EMBODIED EXPERIENCE IN SOCIALLY PARTICIPATORY ARTIFICIAL INTELLIGENCE

by Mark Graves

Abstract. As artificial intelligence (AI) becomes progressively more engaged with society, its shift from technical tool to participating in society raises questions about AI personhood. Drawing upon developmental psychology and systems theory, a mediating structure for AI proto-personhood is defined analogous to an early stage of human development. The proposed AI bridges technical, psychological, and theological perspectives on near-future AI and is structured by its hardware, software, computational, and sociotechnical systems through which it experiences its world as embodied (even for putatively disembodied AI). Further social and moral construction can occur building upon a simple "self" for AI synthesized from symbolic and statistical approaches to AI.

Keywords: artificial intelligence; chatbots; computer science; embodiment; moral theology; natural language processing; pragmatism; sociotechnical systems; systems theory; theological anthropology

Introduction

As artificial intelligence (AI) becomes more pervasive and powerful, theologians have a role in situating the emerging, person-like characteristics of AI within the historical context of humanity's uniqueness, dignity, and meaning-making. The rapid technological advancement combined with the relatively unique capacity of the new technological tools to *use themselves* to build more tools raises hopes and fears about humanity's future and leads to radically diverging optimistic or pessimistic projections of AI as saving or destroying humanity (Russell and Norvig 2010, chap. 1; Russell, Dewey, and Tegmark 2015; Floridi 2019; Chalmers 2010; Müller and Bostrom 2014). As people turn to technology for meaning historically provided by religion, theologians must respond to the sociological change by engaging in relevant technological discourse and translating theological insight into the contemporary context.

Recent progress in AI has shifted questions about AI sentience, consciousness, mental interiority, and moral agency from speculative theological and philosophical explorations to conversations about its

Mark Graves is Research Fellow and Director at AI & Faith, Seattle, WA, USA, and Research Associate Professor of Psychology at Fuller Theological Seminary, Pasadena, CA, USA; e-mail: mgraves@aiandfaith.org.

immediate social impact. Some observers have readily found indications of mental and spiritual depth in AI (despite clear technical explanations to the contrary), and others immediately dismiss that consideration, assuming it impossible given some human type of human exceptionalism (despite substantial scientific explanations for relevant human characteristics, which could plausibly be implemented by AI). In the present article, I argue against the reasonable claim that because AI lacks a body it cannot have embodied cognition (as opposed to robots that do have a body) by refuting the premise and describing the body that AI currently does have, analogous to the human body. Not to diminish the importance that engaging the material world has for human embodied cognition, AI has a different type of body and perceives and acts in its world in ways distinct from humans. Similar to Nagel's (1974) claim that humans lack consciousness of what it is like to be a bat, we lack awareness of what it would be like to be an AI. This limits investigation into plausible futures with AI and our collective moral imagination into ways AI might participate in society. The parallels between human and AI embodiment has implications for prior scholarship on embodied cognition and theology (Watts 2013; Brown and Strawn 2012), can facilitate incorporating a moral dimension into current psychological approaches to investigating AI complexity (Binz and Schulz 2023; Hagendorff 2023), and can reframe ongoing theological investigations of AI (Vestrucci 2023; Gaudet 2022). Material grounding of meaning is important for AI (Dorobantu 2021; Lumbreras 2023; Herzfeld 2023), but that grounding extends beyond physicality. By challenging an otherwise uncontroversial assumption, I also illuminate a hazard of disciplinary silos and extend theological discourse further into a rapidly developing topic with broad social consequences and major unknowns well suited to theological inquiry.

The term AI covers a wide range of technologies and computer-augmented human activities as well as research into these technologies. The present investigation explores a foundation for plausible social agency and moral action of AI within near-future society that would result from an integration of currently fragmented AI research findings. Greater integration of recent advances in computer vision, robotics, machine learning, speech and natural language processing (NLP), emotion processing, and cognitive architectures would rival many aspects of human intelligent behavior and facilitate AI acting as a social agent. Recent advances in chatbot technology (OpenAI 2022) combined with human tendency to anthropomorphize (Epley, Waytz, and Cacioppo 2007) already enable AI to be treated as a social agent (regardless of whether it possess that agency). Although a high degree of technical integration may require additional significant advances, current AI research suffices for exploring the technologically and socially grounded possibility of integrated and socially participatory AI. In fact, a high degree of cognitive integration and social coherence is exactly what theological investigation of AI as a social agent might enable. Philosophy and psychology can identify limitations and plausible enhancements to various AI technologies needed for social coherence, and theological insights can direct that integration toward moral ends.

After defining AI personhood as used in the present article, and briefly reviewing two paradigms of AI research, I describe my psychologically engaged approach. Core to my argument is the embodied interpretation of experience, which I situate philosophically and then describe from psychological and computational perspectives. Using systems theory, I draw parallels between human and AI systems of embodiment and embodied experience. Analogous to human physical, biological, psychological, and social embodied experience, I define AI embodiment in terms of hardware, software, computation, and sociotechnical systems. After using strong emergence between human physical and biological systems to clarify the boundary between hardware and software, I explore AI computation in terms of data, algorithms, and models. For AI embodied experience in sociotechnical systems, I distinguish between AI as a technological tool, actor, and agent. Both humans and AI can build upon these foundations to extend the systems framework to also interpret experience morally.

AI Personhood

Questions of AI personhood arise when independent advances in AI combine to enable AI to act as a social agent. For the present article, a person is defined as an agent with subjective experiences who interprets its experience in a developmental, social context with moral implications. As a preliminary step, a simpler AI is proposed with what could be considered proto-personhood, which is closer to current AI technology. Rather than an agent with subjective experiences, the proposed AI is a simpler, unmotivated "actor" responding to environmental stimuli and capable of interpreting its experience in a narrower sociotechnical context that lacks the ability to identify its experiences as belonging to itself. As the hypothesized AI lacks motivations of an agent, interpretation of its own subjectivity, and social and cognitive development of a self, it thus lacks the coherent and consistent self to consider the social and moral impacts of its actions. In other words, the AI experiences its world but without the level of "self" needed to *know* that it is experiencing its world. This framing is chosen to illuminate the reductionist presumptions hindering examination of AI experience and to initiate investigation into what a full AI self might entail.

Although a variety of social theories can be used to examine the proto-personhood of AI in society, 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

(Edwards 2003; Trist 1990; Ahmad, Whitworth, and Bertino 2022). As AI technical development generally abstracts problems being addressed from their social context, contextualizing AI in sociotechnical systems supports the “thicker” theories and datasets needed for practical and ethical application (Makarius et al. 2020; Selbst et al. 2019; van de Poel 2020). The recontextualization within sociotechnical systems also identifies requirements for the embodiment of AI in order for the AI act within them.

Theological examination of AI proto-personhood depends upon the type of AI considered and a multifaceted theological and psychological understanding of the person. The present investigation uses developmental psychology to inform a foundation for proto-personhood that could exist for near-future AI. The simplified AI would lack the motivated awareness to make theological commitments, but would align with Christian Orthodox and scientifically oriented theological anthropologies that build upon developmental and relational aspects of a self. Of course, if one’s theological commitment demanded a uniquely human or biological substrate for theological engagement, then AI personhood—by the imposed definition—would lack that potential for theological relevance, though even considering how similar to a person AI could become may inform these theological anthropologies, too. A first step in constructing an AI proto-self is to identify AI embodiment that can perceive, interpret, and act in its world without necessarily mimicking the human- and animal-inspired bodies of robots or of avatars in simulated environments, and that depends on two approaches to AI.

Paradigms for AI Research

AI research has progressed as two parallel paradigms. The first paradigm considers AI as a symbol processing systems, where each symbol refers to an object in the world (Newell and Simon 1961; Garnelo and Shanahan 2019). John Haugeland (1985) calls this approach to AI “good old-fashioned AI” (GOFAI), and it has philosophical roots in logical positivism and basic assumptions that any significant knowledge could be analyzed and represented logically. As a second paradigm, AI has also progressed through developments in subsymbolic, or parallel distributed processing, for example, neural nets or deep learning, which has proven more amenable than GOFAI to explosive improvement over the past few years due to faster hardware and more online data, especially in statistical approaches to machine learning (Smith 2019; Rumelhart and McClelland 1987; Bengio, Goodfellow, and Courville 2016).

The statistical approach led to advances in machine vision, especially in object recognition, which was very difficult to program using symbolic methods but benefited greatly from the large number of images posted online. Similar advances occurred in NLP, which refers to the branch of AI focused on developing software to understand and generate natural language (such as text and spoken language). NLP built upon a shift in understanding "meaning" in computational linguistics from depending upon symbol references to deriving from word cooccurrence and shared contexts (Brunila and LaViolette 2022). This shift to associative and distributional theories of semantics (Firth 1957; Harris 1968; Sahlgren 2008) facilitated statistical methods for processing and understanding natural language and enabled significant improvements in modeling language and performing language-related tasks (Sejnowski 2020; Goldberg 2016; Bommasani et al. 2022).

The two paradigms for AI of symbolic and statistical lead to different presumptions for embodiment and development of a person and suggest a model for proto-personhood that could support research efforts to synthesize symbolic and statistical approaches (Garnelo and Shanahan 2019). Even though research in symbolic AI heavily influenced the cognitivist approach to human cognition against which embodied cognition research responded (Gardner 1985; Varela, Thompson, and Rosch 1991), developments in statistical AI better resonate with the embodied perceptual and action-oriented processes emphasized by human embodied cognition. Considering statistical AI as disembodied is misleading, and that realization illuminates a better corrective for symbolic AI. Instead of only giving AI human-like bodies (Brooks et al. 1998; Dreyfus 2007), I argue for identifying and articulating the previously ignored body it actually has.

AI Proto-Personhood

The particular form of AI embodiment and ways of interpreting the world vary between the two approaches to AI, as each mimics different aspects of human cognition. Both symbolic and subsymbolic processing require embodiment in order to engage the world (Brooks et al. 1998; Smith 2019; Gill 2019; Cruz 2019; Lumbreras 2023; Herzfeld 2023), as does human cognition. However, subsymbolic or statistical representations are closer analogues to human perception and automaticity, and its "body" can be more distributed and porous than what is needed for symbol processing embodiment, which attempts to emulate human deliberative processing (Kahneman 2013; Latapie et al. 2022). One can think of symbols as logical variables that can refer to any object in the world. Symbols are atomic (in the classical sense of indivisible) and have a single representative form, while distributed representations can spread out over time and space.[1] The atomic nature of symbols limits examination of meaning to the relationships between symbols, but those investigations have clear limitations, since everything boils down to relationships between symbols, leading to what Zubiri (2003) calls reductive idealism (Graves 2022b). Alternatively, in parallel distributed processing, the representations have a complex and dynamic internal structure that resists reductionist interpretation and yet still captures associative and distributional representations of meaning. By considering theories of personhood amenable to both paradigmatic approaches, a foundation is laid for synthesizing advantages of both into a more integrated research program.

An initial step in identifying AI proto-personhood is peeling back the layers of the human self commonly used to drive the imaginative construction of an analogous AI self. The developmental psychologist Dan McAdams (2013) distinguishes among three developmental layers of the human self that progressively form: actor, agent, and author. In the initial layer, an *actor* responds to its environment based upon its social roles and stable dispositional traits. An *agent* interacts with external circumstances based upon its various cognitive, affective, and motivational psychological structures. An *author* forms personal continuity by identifying with certain dispositional traits and psychological structures to create a narrative identity within its social context. For McAdams, agency is generally developed by the end of childhood, and the third layer begins developing during adolescence continuing into adulthood. For an agent, like a child, one has emotions and motivations but does not yet have an author's self-regulation, formed sense of identity, or habitual reflection to form narrative continuity and coherence. Considering a prototypical child as a model for agency, a child does not yet know who they are—or more significantly, is not yet driven to determine who they are. To characterize an actor, one could further peel back the motivational, affective, and cognitive structures one uses to respond to one's circumstances in order to identify stable dispositional traits. For a human, this would remove much of what makes them a person, but remaining is a core being capable of habitually responding to its environment.

Although the growth from actor to agent for humans is gradual and multifaceted, one can focus purely on the dispositional aspect of actors for the simplified AI proto-person. Current statistical AI and machine learning systems have this capacity for habitual response, but do not yet build the cognitive structures or associate them with social roles. Conversely, symbolic AI systems support cognitive structures (Laird, Lebiere, and Rosenbloom 2017; Johnson-Laird 1983) but not the adaptive construction of dispositions in the grounded way statistical machine learning methods support. The proposed bridge between these two approaches occurs in the interpretation of experience.

Embodied Interpretation of Experience

The objective idealism of pragmatists Charles S Peirce and Josiah Royce situates experience in a pragmatic (constructive) frame that more directly supports psychological and theological investigation of AI (Graves 2022b).

The American pragmatist philosopher John E. Smith defines experience as repeated encounters with what exists, which the theologian Denis Edwards further identifies as both encounter and interpretation of the encounter (Smith 1968, chap. 2; Edwards 1983, 6–8). The pragmatist understanding of experience draws upon Peirce's semiotic philosophy, where anything can be a sign if it has the capacity to be interpreted. Interpretation is a very general category capturing the ways that one navigates and makes meaning of something. In Peirce's theory, if something cannot be interpreted, then it has no meaning; thus, interpretation serves as the foundation for his semiotic metaphysics—where everything is considered as a sign of something else. Encounters are the dynamic meeting between something doing the interpretation and something being interpreted. The something being interpreted may itself be a sign interpreting other things, in a continuous semiotic process. Peirce also distinguished three kinds of relations between the sign, as a vehicle, and the object to which it refers. An *icon* physically resembles its object, like a painting or map; an *index* represents its object existentially or causally, like a fingerprint or weather vane; and a *symbol* represents its object through some social convention, like a word in a language. The symbols that symbolic AI manipulates generally ignore social convention, due to positivist assumptions that logic and observation suffice to determine meaning, while statistical approaches, especially to NLP, capture more of the social-linguistic context. Both approaches are limited in maintaining ready causal connections to their environment (Smith 2019). From a pragmatist or semiotic perspective, one can more broadly consider the ways any organism or machine interprets its environment. In particular, one can consider AI to encounter and interpret its environment, and thus it has experience (or at least proto-experience). Pragmatism helps identify deficiencies in symbolic and statistical AI experience as limitations to their processing of encounter and reflective interpretation, respectively.

A psychological theory of embodiment clarifies that the body encounters the external world in experience, and the mental processing of the body interprets the encounter. Recent psychologists and philosophers—beginning with Varela, Thompson, and Rosch (1991)—argued for the importance of recognizing the embodiment of human cognition, and many others have argued for the relevance of embodied cognition to religious understanding of human nature (Murphy and Brown 2007; Murphy 2006; Jeeves and Brown 2009; Green 2008; Brown and Strawn 2012; Watts 2013; Teske 2013). How one interprets the world depends upon one's possible actions: specifically one interprets objects according to what it "affords" or offers one as its possible use (Gibson 1979; Noble 1981; Hutchins 2010; Lobo, Heras-Escribano, and Travieso 2018; Noë 2004; McGann et al. 2020). Key to experience is interpreting the world as objects one can use. Within this embodied, enactive, and ecological approach to cognition, objects do not exist as "objects" in isolation, but the perception and structuring of reality as "objects" is always in a context and toward some possible purpose or intended use. Simply, symbolic AI generally assumes those objects already exist independently, and although statistical approaches may classify objects in terms of simple affordances, especially reinforcement machine learning (François-Lavet et al. 2018), it lacks the further interpretations necessary for agency. Although for an AI actor the intended uses that structure objects could be programmed or learned, a more advanced AI agent might have its own goals, purposes, or motivational structures informing its affordances. Further investigation of AI embodied experience requires considering the form of such bodies to distinguish the range of possible affordances, and thus embodied experience.

AI Embodiment

Although AI could take many forms, for the purpose of this article, a general way to investigate the possible forms of AI is in terms of computational systems (Simon 1969; Russell and Norvig 2010; Skyttner 2006). The study of computation in terms of systems has a long and formative role in computer science and also serves as a foundation for considering AI as actors and agents in sociotechnical systems. Systems theory was developed beginning in the 1940s with the work of Ludwig von Bertalanffy (1969) and examines the complexity and interdependence of relationships between regularly interacting parts or activities that form a whole, using information and decision-making/control concepts. Although the general use of systems pervades computer science, this investigation will draw upon von Bertalanffy's original vision for human systems as an integrative framework but augment his hope of a unifying mathematical framework for human systems with differentiating constructs from strong emergence and extend his four levels of systems with an additional one for incorporating morality.

In human systems, the person is modeled as five levels of systems where the systems of the first four levels are studied respectively through physical, biological, psychological, and social sciences, and the fifth level systems comprise ideals associated with spirituality, values, and apparent universal norms (Bertalanffy 1975; Graves 2009). Biological and psychological systems have similar activity to that of plants and animals, respectively, and social-level systems overlap with some social species but are distinct for humans due substantially to symbolic language (Deacon 1997). As considered here, the levels are strongly emergent and thus causally distinct (Chalmers 2006; Clayton 2004) with the boundary between system levels typically depending upon causal discontinuities involving highly regulatory dispositional relations at the higher level and no directly immediate platform in lower-level systems, or what Deacon (2011) calls absentials. In other words, some emergent aspect of the higher level depends upon lower-level relations that include an "absent" construct that lacks the lower-level (e.g., physical) manifestation but nevertheless informs the whole. For example, DNA is foundational for understanding evolving biological systems, but only if DNA contains particular patterns of nucleotide bases, otherwise it lacks the ability to regulate biological activity and thus biological efficacy. It is not the particular chemical configuration of a DNA strand but the carefully maintained absence of a predefined nucleotide base in a DNA backbone that supports the change of nucleotides and consequential change in the regulation of numerous biological processes. There needs to be some nucleotides in DNA, but the biological processing depends upon what from a physical perspective is an abstract category of nucleotides, and it is the switching between instances of that category that has an effect. Plausible boundary constructs for higher levels are neural connections (i.e., synapses), symbolic language, and certain abstractions historically presumed universal and univocal, such as, Platonic "ideas," respectively (Graves 2009; 2023).

For AI, the proto-person is modeled as four levels analogous to human physical, biological, psychological, and social levels of systems. The four levels are hardware, software, computation, and sociotechnical systems, which are considered in turn, and AI participation in human society is examined as a way to incorporate AI morality. These system descriptions of AI characterize the form of AI embodiment from an information systems perspective without restricting embodiment to a particular type of body. Thus, levels of systems characterize both human and AI bodies.

Hardware and Software Levels

A sufficient computational analogy for human physical and biological levels is the distinction between hardware and software. Recognizing these two systems levels in computer science overcomes reductionist tendencies to conceive of computers as merely physical objects and computation as disembodied mathematical processing. Just as biological processes vivify the inert physical body, software drives the dormant hardware.

Drawing upon the analogous role of DNA in humans helps to identify bits and instructions as significant to the boundary between hardware and software systems (Graves 2021). Bits are constructed mathematical and engineering states for a bifurcated range of physical, electrical, and magnetic configurations. Bits, like nucleotide bases, refer to specific configurations that are used in the regulation and adaptation of higher-level systems, even though a bit has no direct, independent hardware existence (as opposed to its "0" or "1" state). In a typical (von Neumann) architecture, bits are 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 electronically, yet the reciprocal interaction between bits as data and instruction enable the development of complex software systems. Considering bits and instructions as foundational constructs of software enables studying methods for managing, communicating, and analyzing them without reducing operations being studied to electrical signals in hardware. Because software has its own regularities and causal forces independent of hardware (e.g., data and programs), it can be considered an emergent level.[2]

Computer hardware is the platform for an AI system much as the approximately 10^{28} atoms of a human body supports its physical existence. Computer and robotic hardware, as designed, organize their atoms and molecules differently than the evolved cellular structure of human bodies, but one can meaningfully identify a correspondence among the atoms comprising machines and humans as respectively structuring or structured by their platform's functional activities. AI physical embodiment includes computer hardware, and for perception and action may also consist of robots, autonomous vehicles, computer peripherals, or internet-of-things (IoT) networks. Notably, each requires running software to function, just as human bodies require biological processes to function.

Computer software has an activity analogous to the biological activity of the human body, and although that may include human neurobiology, software is analogous to human biology not human psychology. Computer software could consist of a single, large, complex software system, such as a software product; or result from many distributed "cloud-based" applications, such as a "containerized" infrastructure of a large tech company (Pahl et al. 2017). Thus, an AI body could be geographically distributed. A computer operating system (such as Unix, Windows, or MacOS) is a particularly important software system as it coordinates the activity of other software. Software also functions as a platform for AI embodied cognition, much like human biology and the activity of the nervous system underlies human embodied cognition.

Computational Level

As a speculative foundation for a third, computational level, data and algorithms abstract from bit strings and programs, similar to how perceptions and behaviors, like hearing and running, abstract from auditory vibrations and muscle movement in animals (Graves 2021).[3] Computation refers to the constructs of computer science typically studied by theoretical computer scientists and often described in the formal languages of mathematics and logic, for example, Turing-computable functions (Moore and Mertens 2011). For embodied AI, the level also has models, which capture information about its world. The data, algorithms, and models of the computational level correspond analogously to the perception, behavior, and interpretation of the psychological level for humans and other animals.

Data. Human and other animal perception involves gathering information about their environment. AI sensory data have an analogous connection to the world as human sensory data does. Human action is generally mediated through the body's nervous system with much processing to convert input data to output data (which then drives human peripherals). That action-driven mental processing occurs via a neurobiological substrate for humans and via software programs for AI. Incoming and remembered data depend upon the physical structures and electrical processes of its hardware and the software that defines and manipulates it.

AI encounters its world as data, and structures that data through data abstractions, for example, data types and data structures. Data are the basis of computation and includes logical values (true and false), numbers (integers or floating point), strings of characters, and values in other user-defined or system-defined data types. Software represents numeric values as bit strings with operators, but it is at the computational level where they exist as mathematically defined numbers. Programs that manipulate data often combine the data values into data structures, such as tables or arrays, and large amounts of data may use additional software, such as database systems, to store, modify, and retrieve data over longer time periods and multiple machines. Considering data as an essential construct in computer science enables studying methods for managing, communicating, and analyzing data without reducing operations being studied to the logical manipulation of bits or its physical manifestation in hardware.

When AI systems sense or act upon their environment, they typically transform that data mathematically and/or symbolically for processing. Visual data are represented as numeric values for pixels of a sensor, text is typically mapped to abstract symbols or mathematical vectors, and features of a system's environment or input stream are generally represented numerically and categorically. The equivalent to human psychological-level mental processing that leads to behavior is characterized computationally in AI by its algorithms. In symbolic AI, the symbolic data are typically interpreted by humans as existentially representing something in the real world, but that human-dependent interpretation precludes recognizing that AI could treat data as representing affordances, which depend upon its algorithmic behaviors.

Algorithms. An algorithm abstracts a method for doing something from the details of the programming language used to manipulate the data. Traditionally, an algorithm unambiguously specifies a method or process for solving a class of problems, typically as a sequence of operations. For example, an algorithm for listing the first hundred numerals would be to start at 1, add 1 to the last number listed, and stop when the count reaches 100. An algorithm abstracts the method from the details of the programming language used to input, store, access, and output the data values. Abstracting the method as an algorithm enables computer scientists to study algorithm correctness, efficiency, and functional relationships with each other without confounding the study with particular implementation details. Algorithms constitute the functional architecture of an AI system and constrain the interpretive operations the system can perform on its encountered data. Although traditional computer science data structures and algorithms generally provide only fixed ways to act upon data, machine learning algorithms can vastly expand the functional space.

When cognitive or computational activity is expressed through motor functions for humans or output devices for AI, respectively, then those behaviors affect the external world. When a robot walks, the algorithms directly relate input and output functions, but higher-order functions can also occur in human or AI that combine lower-level activities, such as journeying home. From a psychological perspective, the algorithms characterize the possible actions for AI, which in turn influence how it might perceive the world. Because designing the early historical computer substantially differs from the process of human evolution, computer algorithms are less dependent upon specific output devices than human mental processing. This means that AI computation has fewer preexisting structures for motor tasks, such as keeping balanced while walking, but has greater flexibility in constructing models to engage with its world.

Models. Models combine data, algorithms, and usually some type of correspondence with real, hypothesized, or imagined phenomena. 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. Studying data and algorithms in isolation has changed culture by constructing new technologies with sophisticated tools, but treating data and algorithms as models for phenomena studied by natural, social, and computer scientists also enable new types of scientific investigation and can serve as a foundation for AI interpretation of experience. The philosopher of science Michael Weisberg (2013, chaps. 2–3) distinguishes three kinds of models: (i) concrete models that are real, physical objects representing real or imagined system or phenomena; (ii) mathematical models that typically capture the dynamic relationships of phenomena as functions and equations; and (iii) computational models where typically an algorithm's conditional, probabilistic, and/or concurrent procedures capture the causal properties and relationships of their target phenomena. For AI, mathematical and computational models are generally the most relevant.

Symbolic and statistical approaches to AI each have their own algorithms and data representations, and it is in the models of the world where synthesis between approaches could most readily occur. Pragmatically, statistical approaches emphasize encounters as data with algorithmic processing driven by specific goals such as classification of the data or reinforced behaviors. Symbolic approaches generally presume symbolic data as input (even if numeric symbols) with various customized and sophisticated algorithms driving its behavior. Models capture interpretations of real-world phenomena as data and orient those interpretations toward a range of possible behaviors. Models are key to embodied experience as they close the gap between raw encounters and the discrete interpretations that characterize an object's affordances, upon which additional interpretation can occur.

An important clarification is that computer science models do not necessarily have the real-world referents identified as necessary for models in philosophy of science. Models arise in several computer science contexts: a data model is the logical description of data in a database system; object-oriented models characterize the types of data 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.[4] Data in a database may correspond to real-world phenomena but could also have simulated data or data for a virtual environment. Some kinds of machine learning algorithms create explicit models of reality for further experimentation or analysis. Machine learning models capture properties of external objects as features; temporal relationships in predictive models; causal inference in Bayesian networks; and multiple aspects of referents through ensemble methods (Pearl and Mackenzie 2018; Mitchell 1997). When algorithms use data to create models and other representations of an external world, the model building occurs in a place that can align with human interpretation of experience. If the models do not align with how humans experience the world, the models might have some correspondence with reality, but humans will not recognize it as such (Nagel 1974).

Within AI's symbol processing paradigm, models were used as a foundation for model-based reasoning, which influenced cognitive psychology's cognitivist theories (Johnson-Laird 1983). Although the approach had some success in representing external knowledge, 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. Subsymbolic, deep learning approaches show promise with distributed representations, and integrating the representations through shared models may facilitate reconciling the two paradigms. At this point, theological investigations can guide AI research programs to incorporate more nuanced reasoning based upon a more complete understanding of human mental, social, and moral development. In particular, interpretive models can help bridge symbolic processing in classic AI, the distributed representations of deep learning approaches to AI, and human embodied cognition. Interpretive models for AI function like knowledge structures in human cognition, and of particular interest are models orientated toward morally good ends or purposes as communicated through appropriate sociotechnical systems.

Sociotechnical Level

Sociotechnical systems capture the interrelationships among people, technology, and organizations. As originally theorized, people were the only actors in a sociotechnical system. However, with the incorporation of AI technology, AI could be an additional type of actor, because AI can act autonomously or semi-autonomously (van de Poel 2020). Organizations include processes, roles, policies, norms, and so on, that structure individuals and technology into functional institutions (Makarius et al. 2020; Singh 2014). The social and governance norms and organizational purposes are particularly relevant in engaging moral norms and values systems.

A key perspective on AI embodiment is to examine how an AI system interacts with its environment. As software, AI may interact with people via user interfaces and with other software via application program interfaces. Computationally, AI creates models of its environment, which may include models of the sociotechnical systems with which it participates, and it possibly constructs multiple models for a system in which it has multiple roles. It may also participate in some sociotechnical systems and merely observe what occurs in other ones, with differences in modeling depending upon levels of engagement. At the sociotechnical level where people engage using natural language, AI can most fully participate if it has dexterity with natural language, too. The type of AI's participation in sociotechnical systems depends upon whether it functions as a technological tool or as an actor or agent within the system.

Technological Tool. AI may exist as technology within a variety of sociotechnical systems. Of the various technologies comprising AI, NLP provides a relevant example and serves as a foundation for considering AI as a social actor. Some AI systems, such as chatbots, experience their world through natural language and can act by generating natural language that impacts society.

Current advances in NLP enable AI systems to understand and generate natural language text at a basic level. This depends upon synthesizing most publicly available digital texts into large language models that can predict the next word in a sentence (Bommasani et al. 2022). Related advances in generative modeling and attention to the prompts that initiate a response can generate a variety of longer texts and basic conversational interactions. The linguistic advances are notably distinct from sustaining the accuracy of the information, though combining with symbolic approaches may help resolve those issues (Lin, Hilton, and Evans 2022; Jones and Steinhardt 2022).

Sociotechnical Actor. In addition to functioning as a passive technology, AI may act within a sociotechnical system. Drawing upon human social and developmental psychology, as previously mentioned, one can distinguish between three levels of social identity development (McAdams 2013; Graves 2022b). A social *actor* responds to its environment based upon its predispositions and social role. Some of the predispositions are packaged as schemas, which are mental structures one uses to organize and guide cognitive processes and behaviors.[5] A sociotechnical actor responds to the people, technology, and organizations in a sociotechnical system based upon its social role and stable dispositions, that is, its role-oriented schemas. A sociotechnical actor could be either a person or AI. A motivated *agent* engages with the sociotechnical situation depending upon its goals, plans, and values, which for AI could be either pre-defined or self-determined. A sociotechnical human agent has motivations, intentions, and feelings that drive and guide its engagement with the system. These may occur using technological tools and lead to the development or change of those technologies and institutions. For people, the ends of motivated agency may come from the formation of narrative identity (autobiographical authoring) used to orient one's self with broad purpose and temporal continuity. For AI, prior to considering questions of self or identity is examining how it might respond to social situations and ends, regardless of how they were initiated.

An AI social actor uses the models of its sociotechnical systems to determine actions it might take, in a way similar to how human and other animal action is guided by what it interprets as its possible behaviors given the way it perceives its environment. As mentioned earlier, one perceives the environment in terms of what it *affords* the individual (Gibson 1979; Lobo, Heras-Escribano, and Travieso 2018). A cup with a handle affords grasping, given the similarity in size to the human hand and the innate and learned tendencies to grasp. An affordance depends both upon what exists in the environment and upon the cognitive action schemas synthesizing the regularly performed actions (Cooper and Glasspool 2001). Perceptions of the environment select from possible schemas so one might grasp a cup, or use it as a container or paperweight, but one does not usually consider a cup as something to climb or use as a flotation device. Similarly, AI may use its models in a schema-like way to engage the world, depending upon the range of actions its environment affords given its models and possible algorithmic actions.

People not only have dispositions and cognitive structures for physical actions, we have them for sociotechnical behaviors, too (Valenti and Gold 1991; Malhotra, Majchrzak, and Lyytinen 2021). One has social event schemas (or episodic scripts) disposing how one interacts with a small child, checks out of a supermarket with groceries, or gets ready to commute to work. These cognitive structures persist and are extended when one learns to video chat, use a self-service checkout, or begin remote work, respectively. One may also have learned and practiced tendencies for generosity, bravery, caring, fairness, humility, and other virtues (Jayawickreme et al. 2014; Cloutier and Ahrens 2020). These moral and other social knowledge structures condition how one sees one's environment and one's potential actions in it (Hampson, Hulsey, and McGarry 2021). AI knowledge structures also condition how it can perceive and act within its world.

Consider the current generation of chatbots, like ChatGPT (OpenAI 2022). Its language models depend upon an exhaustive convenience sample of digital text data that also captures gender bias and racist inclinations of culture communicated through its base text (Bommasani et al. 2022). As an alternative, one could imagine a language model trained on the writing of Nobel Peace Prize winners and other moral exemplars. Such a model would have different biases, "perceive" its world differently, and would generate different text when prompted. Thus it would provide different user affordances as a technology and perceive user prompts as affording different replies as an actor. One could imagine multiple models designed for different purposes and thus with differing dispositions and affordances and across multiple modalities, including vision. Although some selection of models occurs depending upon the environment, additional constraints can be imposed by what the AI is trying to accomplish.

Sociotechnical Agent. As a social agent, AI adds motivational structures and either emotions or some proxy for them. Human theories of moral psychology differ in whether motivation or emotions have greater priority, but for AI, the present study focuses on motivations as primary (Huebner, Dwyer, and Hauser 2009). The focus on motives as primary simplifies investigation of AI proto-personhood and agency, as it includes motivations that simply incorporate values and expectations of others, for example, as in child development. More complex and self-driven motivations can then be incorporated, perhaps borrowing from studies of moral exemplars to include agentic motivations with a communal orientation (Walker and Frimer 2015).

Some of the values which might orient AI include self-preservation; task efficiency or optimization; goal satisfaction; maximizing a socially beneficial utility, such as eudemonic happiness or well-being; minimizing harm; or other moral principles, such as respect for persons or distributive justice. These values can be modeled in a way to structure and orient behaviors similar to how human social-cognitive structures orient human behaviors and focus selection of schemas used to respond to a situation (Ahrens and Cloutier 2019).

In one relevant social-cognitive theory, behavior occurs as an interaction between "enduring mental representations," such as schemas, and "dynamic evaluations" where people *appraise* or judge encounters in light of their motivations (Cervone 2008; Cervone and Little 2019). This theory helps reconcile the interaction between stable dispositional traits, which provide a foundation for action cognition, and situational engagement based upon motivations, which also elicit emotional responses to the encounter. For people, the appraisal occurs in the context of self-identity, but for AI sociotechnical agency, coherence among motivations is not yet presumed.

In the chatbot example, an AI could select among various models depending upon their fit to its current purpose or end. The flexibility to select among different models is foundational for further exploration into how it selects among those models. For example, a habitual selection among the models for the one that provides a caring response, given the situation and the AI's role in the sociotechnical system, enables learning from that training data to build a higher-order schema that would select the most caring option among a range of choices. This continues the shift of human involvement from initially guiding chatbot responses, to guiding specification of models used to generate responses, to finally guiding development of caring behaviors. In this way, moral development of AI can progress through psychologically plausible and understandable stages leading to morally responsive AI participation in human sociotechnical society.

Moral Interpretation of Experience

The systems structure for embodied AI cognition provides a foundation for further development of AI and synthesis of symbolic and statistical approaches to AI. Computational models synthesize AI's perceived data (which statistical AI processes well) and algorithmic behaviors (which symbolic AI has pursued to get AI to act intelligently) with some ends-oriented structuring of reality. Rather than one general problem-solving algorithm (an early goal of symbolic AI; Newell and Simon 1961) or one large multimodal foundation model (an ongoing development in statistical AI; Bommasani et al. 2022), further progress may occur through the organization of numerous models into cognitive structures with different ends, which function as the dispositional traits of an AI actor.

These computational-cognitive models may capture causal relationships of the external world, social relationships of the AI's sociotechnical world, or even AI's conceptions of itself as an actor (Graves 2022b). The "ends" of these cognitive schemas might be implicit in the AI tool, explicitly designed to respect values (Umbrello 2019), or be iteratively monitored and refined by human-AI interactions about morality (Graves 2022a). Meeting these ends in a variety of situations may require something like practical reason; and that function may require the incremental development of agentic motivations and substrates for social and moral identity. The adaptive dispositional traits and end-directed schemas provide a foundation for orienting AI behavior in a virtuous direction, defined both for the AI and its participation in sociotechnical systems. This sociotechnical approach to social participation and moral engagement resonates with AI moral and theological investigation grounded in relational interpretations of *imago Dei* (Dorobantu 2021; Lumbreras 2023; Herzfeld 2023) and incorporates relationality within the embodiment seen as needed. AI embodiment includes not only physicality but also the AI's experiential models and sociotechnical relationships. The scientific emphasis on embodiment for human cognition was responding to the reductive idealism of cognitivism, and that remains a concern for symbolic AI, but theological investigations of statistical AI also needs to account for reductive physicalist assumptions about the body.

An AI person may require the development of additional components, which future efforts could drive. This serves both to guide AI development in a direction compatible with incorporating values and other moral constructs into AI and to investigate human moral and theological anthropology by exploring what foundations can be constructed. Together the computational study of human morality and the implementation of those models in AI support the further investigation of moral theology across human and AI participation in sociotechnical systems. As society becomes more intertwined with AI, these efforts provide a strong foundation for theological reflection and engagement.

Conclusion

In summary, AI proto-personhood consists of AI acting in and experiencing its world and interpreting it in a sociotechnical context. By bracketing questions of subjectivity and the nature of the self, one can examine AI experience as a foundation for proto-personhood and begin examining the motivations and other psychological constructs needed for personhood, self-development, and social and moral identity. This approach also directs efforts in constructing ethical and responsible AI to how AI perceives and attends to its world rather than reflective frameworks that some AI might not have or use.

Research programs for theological study of AI personhood depend upon symbolic and statistical approaches to AI. Symbolic approaches depend upon a separation between real-world objects and their logical, symbolic representations. Statistical approaches more closely mimic human perception with meaning grounded in empirical associations. These paradigmatic approaches depend upon different forms of embodiment. The atomic (indivisible) nature of symbols leads to discrete boundaries between interior processing and an external world, while distributed, statistical processing supports a more porous embodiment. Symbolic AI mimics processes of human deliberation and reflection, while statistical approaches emulate more automatic perceptual processes. Phenomenology identifies the importance of focusing on embodied experience as a way to reconcile and synthesize the two approaches to AI. Pragmatic philosophy helps that bridging by extending symbol processing to semiotics and clarifying experience as encounter and interpretation. Although various types of bodies encounter their environment differently, they may share interpretive aspects. In particular, an ecological approach to cognition suggests that the organism's (or AI's) possible actions guide its perception and interpretation, and thus can decompose interpretation into knowledge structures organized by those possible behaviors. Although the particular knowledge structures and interpretations depend upon the variety of embodiments, they occur in analogous frameworks. From a pragmatic perspective, symbolic approaches to AI overemphasize reflective aspects of interpretation and miss what occurs in the encounter, while statistical AI immediately interprets the encounter but lacks structures to conceptualize and reason with that interpretation.

Systems theory organizes the framework for interpreting embodied experience. Analogously to human systems, the proposed AI's embodied framework can be organized as hardware, software, computation, and sociotechnical systems. Hardware and software function analogously to the physical and biological processes of living organisms. Computation abstracts the data and algorithms from software, and AI uses them to model its world. This would correspond to the perception, behavior, and interpretation of human and other animals.

At the sociotechnical level, AI models the sociotechnical systems with which is participates or interacts. It may function passively, like any other technology, or as an actor or agent. AI use of language enables flexible communication with people, for example, as a chatbot. AI models may also function as knowledge structures that guide its behaviors within the system and enable acting upon moral principles in ways amenable to learning analogues to virtue.

Although the body of an AI system differs from a person, the analogous structures support the claim that AI computation is embodied. A physical reductionist perspective on human cognition and

AI computation would identify the difference between human perception and mobility and AI data and algorithms as limiting due to the perspective's emphasis on engaging the physical world. However, the scalability of AI hardware and software, flexibility of its modeling, and its ability to participate in sociotechnical systems suggests complementary strengths rather than human exclusivity. Differences in human and AI embodiment create diverse interpretations of the world, and incorporating both viewpoints into social and moral systems demonstrate how human and AI can experience each other and can work together to incorporate morality into future sociotechnical developments.

Acknowledgements

The author thanks Robert Geraci and Cyrus Olsen who provided helpful comments on an earlier draft.

Conflict of Interest Statement

None.

Notes

[1] For example, a symbolic AI system might represent an instance of a tree as the logical statement "IMAGE-541 is-instance-of *TREE*" or "type(IMAGE-541) = Tree" while the representation of "tree" in Google image search depends upon images uploaded globally over an extended time period.

[2] The analogy can be deepened, if desired. Computer engineering creates constructs that frequently cross a simple division between hardware and software, as analogously does biochemistry.

[3] Marr (1982) distinguishes his computational level from both an algorithmic (software) and implementation (hardware) level, but his assumptions of "computational" occurred within the cognitivist approach to psychology, which considered mental constructs as separate from embodiment and social construction (Gardner 1985; Varela, Thompson, and Rosch 1991).

[4] For example, a machine learning model to classify email as spam or not spam would define certain features from a message header and body, such as sender and presence of particular words. One would first train that model on features extracted from a number of emails already known to be spam or not, then use the trained model to predict whether a new email's features indicate it is likely spam.

[5] Schemas were identified by early cognitive psychologists and adapted as a target for early, symbolic AI research, which helped solidify the cognitivist approach to cognition, which views human cognition as information processing. Thus, for the high-level treatment in the present investigation, one can consider both people and AI as having schemas.

References

Ahmad, Adnan, Brian Whitworth, and Elisa Bertino. 2022. "A Framework for the Application of Socio-Technical Design Methodology." *Ethics and Information Technology* 24 (4): 46. https://doi.org/10.1007/s10676-022-09651-0.

Ahrens, Anthony H., and David Cloutier. 2019. "Acting for Good Reasons: Integrating Virtue Theory and Social Cognitive Theory." *Social and Personality Psychology Compass* 13 (4): e12444. https://doi.org/10.1111/spc3.12444.

Bengio, Yoshua, Ian Goodfellow, and Aaron Courville. 2016. *Deep Learning*. Cambridge, MA: MIT Press.

Bertalanffy, Ludwig von. 1969. *General System Theory: Foundations, Development, Applications*. New York: G. Braziller.

———. 1975. *Perspectives on General System Theory: Scientific-Philosophical Studies*. New York: G. Braziller.

Binz, Marcel, and Eric Schulz. 2023. “Using Cognitive Psychology to Understand GPT-3.” *Proceedings of the National Academy of Sciences of the United States of America* 120 (6): e2218523120. https://doi.org/10.1073/pnas.2218523120.

Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, et al. 2022. “On the Opportunities and Risks of Foundation Models.” *arXiv*. https://doi.org/10.48550/arXiv.2108.07258.

Brooks, Rodney A., Cynthia Breazeal (Ferrell), Robert Irie, Charles C. Kemp, Matthew Marjanović, Brian Scassellati, and Matthew M. Williamson. 1998. “Alternative Essences of Intelligence.” In *Proceedings of the Fifteenth National/Tenth Conference on Artificial Intelligence/Innovative Applications of Artificial Intelligence*, 961–68. AAAI ’98/IAAI ’98. Menlo Park, CA: American Association for Artificial Intelligence.

Brown, Warren S., and Brad D. Strawn. 2012. *The Physical Nature of Christian Life: Neuroscience, Psychology, and the Church*. New York: Cambridge University Press.

Brunila, Mikael, and Jack LaViolette. 2022. “What Company Do Words Keep? Revisiting the Distributional Semantics of J. R. Firth & Zellig Harris.” In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4403–17. Seattle, WA: Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-main.327.

Cervone, Daniel. 2008. “Explanatory Models of Personality: Social-Cognitive Theories and the Knowledge-and-Appraisal Model of Personality Architecture.” In *The SAGE Handbook of Personality Theory and Assessment, Vol 1: Personality Theories and Models*, 80–100. Thousand Oaks, CA: Sage Publications, Inc. https://doi.org/10.4135/9781849200462.n4.

Cervone, Daniel, and Brian R. Little. 2019. “Personality Architecture and Dynamics: The New Agenda and What’s New about It.” *Personality and Individual Differences*, Dynamic Personality Psychology, 136 (January): 12–23. https://doi.org/10.1016/j.paid.2017.07.001.

Chalmers, David. 2010. “The Singularity: A Philosophical Analysis.” *Journal of Consciousness Studies* 17 (9–10): 7–65.

Chalmers, David J. 2006. “Strong and Weak Emergence.” In *The Re-Emergence of Emergence*, edited by Philip Clayton and Paul Davies, 244–56. Oxford: Oxford University Press.

Clayton, Philip. 2004. *Mind and Emergence: From Quantum to Consciousness*. New York: Oxford University Press.

Cloutier, David, and Anthony H. Ahrens. 2020. “Catholic Moral Theology and the Virtues: Integrating Psychology in Models of Moral Agency.” *Theological Studies* 81 (2): 326–47. https://doi.org/10.1177/0040563920928563.

Cooper, Richard, and David Glasspool. 2001. “Learning Action Affordances and Action Schemas.” In *Connectionist Models of Learning, Development and Evolution*, edited by Robert M. French and Jacques P. Sougné, 133–42. Perspectives in Neural Computing. London: Springer.

Cruz, Joe. 2019. “Shared Moral Foundations of Embodied Artificial Intelligence.” In *AIES ’19: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. New York: Association for Computing Machinery. https://doi.org/10.1145/3306618.3314280.

Deacon, Terrence W. 1997. *The Symbolic Species: The Co-Evolution of Language and the Brain*. New York: W.W. Norton.

———. 2011. *Incomplete Nature: How Mind Emerged from Matter*. New York: W.W. Norton.

Dorobantu, Marius. 2021. “Human-Level, But Non-Humanlike: Artificial Intelligence and a Multi-Level Relational Interpretation of the Imago Dei.” *Philosophy, Theology and the Sciences* 8 (1): 81–107. https://doi.org/10.1628/ptsc-2021-0006.

Dreyfus, Hubert L. 2007. “Why Heideggerian AI Failed and How Fixing It Would Require Making It More Heideggerian.” *Philosophical Psychology* 20 (2): 247–68.

Edwards, Denis. 1983. *Human Experience of God*. New York: Paulist Press.

Edwards, Paul N. 2003. "Infrastructure and Modernity: Force, Time, and Social Organization in the History of Sociotechnical Systems." In *Modernity and Technology*, edited by Thomas J. Misa, Philip Brey, and Andrew Feenberg, 185–226. Cambridge, MA: MIT Press.

Epley, Nicholas, Adam Waytz, and John T. Cacioppo. 2007. "On Seeing Human: A Three-Factor Theory of Anthropomorphism." *Psychological Review* 114 (4): 864–86. https://doi.org/10.1037/0033-295X.114.4.864.

Firth, John. 1957. "A Synopsis of Linguistic Theory 1930–1955." In *Special Volume of the Philological Society*. Oxford: Oxford University Press.

Floridi, Luciano. 2019. "What the Near Future of Artificial Intelligence Could Be." *Philosophy & Technology* 32 (1): 1–15. https://doi.org/10.1007/s13347-019-00345-y.

François-Lavet, Vincent, Peter Henderson, Riashat Islam, Marc G. Bellemare, and Joelle Pineau. 2018. "An Introduction to Deep Reinforcement Learning." *Foundations and Trends in Machine Learning* 11 (3–4): 219–354. https://doi.org/10.1561/2200000071.

Gardner, Howard. 1985. *The Mind's New Science : A History of the Cognitive Revolution*. New York: Basic Books.

Garnelo, Marta, and Murray Shanahan. 2019. "Reconciling Deep Learning with Symbolic Artificial Intelligence: Representing Objects and Relations." *Current Opinion in Behavioral Sciences*, SI: 29: Artificial Intelligence (2019), 29 (October): 17–23. https://doi.org/10.1016/j.cobeha.2018.12.010.

Gaudet, Matthew J. 2022. "An Introduction to the Ethics of Artificial Intelligence." *Journal of Moral Theology* 11 (Special Issue 1). https://doi.org/10.55476/001c.34121.

Gibson, James J. 1979. *The Ecological Approach to Visual Perception*. Boston: Houghton Mifflin.

Gill, Karamjit S. 2019. "From Judgment to Calculation: The Phenomenology of Embodied Skill." *AI & SOCIETY* 34 (2): 165–75. https://doi.org/10.1007/s00146-019-00884-0.

Goldberg, Yoav. 2016. "A Primer on Neural Network Models for Natural Language Processing." *Journal of Artificial Intelligence Research* 57:345–420.

Graves, Mark. 2009. "The Emergence of Transcendental Norms in Human Systems." *Zygon: Journal of Religion and Science* 44 (3): 501–32.

———. 2021. "Emergent Models for Moral AI Spirituality." *International Journal of Interactive Multimedia and Artificial Intelligence* 7 (1. Special Issue on AI, Spirituality, and Analogue Thinking): 7–15. https://doi.org/10.9781/ijimai.2021.08.002.

———. 2022a. "Apprehending AI Moral Purpose in Practical Wisdom." *AI & SOCIETY*. https://doi.org/10.1007/s00146-022-01597-7.

———. 2022b. "Theological Foundations for Moral Artificial Intelligence." *Journal of Moral Theology* 11 (Special Issue 1): 182–211. https://doi.org/10.55476/001c.34130.

———. 2023. "Interaction in Emergent Human Systems." *Theology and Science* 21 (2): 331–39. https://doi.org/10.1080/14746700.2023.2188377.

Green, Joel B. 2008. *Body, Soul, and Human Life : The Nature of Humanity in the Bible*. Grand Rapids, MI: Baker Academic.

Hagendorff, Thilo. 2023. "Machine Psychology: Investigating Emergent Capabilities and Behavior in Large Language Models Using Psychological Methods." arXiv.org. https://arxiv.org/abs/2303.13988v1.

Hampson, Peter J., Timothy L. Hulsey, and Phillip P. McGarry. 2021. "Moral Affordance, Moral Expertise, and Virtue." *Theory & Psychology* 31 (4): 513–32. https://doi.org/10.1177/09593543211021662.

Harris, Zellig. 1968. *Mathematical Structures of Language*. New York: Interscience.

Haugeland, John. 1985. *Artificial Intelligence: The Very Idea*. Cambridge, MA: MIT Press.

Herzfeld, Noreen. 2023. *The Artifice of Intelligence: Divine and Human Relationship in a Robotic Age*. Minneapolis, MN: Fortress Press.

Huebner, Bryce, Susan Dwyer, and Marc Hauser. 2009. "The Role of Emotion in Moral Psychology." *Trends in Cognitive Sciences* 13 (1): 1–6. https://doi.org/10.1016/j.tics.2008.09.006.

Hutchins, Edwin. 2010. "Cognitive Ecology." *Topics in Cognitive Science* 2 (4): 705–15. https://doi.org/10.1111/j.1756-8765.2010.01089.x.

Jayawickreme, Eranda, Peter Meindl, Erik G. Helzer, R. Michael Furr, and William Fleeson. 2014. "Virtuous States and Virtuous Traits: How the Empirical Evidence Regarding the Existence of Broad Traits Saves Virtue Ethics from the Situationist Critique." *Theory and Research in Education* 12 (3): 283–308. https://doi.org/10.1177/1477878514545206.

Jeeves, Malcolm A., and Warren S. Brown. 2009. *Neuroscience, Psychology, and Religion*. Conshohocken, PA: Templeton Foundation Press.

Johnson-Laird, Philip Nicholas. 1983. *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Cambridge, MA: Harvard University Press.

Jones, Erik, and Jacob Steinhardt. 2022. "Capturing Failures of Large Language Models via Human Cognitive Biases." In *NeurIPS 2022*. https://doi.org/10.48550/arXiv.2202.12299.

Kahneman, Daniel. 2013. *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.

Laird, John E., Christian Lebiere, and Paul S. Rosenbloom. 2017. "A Standard Model of the Mind: Toward a Common Computational Framework across Artificial Intelligence, Cognitive Science, Neuroscience, and Robotics." *AI Magazine* 38 (4): 13. https://doi.org/10.1609/aimag.v38i4.2744.

Latapie, Hugo, Ozkan Kilic, Kristinn R. Thórisson, Pei Wang, and Patrick Hammer. 2022. "Neurosymbolic Systems of Perception and Cognition: The Role of Attention." *Frontiers in Psychology* 13 (May): 806397. https://doi.org/10.3389/fpsyg.2022.806397.

Lin, Stephanie, Jacob Hilton, and Owain Evans. 2022. "TruthfulQA: Measuring How Models Mimic Human Falsehoods." In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3214–52. Dublin, Ireland: Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.229.

Lobo, Lorena, Manuel Heras-Escribano, and David Travieso. 2018. "The History and Philosophy of Ecological Psychology." *Frontiers in Psychology* 9:2228. https://doi.org/10.3389/fpsyg.2018.02228.

Lumbreras, Sara. 2023. "Lessons from the Quest for Artificial Consciousness: The Emergence Criterion, Insight-Oriented AI, and Imago Dei." *Zygon: Journal of Religion and Science*. https://doi.org/10.1111/zygo.12827.

Makarius, Erin E., Debmalya Mukherjee, Joseph D. Fox, and Alexa K. Fox. 2020. "Rising with the Machines: A Sociotechnical Framework for Bringing Artificial Intelligence into the Organization." *Journal of Business Research* 120 (November): 262–73. https://doi.org/10.1016/j.jbusres.2020.07.045.

Malhotra, Arvind, Ann Majchrzak, and Kalle Lyytinen. 2021. "Socio-Technical Affordances for Large-Scale Collaborations: Introduction to a Virtual Special Issue." *Organization Science* 32 (5): 1371–90. https://doi.org/10.1287/orsc.2021.1457.

Marr, David. 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Cambridge, MA: MIT Press.

McAdams, Dan P. 2013. "The Psychological Self as Actor, Agent, and Author." *Perspectives on Psychological Science* 8 (3): 272–95.

McGann, Marek, Ezequiel A. Di Paolo, Manuel Heras-Escribano, and Anthony Chemero. 2020. "Editorial: Enaction and Ecological Psychology: Convergences and Complementarities." *Frontiers in Psychology* 11:617898. https://doi.org/10.3389/fpsyg.2020.617898.

Mitchell, Tom M. 1997. *Machine Learning*. New York: McGraw-Hill Education.

Moore, Cristopher, and Stephan Mertens. 2011. *The Nature of Computation*. New York: Oxford University Press.

Müller, Vincent C., and Nick Bostrom. 2014. "Future Progress in Artificial Intelligence: A Survey of Expert Opinion." In *Fundamental Issues of Artificial Intelligence*, 555–72. Synthese Library. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-26485-1_33.

Murphy, Nancey C. 2006. *Bodies and Souls, or Spirited Bodies?* New York: Cambridge University Press.

Murphy, Nancey C., and Warren S. Brown. 2007. *Did My Neurons Make Me Do It?: Philosophical and Neurobiological Perspectives on Moral Responsibility*. Oxford: Oxford University Press.

Nagel, Thomas. 1974. "What Is It Like to Be a Bat?" *Philosophical Review* 83 (4): 435–50.

Newell, Allen, and Herbert A. Simon. 1961. *GPS, a Program That Simulates Human Thought. Lernende Automaten*. Munich: Oldenbourg.

Noble, Wiliam G. 1981. "Gibsonian Theory and the Pragmatist Perspective." *Journal for the Theory of Social Behaviour* 11 (1): 65–85. https://doi.org/10.1111/j.1468-5914.1981.tb00023.x.

Noë, Alva. 2004. *Action in Perception*. Cambridge, MA: MIT Press.

OpenAI. 2022. "ChatGPT: Optimizing Language Models for Dialogue." *OpenAI (blog)*. November 30, 2022. https://openai.com/blog/chatgpt/.

Pahl, Claus, Antonio Brogi, Jacopo Soldani, and Pooyan Jamshidi. 2017. "Cloud Container Technologies: A State-of-the-Art Review." *IEEE Transactions on Cloud Computing*.

Pearl, Judea, and Dana Mackenzie. 2018. *The Book of Why: The New Science of Cause and Effect*. New York: Basic Books.

Rumelhart, David E., and James L. McClelland. 1987. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations, Psychological and Biological Models*. Cambridge, MA: MIT Press.

Russell, Stuart, Daniel Dewey, and Max Tegmark. 2015. "Research Priorities for Robust and Beneficial Artificial Intelligence." *AI Magazine* 36 (4): 105–14. https://doi.org/10.1609/aimag.v36i4.2577.

Russell, Stuart, and Peter Norvig. 2010. *Artificial Intelligence : A Modern Approach*. Upper Saddle River, NJ: Prentice Hall.

Sahlgren, Magnus. 2008. "The Distributional Hypothesis." *Rivista Di Linguistica* 20 (1): 33–53. https://www.italian-journal-linguistics.com/app/uploads/2021/05/Sahlgren-1.pdf.

Sejnowski, Terrence J. 2020. "The Unreasonable Effectiveness of Deep Learning in Artificial Intelligence." *Proceedings of the National Academy of Sciences of the United States of America* 117 (48): 30033–38. https://doi.org/10.1073/pnas.1907373117.

Selbst, Andrew D., Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. "Fairness and Abstraction in Sociotechnical Systems." In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 59–68. FAT* '19. New York: Association for Computing Machinery. https://doi.org/10.1145/3287560.3287598.

Simon, Herbert Alexander. 1969. *The Sciences of the Artificial*. Cambridge, MA: MIT Press.

Singh, Munindar P. 2014. "Norms as a Basis for Governing Sociotechnical Systems." *ACM Transactions on Intelligent Systems and Technology* 5 (1): 21:1–21:23. https://doi.org/10.1145/2542182.2542203.

Skyttner, Lars. 2006. *General Systems Theory: Perspectives, Problems, Practice*. 2nd ed. Singapore; River Edge, NJ: World Scientific.

Smith, Brian Cantwell. 2019. *The Promise of Artificial Intelligence: Reckoning and Judgment*. Cambridge, MA: The MIT Press.

Smith, John Edwin. 1968. *Experience and God*. New York: Oxford University Press.

Teske, John A. 2013. "From Embodied to Extended Cognition." *Zygon: Journal of Religion and Science* 48 (3): 759–87.

Trist, Beulah. 1990. *The Social Engagement of Social Science, Volume 2: A Tavistock Anthology–The Socio-Technical Perspective*. Vol. 2. Philadelphia: University of Pennsylvania Press.

Umbrello, Steven. 2019. "Beneficial Artificial Intelligence Coordination by Means of a Value Sensitive Design Approach." *Big Data and Cognitive Computing* 3 (1): 5. https://doi.org/10.3390/bdcc3010005.

Valenti, S. Stavros, and James M. M. Gold. 1991. "Social Affordances and Interaction I: Introduction." *Ecological Psychology* 3 (2): 77–98. https://doi.org/10.1207/s15326969eco0302_2.

van de Poel, Ibo. 2020. "Embedding Values in Artificial Intelligence (AI) Systems." *Minds and Machines* 30 (3): 385–409. https://doi.org/10.1007/s11023-020-09537-4.

Varela, Francisco J., Evan Thompson, and Eleanor Rosch. 1991. *The Embodied Mind : Cognitive Science and Human Experience*. Cambridge, MA: MIT Press.

Vestrucci, Andrea. 2023. "Introduction: Five Steps Toward a Religion–AI Dialogue." *Zygon: Journal of Religion and Science*. https://doi.org/10.1111/zygo.12828.

Walker, Lawrence J., and Jeremy A. Frimer. 2015. "Developmental Trajectories of Agency and Communion in Moral Motivation." *Merrill-Palmer Quarterly* 61 (3): 412–39.

Watts, Fraser. 2013. "Embodied Cognition and Religion." *Zygon: Journal of Religion and Science* 48 (3): 745–58.

Weisberg, Michael. 2013. *Simulation and Similarity: Using Models to Understand the World*. New York: Oxford University Press.

Zubiri, Xavier. 2003. *Dynamic Structure of Reality*. Translated by Nelson R. Orringer. Champaign, IL: University of Illinois.