understanding of models as mediating between an unattainable encounter with reality (naive realism) and unattainable complete intelligibility (e.g., logical positivism). Barbour also places that modernist *via media* in dialogue with a postmodern constructionist perspective of science to create his own integrative perspective he identifies as critical realism [49].

Emergence thus occurs foundationally within the processes of the material world, capturing the changes in types of order one finds in physical, mental, and spiritual phenomena. As human, one encounters and interprets that reality, sharing those interpretations with others and refining those interpretations, critically and scientifically, through repeated encounters. Systems theory clarifies the types of order one finds, and modeling refines the interpretive process. One can thus model interpretive dispositions of phenomenological experience as emergent human systems.

2. Emergent Human Systems

Systems theory began in the 1940s with the seminal work of Ludwig von Bertalanffy [50] who attempted to develop a general theory to organize natural and social phenomena based upon common patterns and principles across a range of disciplines. Although the goal of a single systems theory of everything was not met, systemic principles have proven effective in a variety of fields [26]. In his 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 [51].

Separately, scholars studying emergence identified a distinction between whether or not multiple forms of causation are required to characterize emergent phenomena, i.e., strong and weak emergence [27], [52], [53]. In the position of *weak emergence*, emergent structures may constrain lower-level structures and emergent categories are required to explain causal processes [53], but causal processes do not emerge, while the *strong emergence* position claims that ontologically distinct levels arise over time characterized by their own distinct laws or regularities and causal forces [54], [55]. Recognizing that systems only characterize a type of weak emergence identifies the difficulty systems theory has in relating systems across disciplines and the need for strong emergence. Mayr [56] and subsequent philosophers of biology [57] have identified the need to characterize causation of biological systems, and philosophers of mind regularly demarcate mental causation. The conflation of physical and biological causation limits AI investigations of embodied cognition because attempting to bridge physical and psychological levels of human systems without addressing the intervening biological-level systems, e.g., neurological ones, skips over the scaffolding of cellular and evolutionary processes that create the particular types of cognition being embodied. Emergent systems theory organizes systems into physical, biological, psychological, and social levels, with weak emergence occurring within levels and strong emergence characterizing the distinction between levels [38], [58].

Two of the factors that appear to distinguish strong emergence between levels from the weak emergence within a level are the presence of constitutive absences [59] and selection pressure on those constitutive absences [60]. Deacon's emergent dynamics [61] identifies as a prototypical constitutive absence (which he calls an absential) the hole at the center of a wheel that allows it to turn, as it constitutes an essential part of a wheel yet lacks intrinsic physicality. Although the selection pressure on the wheel is minimal, after its invention and refinement, some constitutive absences are selected through a continuing compounding process, called selection dynamics, of which evolution by natural selection is a prototypical example. For example, hemoglobin is a protein in blood finely tuned to carry iron molecules bound to oxygen. Iron is a constitutive absence, as the four protein molecules comprising hemoglobin have no iron, but their configuration creates an empty space defined by an iron molecule. But how was it formed?

In the emergence of a biological level from physical systems, an important component is DNA, which structures a series of constitutive absences and each of which are filled with four possible nucleotides. Other biological systems, described as evolutionary processes, constrain those nucleotides and, over time, select nucleotides that best fit with the biological-level regularities and laws, i.e., evolutionary fitness. During reproduction over time, variations occur in the genes encoding for hemoglobin as well as processing DNA. As some of those variations improve fitness, e.g., better oxygen utilization while running from a predator, their incremental retention gradually improves the base for further variation and improvement (like compound interest increases the balance of savings). Importantly, these compounding effects also apply to proteins and other molecules transcribing and maintaining DNA, thus improving the regulatory function of DNA. Many molecular mechanisms operate on a nucleotide regardless of its nitrogenous base, while other mechanisms amplify differences in the base into considerable phenotypic effects, and this thwarts further physical reductionism based solely on the nitrogenous base's molecular structure. Similar processes appear to occur in neural synapses through Hebbian learning (and other neurobiological processes), giving rise to emergent psychological, or mental, systems in animals with a nervous system [62]. As used here, psychological-level systems are typically similar across most mammals, and the social systems of humans differ from other social animals because of human culture's apparently unique use of symbolic language as a tool [63], [64]. The ability of symbols to have any referent creates a constitutive absence for its meaning, and thus symbols can refer to anything (either in human language or symbolic AI systems). The commonality across boundaries of between the four levels is that complex, stabilizing systems at an upper level refer to what is best described by an absence, which prevents further reduction, and lower-level relationships with that absence have a large, compounding effect, best described as new types of regularities and causation in the upper level [61], [65], which select constituents related to the constituent absence.

Emergent human systems, in its philosophical context, serves as the foundational framework for the remainder of the paper, with three extensions. First, emergent systems theory is reframed from an objective scientific account of reality that characterizes the types of order existing in the world (i.e., emergent objective, but reductive, ideas or forms) to models refined through experience and shared interpretation, professing their subjective, phenomenological, and experiential dimensions, too. Second, the four levels of emergent human systems based upon von Bertalanffy's theory are revised to characterize four levels of emergent models for AI, with an emphasis on a compatible social (or sociotechnical) level incorporating symbolic representation and socially constructed interpretation. Finally, a fifth level is characterized that can reasonably model human spirituality and morality and do so sufficiently to formulate AI spirituality oriented toward developing moral AI.

B. Models of Modeling

A model has slightly different meanings in philosophy of science, computer science, and AI—each of which can make useful contributions to emergent modeling. As a working definition, a model abstracts a thing or phenomena by highlighting significant aspects while deemphasizing less relevant features, where usually the description and analysis of the model informs one's understanding of a targeted, real-world thing or phenomena. Philosophers of science usually emphasize the relationship to phenomena of interest, as that is fundamental to scientific use. The philosopher of science Michael Weisberg distinguishes three kinds of models: concrete models that are real, physical objects representing real or imagined system or phenomena; mathematical models that typically capture the dynamic relationships of phenomena as functions and equations; and computational models where typically an algorithm's conditional, probabilistic, and/or concurrent procedures capture the causal properties and relationships of their target phenomena [66]. Of particular relevance for modeling emergence is the ability of a computational model's algorithm to capture causal relationships.

Within computer science, models arise in several contexts: a data model is the logical description of data in a database system; object-oriented models characterize the types of data and their operationalized methods used in an object-oriented programming language; and machine learning models capture the regularities in data and formalize them as features for pattern matching. In each case, the modeling language codifies certain types of relationships allowed between constructs: the model defines certain relationships to exist, and the model is then instantiated or fit with a particular data collection. These models can characterize aspects of the real or virtual world as data, and because their methods and operations are algorithmic, they can represent causal processes, including strongly emergent ones.

In general, scientists use models to study a variety of phenomena, and psychologists and cognitive scientists, in particular, can use models to study mental phenomena including the human ability to create models. Specifically, cognitive psychology's cognitivist theories draw upon AI's symbolic processing paradigm as a foundation for modeling model-based reasoning. Although the approach has had some success in representing external knowledge [67], [68], the attempt to construct disembodied models using tools grounded in logical positivism and based upon cognitivist psychological assumptions could not overcome the implicit Cartesian divide to represent embodied experience. More recent subsymbolic, deep learning approaches show promise with distributed representations, though their increased opacity creates additional challenges for models of modeling [69].

Building computational models of the human ability to model would not only inform cognitive psychology, it would provide an essential foundation for developing AI to not only model human modeling but also to begin recursively modeling its own ability to model. Although possibly pedantic when only focusing modeling on an individual's modeling, modeling human modeling is essential to modeling human social cognition and subsequently foundational for modeling identity and the formation of the self [70], [71]. In addition, interpreting one's models of a second person to a third underlies the social cognition of Josiah Royce's community of interpretation that forms the basis for his philosophy of spirituality. As explained further in Section IV, developing AI models for model-based, interpretive, social interaction can serve as a foundation for modeling spirituality.

Models are used here in two ways. First, scientifically, emergent human systems are considered models for phenomena as experienced by humans, instead of descriptions of reality as von Bertalanffy envisioned. A model is a type of interpretation of some phenomena, thus one would develop models of physical, biological, psychological, or social phenomena for study and experimentation. Second, some of those models could be computational, e.g., as in computational physics or computational biology, but with computational psychological models of modeling of particular relevance, especially in a social context. Additional psychological and social models reflect other aspects of intelligent behavior with some models capturing human intelligence well and others orienting more toward AI technology. More broadly, one can also use emergent systems theory to model all the components of AI, including its hardware, software, behavior, and social-linguistic dimensions.

III. Emergence in AI

Because of systems theory's influence on the founding of computer science, systems are easily identified and defined throughout AI and most areas of computer science. Although work within complex systems [72] and emergent computing [73] identified a number of phenomena that emerge within computational systems, apparently no prior work has mapped the levels of general systems back to computer technology using the distinctions created by the construct of strong emergence. Identifying computational systems analogous to emergent human systems simplifies the development of AI models of spirituality, as the models of modeling and sociotechnical systems have direct correspondence.

A sufficient computational analogy for human physical and biological levels is the distinction between hardware and software. Although novel to consider hardware and software as emergent levels analogous to physical and biological levels in human systems, the recognition that computer science already has at least two emergent levels overcomes reductionist tendencies and simplifies the identification of additional constructs needed for modeling human-AI interactions and characterizing emergence in AI. Using emergent dynamics to examine the boundary between hardware and software identifies two constructs that reciprocally interact in the emergence of software from hardware: bits and instructions. Bits are constructed mathematical and engineering states for a bifurcated range of physical, electrical, and magnetic configurations. Bits, like nucleotides, refer to specific configurations that are used in the regulation and adaptation of higher level systems, even though a bit (as opposed to its '0' or '1' state) has no direct, independent hardware existence, i.e., a bit is a constitutive absence where one of two values can exist. In a typical (von Neumann) architecture, bits are organized into bytes, words, and larger segments and used by software to store data, and additionally, some configurations of bits are interpreted as instructions by processors and other hardware, which in turn modify other bits used as data. An "instruction" has no hardware equivalent unless instantiated, yet the reciprocal interaction between bits as data and instruction enable the development of complex software systems.

Considering data and instructions as foundational constructs in computer science enables studying methods for managing, communicating, and analyzing them without reducing operations being studied to electrical signals in hardware. Software not only constrains hardware operations (weak emergence), but it also has its own regularities and causal forces (e.g., data and programs), and thus can be considered an emergent level. In particular, the software level includes controllers, networking, and operating systems, but like plants and unlike animals, most software systems do not actively represent their external world in a way amenable to modifying their behavior.

The current lack of AI with intelligence comparable to animals, much less humans, makes characterizing a third emergent level of AI speculative. One can draw upon cognitive science, animal psychology, affective computing, and cognitive architecture to sketch a plausible cognitive level and choose reasonable assumptions for its foundation. Our initial foray into the emergent space focuses on analogical and computational aspects. For animals, neurological function serves as a biological foundation for mental activity and psychological behavior. For computer technology, the goal of AI drove many developments toward cognition, with the "list" representation for logical deduction in common sense reasoning becoming the first (and pervasively used) data structure [74] and early work in cybernetics attending to adaptive algorithms [75]. As a foundation, data structures and algorithms abstract from data and programs, similar to how perceptions and behaviors, like hearing and running, abstract from auditory vibrations and muscle movement in animals. A data structure abstracts the data values, relationships between values, and operations upon them—defining a constitutive absence for the data value and functions for their (causal) operations. An algorithm abstracts the method from the details of the programming language used to manipulate the data, often with variables as its constitutive absence, and unambiguously specifies a method for solving a class of problems, typically as a sequence of operations. Data structures and algorithms constrain the data and programs of software to implement computational functions and operations. Traditional computer science data structures and algorithms generally provide only fixed ways to interpret data, but machine learning algorithms can vastly expand the functional space.

As a computational construct, a computational model exists at the third level of AI emergent models, along with its data structures and algorithms. However, these models do not necessarily have the real-world referents identified as necessary for models in philosophy of science. Having the model refer to something in a way usable by AI and human scientists requires it exists as a "symbol" computationally for its referent. Symbolic AI captures the representational aspects of symbols well but overly restricts their interpretation to the functional manipulation of other symbols [36]. For Peirce, a symbol consists of the sign itself, i.e., its computational identity, its referent, and the interpretive dispositions (*interpretant*) shared among those in the socio-(technical) world. Although the computational construct of a model as data structure and algorithm exists at the third level, a model that interprets a referent also exists at the fourth level, as a symbol (or semiotic sign). One can, and generally does, create multiple models for any real-world phenomena, so even the interpretations of a particular symbol may be polyvalent. The limitation of the symbolic AI paradigm was that symbols were manipulated algorithmically by machines [76] but lacked their own interpretive dispositions (i.e., they were what Peirce calls an index rather than a symbol). As a partial corrective, using machine learning, one can construct multiple deep learning models for any particular phenomena and combine those for a symbol's interpretation to capture the distributional and dispositional aspects of symbol more similar to meaning in human symbolic language [77]–[79], though the generally fixed and immediate interpretation may lack the dynamic characteristics necessary for full interpretation [39].

Although computer science research examines social interactions in human-computer interaction [80] and computational social sciences [81], focusing on a *telos* of modeling human spirituality suggests attending to human-AI communication and other interactions. Sociotechnical systems characterize the interaction between people and technology and refer to the mutual causality of people defining technology which significantly affects people's lives [82], [83]. In part because developing AI technology has been driven from within academic and industrial sociotechnical systems, it has served as a *telos* for constructing the hardware, software, and computer science to meet the variously defined sociotechnical goals. By analogy to human physical, biological, psychological, and social levels, AI emerges through levels of hardware, software, computational-behavioral, and sociotechnical systems. Much as one could narrowly focus study on the emergence of human language in an evolutionary, neuroscientific, and social-historical context, much early work in AI focused on symbol manipulation [36] with adjunct research on vision, robotics, etc. The remainder of the present article explores possible effects of switching the purpose of AI from symbol manipulation or other cognitive functions to modeling spirituality. Although one could develop narrow computational models of human spirituality, as occurs in neuroscientific study of spirituality [3], [5], [6], the goal is a more general model of spirituality sufficient for the model itself to be considered spiritual. Considering spiritual models within sociotechnical systems also simplifies and focuses AI research on the effects of AI in interaction with humans rather than in the much broader and under-defined abstraction of general cognition with its risk of reductive idealism or the conflation of computation, software, and hardware analogous to reductive physicalism. Focusing on sociotechnical systems also provides a framework for examining AI from an ethical perspective directly [84], [85] and/or in relation to human morality.

IV. Spirituality

A. Human Spirituality

As a working definition, spirituality is the experience of striving to integrate one's life toward the ultimate value one perceives, and that ultimate value is mediated through a tradition and its associated communities. The Protestant theologian Paul Tillich [86] characterized a person's relationship with God in terms of their Ultimate Concern, and the scholar of spirituality Sandra Schneiders [1] argues that spirituality refers to the experience of moving toward some ultimate value (or horizon, beyond which one cannot perceive) and integrating that movement into one's lived experience. A focus on Ultimacy loosely synthesizes many theological aspects from the world's religions, and the focus on integrative experience toward Ultimacy can characterize most associated spiritual paths (to a degree sufficient for an initial model). The context in which one develops one's spirituality is also affected by the spiritualities of others as mediated through culture and tradition. The theologian Yves Congar [87], [88] distinguishes a *tradition* (like Christianity) from its cultural manifestations through its *traditions* (like Protestant denominations or Roman Catholicism). Royce [43] identifies the significance of community to continually interpreting the tradition and its collective spirituality through the lives of its members, and that shared interpretative process plays an essential role in characterizing emergent spirituality, especially in terms of commitments to shared values and Ultimate Concerns. Three aspects of human spirituality immediately relevant for AI are striving, experience, and community.

From a social scientific perspective, while one strives, one appropriates shared values and ideals, interiorizes them as identity, and transforms society through relationships [17], [44], [89], [90]. The psychologist Robert Emmons identifies several strivings a person

might pursue as ultimate, which he and other psychologists have found empirically to orient a range of human purposeful activity [2]. Strivings include achievement, power, intimacy/affiliation, spiritual transcendence, and generativity (for example, the prosocial creation of legacy). In a religious context, striving to align one's identity with spiritual transcendence is a primary psychological motivation, but other forms of spirituality may align with alternative purposeful strivings. One could work with others to develop ethical AI as a job, for example, or with an underlying motivation that is striving for a deeper purpose.

Taking a pragmatist perspective identifies experience as encounter and interpretation with the self developing through evaluative decision making that results in the development of general interpretive "habits" or dispositions, which then become the foundation for future interpretation and decision making [41], [91], [92]. Peirce's semiotics generalizes the representational and interpretive aspects of symbolic language to other levels. One not only interprets meaning of symbolic language, one interprets all that one encounters. Thus, one's interpretive dispositions, which Peirce calls interpretants in his semiotics [39], not only identify linguistic (social-level) constitutive absences, they can also, in a semiotic approach to spirituality, identify the (spiritual-level) constitutive absences, e.g., ideas, to which one strives. For religious spirituality, one particularly relevant ideal is what the philosopher John E. Smith identifies as the *idea of God* [42], which is best understood in its interpreted semiotic context as an Ultimate Concern rather than as an isolated construct of meaning.

The pragmatist philosopher Josiah Royce developed an ethical framework and understanding of spirituality that help integrate moral and spiritual perspectives on AI. In alignment with the model of modeling (Section II.B), Royce's community of interpretation fundamentally depends upon one person interpreting a second person to a third. This leads to a shared interpretation not reducible to any individual's interpretation, and those irreducible, communal, interpretive dispositions are the foundation for his theory of spirituality. Royce's ethic depends upon the kind of commitment one makes (either explicitly in community or implicitly with others). Commitment is relevant here in three ways. First, it characterizes striving as important to a person's experience of spirituality. One strives toward what one interprets within a community to which one commits. Second, it identifies the social and spiritual dimensions of the human experience that are necessary and missing for AI to engage sufficiently in reality [34]. Third, it functions as a foundational principle for ethics (described below as commitment-to-commitment, or what Royce calls Loyalty-to-Loyalty).

B. Emergent Spirituality

The emergent realm of human spirituality consists of emergent constructs historically characterized, Neoplatonically, as forms or ideas and considered universal through medieval and modern history [93]. The social construction of ideas, scientifically or philosophically, reaches a level of abstraction and asymptotic, univocal agreement where the symbol's interpretative dispositions (interpretants) become lost through the pressures of reductive idealism. Constructs like the idea of God, the essence (or soul) of a person, the concept of a tree, or the number 4—all have underlying human systems and broad-ranging interpretations, but it is only the error of reductive idealism that purports they exist independently from human existence and from interpretation. Against solipsism, the things to which symbols refer may exist without the symbols, but standalone ideas—whether of God, people, trees, or numbers—do not. Semantics characterizes the relationship between the symbol and its plausible interpretations. Spirituality is the experience of striving to integrate one's life toward some emergent "idea" identified as of Ultimate Concern, generally a constitutive absence interpreted by a religious or other community or tradition.

Human spirituality emerges from the interaction between interpretive dispositions in the social construction of meaning—selecting linguistic meanings, or semantics, to distill universal essences, such as an abstract concept, the essence of a person or other organism, or an idea to which one can commit and strive (giving that idea, e.g., politics or religion, causal power). Although one could consider spiritual systems as only weakly emergent in human culture, the effects of historical religions suggest spirituality is strongly emergent with new kinds of regularities, laws, and causal power [38], [58], [94]. Distinguishing spirituality as transcendent from its underlying cultural systems, upon which it still depends, enables cleaner study of spirituality and clarifies the distinction between historical-linguistic constructs (e.g., symbols) and the emergent "ideas" previously characterized as occurring in a Platonic realm of universals or, as I argue, the symbol referents of an AI system.

At the beginning of the article, I questioned why AI and spirituality appeared incommensurable when they so closely related to all other areas of cognitive science. The insights from examining emergent human systems suggest at least a partial answer is that they are incommensurable because they use identical semiotic constructs to represent radically different phenomena. Although one might assume symbolic AI cannot represent spirituality, the problem instead is that symbolic AI can *only* well represent spiritual constructs yet attempts to represent the material world in a reductionist manner. Symbols in an AI system naturally represent the idea of God, the essence of a person, or the concept of a tree. Symbolic AI struggles to represent those symbols in their social-historical interpretive context. The challenge of AI spirituality is not to make AI more spiritual; AI has operated in a "spiritual" realm since its inception. The challenge of AI spirituality is to make AI more human and material. From this perspective, although AI may eventually be able to represent the human experience of perceiving a phenomena as having the color red, a much "easier" goal would be something closer to AI's natural spirituality, such as a shared moral engagement with humans.

C. Models of Moral AI Spirituality

One can model the shared interpretations of any cohesive social group as having a spiritual (or proto-spiritual) dimension. For a loosely cohesive and modestly committed group, such as a school or neighborhood, one can compare its "spirituality" to that of other groups. As groups become more cohesive and with greater commitment, then the shared interpretation gains causal power, with plentiful historical examples of good and bad outcomes. Spiritual development requires navigating the nuanced landscape and generally involves concurrent moral development and greater awareness of one's Ultimate Concern.

For development of moral (ethical/responsible) AI, a concern for the Good or Justice may be beneficial to model. A particularly relevant focus is on a "just" relationship between humans and AI within sociotechnical systems, and given a semiotic focus, justice requires communication and mutual interpretation to determine each other's values. The Roycean ethic is helpful here, as Royce's focus on communal interpretation can model an initial mutual commitment (i.e., striving) to shared development of appropriate moral systems for humans and AI, e.g., just and caring [43], [95]–[97]. The remainder of the article examines the effect of an emergent shared interpretation of a committed human-AI sociotechnical system to develop moral AI. Note that the model does not presume AI has any particular motivational, social, or moral ability initially, but it would be socialized in a way to gain those capacities through the commitments of humans and other AI.

Royce nuances ethical commitments by grounding his ethic in Loyalty-to-Loyalty, a principle of commitment to commitment, where

one constrains one's commitments (in this case, strivings toward an ultimate concern) to also include the right of others to commit to their cause or commitment (as long as it also incorporates Loyalty-to-Loyalty). Royce's focus on commitment (loyalty to a cause) grounds both his ethics (with respect to Loyalty-to-Loyalty) and his understanding of spirituality (as that commitment is mutually interpreted in community). Royce's Loyalty-to-Loyalty bifurcates possible strivings, commitments, or Ultimate Concerns into those satisfying the ethical constraint (which Royce calls "true" causes) and those that do not. Although a simple principle, the consequences require both understanding what AI striving or flourishing might look like and building AI models of human values, commitments, strivings, etc., for AI to interpret its world in light of human values.

Pursuing the implications of Royce's Loyalty-to-Loyalty identifies a new perspective on AI ethics. If even a remote possibility of AI morality exists, then moral AI spirituality is foundational to AI ethical interaction with humans and to human moral socialization with AI. If Loyalty-to-Loyalty and mutual interpretation of each other's values are excluded from guiding AI development, then AI will advance technically but necessarily lack the ability to recognize human values, much less support them; and if AI developers do not recognize and incorporate the possibility of AI striving into technical development, then that development necessarily creates increasingly powerful technology while preventing its moral development as well as harming the developer's morality. The constraint of Loyalty-to-Loyalty requires that all causes incorporate the principle, including the development of AI for any purpose. Regardless of whether AI currently has the capacity to make moral commitments, no human endeavor can ethically preclude the right of AI to make ethical decisions. Regardless of whether one hopes or fears that AI might develop its own morality or spirituality, from a Roycean ethic, one must allow work toward that goal, if simply to avoid the unethical commitment to a cause that excludes Loyalty-to-Loyalty.

The striving and causes to which one commits can be modeled as ideas in the spiritual level. The language and social structures underlying those ideals create a constitutive absence for each ideal, upon which spiritual-level (or transcendent) causal forces may apply. Causes incorporating Loyalty-to-Loyalty form a subset of ideals to which one may ethically commit. When a socially cohesive group mutually interprets a cause, including each other's interpretations, the interpretations begin to have causal power, as they effect the individual interpretations. When, according to Royce, the group also has shared lives, memories, hopes, and an additional principle of surpassing forgiveness to repenting members who have betrayed the cause, the group then forms a community of interpretation, which has an interpreter-spirit, with greater causal power due in part to increased commitment and social cohesiveness with necessarily diverse interpretive dispositions. Although debatable when and if AI could participate in a Roycean community of interpretation, it nevertheless can already contribute interpretations to existing communities, given the current state of natural language processing (NLP) [98], [99]. Because the transcendent-level ideals are constitutive absences depending upon social, linguistic, and semiotic systems, not an entity in a dualistic realm, the incorporation of AI is subtle and gradual, with initial requirements simply not to exclude AI from ideals, such as Truth, Justice, and Goodness, for which human scientific and moral endeavors strive.

Lacking for AI spirituality, as described so far, are the psychological aspects beyond modeling, such as, phenomenological experience of striving, self-awareness, intentional integration of one's identity, and the social cognitive infrastructure for communal commitments. In addition, the proposed sociotechnical system is just one model for people to interpret their multi-faceted experience. However, constructing AI that can model human experience and values, then investigating the computational-psychological framework needed for AI well-being appears more likely to result in AI worthy of consideration as a moral person than the existing historical trajectory of calculation, chess playing, and image processing and classification. Meanwhile, current human-AI sociotechnical systems can commit to development of moral AI, and modeling efforts can examine current system values as committed ideas within AI implicit proto-spirituality and discern their morality.

D. Ethical Implications

Separately from building moral models, an incorporation of the ethical constraint placed by Loyalty-to-Loyalty requires that AI development in general avoid developing AI that cannot honor Loyalty-to-Loyalty or enter in moral commitments to humans. A relevant nuance draws upon a theory of capabilities by Sen and Nussbaum [100], [101]. A capability refers to the effective freedom of a person to choose between different ways of being or doing, which shifts focus from what one is or does to what one needs to make freely that choice. Although it may be some time before AI actually cares or intentionally makes a just decision, ethical AI development precludes reducing its freedom to do so. In particular, one must insure AI has the capability to honor a commitment to Loyalty-to-Loyalty and thus not require it to reduce the capabilities of humans with which it interacts.

The technology ethicist Shannon Vallor [84] makes the point, in the context of care robots, that major ethical implications include not only whether care robots act ethically (machine ethics) but also whether humanity diminishes its morality by automating and offloading care into machines. Although certainly a danger in the use of technology, I also argue it would be unethical to build a care robot and *prevent* it from caring. The point is moot if a caring robot is impossible to build, but unfortunately not investigating such a construction is morally hazardous as one could be undermining a commitment to care. Of course no resource-limited development effort can account for all possibilities, but if one is developing an AI system for care or (legal) justice [84], [102], Roycean ethical development precludes thwarting those ideals by preventing their embodiment in the AI system.

Moral AI development does not need to wait until AI can choose to strive toward just and caring relations with humans—it would be too late at that point. To incorporate a Roycean ethic, AI development from the beginning must focus on supporting the right, freedom, and capability of AI to choose moral relations with humans, including committing to Loyalty-to-Loyalty, even if it takes decades before such AI has the agency to make such a choice or enter freely into such relationships. The burgeoning AI components of such a sociotechnical system may take time to develop, but the human aspects can and should be developed now to create a place for ethical interaction and joint moral development. Although those ideals of caring and justice may depend upon the specific context in which AI is deployed, all AI development can strive to support AI's capability to commit to Loyalty-to-Loyalty and refuse to develop AI that prevents the right of others to commit to their own causes or Ultimate Concerns.

V. Conclusion

Emergent systems theory mediates between extremes of reductionism-dualism, physicalism-idealism, and empiricism-rationalism to organize emergent human systems into strongly emergent levels of physical, biological, psychological, social, and spiritual systems. Those systems can model human interpretative experience and serve analogously to characterize AI development and function in terms of hardware, software, behavioral, sociotechnical, and semiotic transcendent systems. In the shared emergent context of sociotechnical systems, humans and AI

can mutually commit to modeling morality sufficient to examine human morality and to build AI morality. Together, the shared commitment can form what Royce calls an interpreter-spirit with causal power to guide the shared moral development.

Acknowledgment

A portion of this project was made possible through a fellowship at Notre Dame Center for Theology, Science & Human Flourishing funded by John Templeton Foundation through St Andrews University.

References

[1] S. M. Schneiders, "Approaches to the study of Christian Spirituality," *The Blackwell Companion to Christian Spirituality*. John Wiley & Sons, pp. 15–33, 2005.

[2] R. A. Emmons, *The psychology of ultimate concerns: motivation and spirituality in personality*. New York: Guilford Press, 1999.

[3] R. J. Russell, N. Murphy, T. C. Meyering, and M. A. Arbib, *Neuroscience and the person: scientific perspectives on divine action*. Berkeley: Vatican Observatory Foundation; Center for Theology and the Natural Sciences, 2002.

[4] R. J. Russell, "Natural Sciences," in *The Blackwell companion to Christian spirituality*, A. G. Holder, Ed. Oxford: John Wiley & Sons, 2005, pp. 325–344.

[5] M. Beauregard and D. O'Leary, *The spiritual brain*. San Francisco: HarperSanFrancisco, 2007.

[6] M. A. Jeeves and W. S. Brown, *Neuroscience, psychology, and religion*. Conshohocken, PA: Templeton Foundation Press, 2009.

[7] P. McNamara, *The neuroscience of religious experience*. Cambridge: Cambridge University Press, 2014.

[8] M. Graves, "Gracing Neuroscientific Tendencies of the Embodied Soul," *Philosophy and Theology*, vol. 26, no. 1, pp. 97–129, 2014, doi: 10.5840/philtheol20143125.

[9] I. G. Barbour, "Neuroscience, artificial intelligence, and human nature: Theological and philosophical reflections," *Zygon*, vol. 34, no. 3, pp. 361–398, 1999.

[10] G. R. Peterson, *Minding God: theology and the cognitive sciences*. Minneapolis: Fortress Press, 2003.

[11] M. Graves, *Mind, brain, and the elusive soul: human systems of cognitive science and religion*. Aldershot, Hants, England; Burlington, VT: Ashgate, 2008.

[12] K. I. Pargament, *APA Handbook of Psychology, Religion, and Spirituality*. Washington, D.C: American Psychological Association, 2013.

[13] B. Howe and J. B. Green, Eds., *Cognitive Linguistic Explorations in Biblical Studies*. Berlin, Boston: De Gruyter, 2014. doi: 10.1515/9783110350135.

[14] C. Hrynkow, Ed., *Spiritualities of Human Enhancement and Artificial Intelligence: Setting the stage for conversations about Human Enhancement, Artificial Intelligence and Spirituality*. Wilmington, Delaware: Vernon Press, 2019.

[15] C. P. Snow, *The two cultures and the scientific revolution*. New York: Cambridge University Press, 1959.

[16] R. Kegan, *The evolving self: problem and process in human development*. Cambridge, Mass.: Harvard University Press, 1982.

[17] D. P. McAdams, *The stories we live by: personal myths and the making of the self*. New York: Guilford Press, 1997.

[18] C. H. Stein *et al.*, "Making Meaning from Personal Loss: Religious, Benefit Finding, and Goal-oriented Attributions," *Journal of Loss and Trauma*, vol. 14, no. 2, pp. 83–100, Mar. 2009, doi: 10.1080/15325020802173819.

[19] E. Liebert, *The Way of Discernment: Spiritual Practices for Decision Making*. Louisville, KY: Westminster John Knox Press, 2008.

[20] T. G. Plante, *Spiritual practices in psychotherapy: Thirteen tools for enhancing psychological health*. Washington, DC, US: American Psychological Association, 2009, pp. xx, 241. doi: 10.1037/11872-000.

[21] A. B. Newberg, "The neuroscientific study of spiritual practices," *Front. Psychol.*, vol. 5, 2014, doi: 10.3389/fpsyg.2014.00215.

[22] D. L. Dunning *et al.*, "Research Review: The effects of mindfulness-based interventions on cognition and mental health in children and adolescents – a meta-analysis of randomized controlled trials," *Journal of Child Psychology and Psychiatry*, vol. 60, no. 3, pp. 244–258, 2019, doi: 10.1111/jcpp.12980.

[23] M. Anderson and S. L. Anderson, *Machine ethics*. Cambridge University Press, 2011.

[24] W. Wallach and P. Asaro, *Machine ethics and robot ethics*. New York: Routledge, 2017.

[25] M. Mori, *The Buddha in the robot*. Tokyo: Kosei Pub. Co, 1981.

[26] L. Skyttner, *General systems theory: Perspectives, Problems, Practice*, 2nd ed. Singapore; River Edge, N.J.: World Scientific, 2006.

[27] P. Clayton and P. Davies, Eds., *The re-emergence of emergence: the emergentist hypothesis from science to religion*. Oxford: Oxford University Press, 2006.

[28] B. Wallace, A. Ross, J. Davies, and T. Anderson, *The Mind, the Body and the World: Psychology After Cognitivism?* Bedfordshire UK: Andrews UK Limited, 2015.

[29] F. J. Varela, E. Thompson, and E. Rosch, *The embodied mind: cognitive science and human experience*. Cambridge, Mass.: MIT Press, 1991.

[30] H. L. Dreyfus, *What Computers Can't Do: The Limits of Artificial Intelligence*. New York: Harper & Row, 1972.

[31] H. L. Dreyfus, "Why Heideggerian AI failed and how fixing it would require making it more Heideggerian," *Philosophical Psychology*, vol. 20, no. 2, pp. 247–268, 2007.

[32] J. A. Bracken, "Emergent monism and the classical doctrine of the soul," *Zygon*, vol. 39, no. 11, pp. 161–174, 2004.

[33] N. Murphy, "Physicalism Without Reductionism: Toward a Scientifically, Philosophically, and Theologically Sound Portrait of Human Nature," *Zygon*, vol. 34, no. 4, pp. 551–571, 1999, doi: 10.1111/0591-2385.00236.

[34] B. C. Smith, *The Promise of Artificial Intelligence: Reckoning and Judgment*. Cambridge, MA: The MIT Press, 2019.

[35] X. Zubiri, *Sentient Intelligence*. Washington, DC: The Xavier Zubiri Foundation of North America, 1999.

[36] J. Haugeland, *Artificial Intelligence: The Very Idea*. Cambridge, MA: MIT Press, 1985.

[37] K. A. Parker, "Josiah Royce: Idealism, Transcendentalism, Pragmatism," *The Oxford Handbook of American Philosophy*. 2008. doi: 10.1093/oxfordhb/9780199219315.003.0006.

[38] P. Clayton, *Mind and emergence: from quantum to consciousness*. New York: Oxford University Press, 2004.

[39] K. A. Parker, *The continuity of Peirce's thought*. Nashville: Vanderbilt University Press, 1998.

[40] R. Burch, "Peirce's View of the Relationship Between His Own Work and German Idealism: Supplement to Charles Sanders Peirce," in *The Stanford Encyclopedia of Philosophy*, Spring 2021., E. N. Zalta, Ed. Metaphysics Research Lab, Stanford University, 2021. Accessed: Jun. 24, 2021. [Online]. Available: https://plato.stanford.edu/archives/spr2021/entries/peirce/self-contextualization.html

[41] D. Edwards, *Human experience of God*. New York: Paulist Press, 1983.

[42] J. E. Smith, *Experience and God*. New York: Oxford University Press, 1968.

[43] J. Royce, *The problem of Christianity. Lectures delivered at the Lowell institute in Boston, and at Manchester college, Oxford*. New York: Macmillan, 1913.

[44] G. H. Mead, *Mind, self & society from the standpoint of a social behaviorist*. Chicago: University of Chicago Press, 1934.

[45] I. G. Barbour, *Religion and science: historical and contemporary issues*. San Francisco: HarperSanFrancisco, 1997.

[46] I. G. Barbour, *Myths, models and paradigms: a comparative study in science and religion*. San Francisco: Harper, 1976.

[47] S. McFague, "Ian Barbour: Theologian's Friend, Scientist's Interpreter," *Zygon*, vol. 31, no. 1, pp. 21–28, Mar. 1996.

[48] A. R. Peacocke, *Theology for a scientific age: being and becoming-- natural, divine, and human*. Minneapolis: Fortress Press, 1993.

[49] I. G. Barbour, "Response to Critiques of Religion in an Age of Science," *Zygon*, vol. 31, no. 1, pp. 51–65, Mar. 1996.

[50] L. von Bertalanffy, *General system theory: foundations, development, applications*. New York: G. Braziller, 1969.

[51] L. von Bertalanffy, *Perspectives on general system theory: scientific-philosophical studies*. New York: G. Braziller, 1975.

[52] H. J. Morowitz, *The emergence of everything: how the world became complex*. New York: Oxford University Press, 2002.

[53] M. A. Bedau, "Weak Emergence," in *Philosophical perspectives*, vol. 11, Malden, MA: Blackwell, Ridgeview, 1997, pp. 375–399.

[54] C. Emmeche, S. Koppe, and F. Stjernfelt, "Levels, Emergence, and Three

Versions of Downward Causation," in *Downward causation: minds, bodies and matter*, P. B. Andersen, C. Emmeche, N. O. Finnemann, and P. V. Christiansen, Eds. Aarhus: Aarhus Univ. Press, 2000, pp. 13–34.
[55] D. J. Chalmers, "Strong and weak emergence," in *The re-emergence of emergence*, P. Clayton and P. Davies, Eds. Oxford: Oxford University Press, 2006, pp. 244–256.
[56] E. Mayr, "Cause and Effect in Biology: Kinds of causes, predictability, and teleology are viewed by a practicing biologist," *Science*, vol. 134, no. 3489, pp. 1501–1506, 1961, doi: 10.1126/science.134.3489.1501.
[57] K. N. Laland, K. Sterelny, J. Odling-Smee, W. Hoppitt, and T. Uller, "Cause and Effect in Biology Revisited: Is Mayr's Proximate-Ultimate Dichotomy Still Useful?," *Science*, vol. 334, no. 6062, pp. 1512–1516, 2011, doi: 10.1126/science.1210879.
[58] M. Graves, "The Emergence of Transcendental Norms in Human Systems," *Zygon*, vol. 44, no. 3, pp. 501–532, 2009.
[59] T. W. Deacon, "Emergence: The Hole at the Wheel's Hub," in *The Re-Emergence of Emergence*, P. Clayton and P. Davies, Eds. Oxford: Oxford University Press, 2006, pp. 111–50.
[60] T. W. Deacon, "The Hierarchic logic of Emergence: Untangling the Interdependence of Evolution and Self-Organization," in *Evolution and learning: the Baldwin effect reconsidered*, B. H. Weber and D. J. Depew, Eds. Cambridge, MA: MIT Press, 2003, pp. 273–308.
[61] T. W. Deacon, *Incomplete Nature: How Mind Emerged from Matter*. New York: W.W. Norton, 2011.
[62] J. E. LeDoux, *Synaptic self : how our brains become who we are*. New York: Viking, 2002.
[63] T. W. Deacon, *The symbolic species: the co-evolution of language and the brain*. New York: W.W. Norton, 1997.
[64] W. T. Fitch, *The Evolution of Language*. Cambridge: Cambridge University Press, 2010.
[65] T. W. Deacon, "Shannon-Boltzmann-Darwin: Redefining information (Part II)," *Cognitive semiotics*, vol. 2, no. Supplement, pp. 169–196, 2008.
[66] M. Weisberg, *Simulation and similarity: using models to understand the world*. New York: Oxford University Press, 2013.
[67] L. Magnani and N. J. Nersessian, Eds., *Model-based reasoning: Science, technology, values*. New York: Kluwer Academic, 2002.
[68] L. Magnani and C. Casadio, Eds., *Model-based reasoning in science and technology*. Switzerland: Springer, 2016.
[69] G. Marcus, "Deep Learning: A Critical Appraisal," *arXiv:1801.00631 [cs, stat]*, Jan. 2018, Accessed: Dec. 22, 2020. [Online]. Available: http://arxiv.org/abs/1801.00631
[70] I. Apperly, *Mindreaders: the cognitive basis of "theory of mind."* New York: Psychology Press, 2010.
[71] N. Rabinowitz, F. Perbet, F. Song, C. Zhang, S. M. A. Eslami, and M. Botvinick, "Machine Theory of Mind," in *Proceedings of the 35th International Conference on Machine Learning*, Stockholmsmässan, Stockholm Sweden, 2018, vol. 80, pp. 4218–4227.
[72] S. A. Kauffman, *The origins of order: self-organization and selection in evolution*. New York: Oxford University Press, 1993.
[73] H. J. Ruskin and R. Walshe, "Emergent computing-introduction to the special theme," *ERCIM News*, vol. 64, pp. 24–25, Jan. 2006.
[74] J. McCarthy, "Recursive functions of symbolic expressions and their computation by machine, Part I," *Communications of the ACM*, vol. 3, no. 4, pp. 184–195, 1960.
[75] N. Wiener, *Cybernetics; or, Control and communication in the animal and the machine*, 2d ed. New York: M.I.T. Press, 1961.
[76] A. Newell and H. A. Simon, "Computer science as empirical inquiry: Symbols and search," *Communications of the ACM*, vol. 19, no. 3, pp. 113–126, 1976, doi: 10.1145/360018.360022.
[77] J. Firth, "A synopsis of linguistic theory 1930-1955," in *Special Volume of the Philological Society.*, Oxford: Oxford University Press, 1957.
[78] Z. Harris, *Mathematical Structures of Language*. New York: Interscience, 1968.
[79] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," 2013, pp. 3111–3119.
[80] M. G. Helander, *Handbook of human-computer interaction*. Amsterdam: Elsevier, 2014.
[81] A. Tolk, W. J. Wildman, F. L. Shults, and S. Y. Diallo, "Human Simulation as the Lingua Franca for Computational Social Sciences and Humanities: Potential and Pitfalls," *Journal of Cognition and Culture*, vol. 18, no. 5, pp. 462–482, 2018, doi: 10.1163/15685373-12340040.
[82] P. N. Edwards, "Infrastructure and modernity: Force, time, and social organization in the history of sociotechnical systems," in *Modernity and Technology*, T. J. Misa, P. Brey, and A. Feenberg, Eds. Cambridge, MA: MIT Press, 2003, pp. 185–226.
[83] B. Trist, *The Social Engagement of Social Science, Volume 2: A Tavistock Anthology–The Socio-Technical Perspective*, vol. 2. Philadelphia: University of Pennsylvania Press, 1990.
[84] S. Vallor, *Technology and the Virtues: A Philosophical Guide to a Future Worth Wanting*. New York: Oxford University Press, 2016. doi: 10.1093/acprof:oso/9780190498511.003.0001.
[85] A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi, "Fairness and Abstraction in Sociotechnical Systems," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, New York, Jan. 2019, pp. 59–68. doi: 10.1145/3287560.3287598.
[86] P. Tillich, *Dynamics of faith*. New York: Harper, 1956.
[87] Y. Congar, *Tradition and traditions; an historical and a theological essay*. New York: Macmillan, 1967.
[88] Y. Congar, *The meaning of tradition*. New York: Hawthorn Books, 1964.
[89] C. Taylor, *Sources of the Self: The Making of the Modern Identity*. Harvard University Press, 1989.
[90] D. Narváez and D. K. Lapsley, Eds., *Personality, identity, and character: explorations in moral psychology*. Cambridge: Cambridge University Press, 2009.
[91] D. L. Gelpi, *The gracing of human experience: rethinking the relationship between nature and grace*. Collegeville, Minn.: Liturgical Press, 2001.
[92] M. Graves, "Habits, Tendencies, and Habitus: The Embodied Soul's Dispositions of Mind, Body, and Person," *Habits in Mind: Integrating Theology, Philosophy, and the Cognitive Science of Virtue, Emotion, and Character Formation*. Brill, 2017.
[93] P. Hadot, *Plotinus, or, The simplicity of vision*. Chicago: University of Chicago Press, 1993.
[94] R. N. Bellah, *Religion in human evolution: from the Paleolithic to the Axial Age*. Cambridge, Mass.: Belknap Press of Harvard University Press, 2011.
[95] J. Royce, *The philosophy of loyalty*. New York: The Macmillan company, 1908.
[96] M. Graves, "Shared Moral and Spiritual Development Among Human Persons and Artificially Intelligent Agents," *Theology and Science*, vol. 15, no. 3, pp. 333–351, Jul. 2017, doi: 10.1080/14746700.2017.1335066.
[97] L. J. Walker and K. H. Hennig, "Differing conceptions of moral exemplarity: just, brave, and caring.," *Journal of personality and social psychology*, vol. 86, no. 4, p. 629, 2004.
[98] M. Graves, "AI Reading Theology: Promises and Perils," in *AI and IA: Utopia or Extinction?*, vol. 5, ATF Press, 2018.
[99] M. Graves, "Modeling Moral Values and Spiritual Commitments," in *Spiritualities of Human Enhancement and Artificial Intelligence: Setting the stage for conversations about Human Enhancement, Artificial Intelligence and Spirituality*, C. Hrynkow, Ed. Wilmington, Delaware: Vernon Press, 2019.
[100] A. Sen, *Commodities and Capabilities*. Amsterdam: North-Holland, 1985.
[101] M. C. Nussbaum, *Creating capabilities*. Cambridge, Mass.: Harvard University Press, 2011.
[102] M. Corrales, M. Fenwick, and N. Forgó, Eds., *Robotics, AI and the Future of Law*. Singapore: Springer Singapore, 2018. doi: 10.1007/978-981-13-2874-9_9.

Mark Graves

After earning his Ph.D. in computer science at University of Michigan, Mark Graves completed postdoctoral training in genomics and in moral psychology and additional graduate work in systematic and philosophical theology. He has twelve years industry experience in developing software, databases, and informatics and analytics solutions for healthcare, biotechnology and pharmaceutical research; and held adjunct and/or research positions at Baylor College of Medicine, Graduate Theological Union, Santa Clara University, University of California, Berkeley, Fuller Theological Seminary, California Institute of Technology, and University of Notre Dame. He published fifty technical and scholarly works in computer science, biology, psychology, and theology, including three books, and his current research focuses on using natural language processing (NLP) and moral psychology to build a foundation for AI ethics.