International Journal of Interactive Multimedia and Artificial Intelligence

September 2021, Vol. VII, Number 1
ISSN: 1989-1660

unir LA UNIVERSIDAD EN INTERNET

Why are we so worried about artificial intelligence? Surely humans are always able to pull the plug? People asked a computer, 'Is there a God?' And the computer said, 'There is now,' and fused the plug.
Stephen Hawking

Special Issue on Artificial Intelligence, Spirituality and Analogue Thinking

INTERNATIONAL JOURNAL OF INTERACTIVE MULTIMEDIA AND ARTIFICIAL INTELLIGENCE

ISSN: 1989-1660 –VOL. 7, NUMBER 1

EDITORIAL TEAM

Editor's Note

I. Introduction to Artificial Intelligence and Spirituality

The development of machines and automation and their integration into daily life has often led to a deeper examination of human nature. Today technological change is accelerating, not only in artificial intelligence (AI), but also in communications, automation, biology and medicine, raising ever deeper questioning of the human condition. Within this context, many people experience a feeling that transcends the self and evokes less immediate views and scopes. Our thirst for transcendence and introspection remains unwavering, despite technological breakthroughs, or perhaps because of them; despite the speed of these breakthroughs, or because of it; despite their diversity, or precisely because of it. This unwavering aspect is in no way a threat to progress, but a part of it, with these transcendental experiences providing us with support and sustenance in our daily lives.

Since their first steps, computation and AI have been interpreted in various and even opposed ways: as progress, assisting humans in several tasks and helping to save time and energy; and as a danger, a threatening development that would dehumanize and even enslave people. The dystopias described in movies over the past twenty years express the fear that AI would become dominant and challenge human capacities.

In 2005 Ray Kurzweil [1, p.25] famously predicted that within decades artificial intelligence would surpass human capabilities, and that "There will be no distinction, post-Singularity, between human and machine". The rapid development of AI technologies has appeared to support his view. However, his conjecture remains contested in part because of a lack of clarity concerning what "intelligence" consists of, with the field continuing "to be marked by noisy and sometimes vitriolic debates" [2, p.3]. Kurzweil's claim implies that there is a single scale of intelligence, applicable to all individual organisms and machines, but this is placed in doubt by the continuing interest in the theory of multiple intelligences, dating back to Gardner [3]. Similarly, Kurzweil implies that intelligence is a quality that can be abstracted from its substrate, running counter to long-standing schools of thought that see intelligence as embodied, embedded in and extended into the environment, and enacted in interactions with the environment [4].

Recent progress in AI has been substantial, but awareness is also emerging of constraints to its capabilities [5]. At the same time, the social and cultural evolution that accompanies this progress invites us to re-order and redefine several human and social dimensions. This might contribute to reaching a degree of maturity in our knowledge and assessments of AI that allows for more nuanced interpretations. The perplexities, fears, doubts, questions concerning the interactions between AI and the human can become more subtle and less dramatic thanks to a more precise understanding of AI developments., one in which the contributions of AI can be recognised, while the issues arising become more subtle and less dramatic.

It is in that context that the present special issue finds its place. The papers seek to shed light on debates such as these, by linking and contrasting AI and the spiritual dimension. They invite the reader to consider the relationship between AI and an aspect of human experience which is generally seen as the polar opposite of computation, that of spirituality. By carefully weighting the connections and the contrasts of AI and the spiritual dimension, the contributions make the case for a more attentive examination of pressing issues that we can no longer ignore, and which require a highly interdisciplinary approach.

II. Ways to Approach Research on the Topic

Research on the relationship between computing and the meaning of human life flourishes proportionally to the increasing digitalization of our world. More and more, reflections on ethics and politics, spiritual values and religious experiences, beliefs, and practices make use of digital media in order to spread their content or express themselves. If we still consider that there is truth in the well-known dictum that "the medium is the message" [6], then it is worth asking how the content of these reflections and practices are changing today.

Every change is the introduction of something new, and this novelty can be interpreted either as the improvement or the worsening of the current situation. Generally speaking, research on either the positive or negative interactions between the advances in AI and the dimension of spirituality and analogue thinking are based on at least three approaches. The first produces analogies between concepts from human studies and concepts from computer science; for instance, speaking of "modeling" for concepts in human sciences, or considering the universe to be intelligently organized in an algorithmic order [7], [8]. The second approach is the application of research on AI and computer science to develop new insights on the extents, limits, and perfectibility of spiritual topics, discussions, or even practices [9]. Finally, the third approach applies sociological, philosophical, aesthetic, or even theological concepts to assess the changes that digitalization introduces in spiritual practices, beliefs, and cultures [10].

III. Aim of this Special Issue

This special issue analyzes the current state of the art, and it addresses all three models of the research. By doing so, the issue will place the general question of the distinction between human and machine into sharper relief.

In the issue, authors provide diverse insights to the topic. Graves uses general systems theory to organize models of human experience, yielding insight into human morality and spirituality, upon which AI modelling can also draw. Krüger discusses how the concept of singularity is reviewed from a cultural studies perspective, first with regard to the cosmological singularity and then to the technological singularity. Vestrucci et al. analyze and debate current topics of investigation on the relationship between AI and the concept of belief, such as: The modelling of belief, the exploration of belief in automated reasoning environment, with specific emphasis on religious belief. Calderero provides an open, synergetic, harmonious vision of the role of technology and the humanities, especially those most focused on the study of the intangible. He argues that it is necessary for the progress of knowledge and, therefore, for the mutually beneficial care of humanity and nature. Dorobantu demonstrates that the similarity between the midwife proposal and the modern Christian anthropology and cosmology is only superficial. Compared to the midwife hypothesis, Christian theological accounts define the cosmic role of humanity quite differently, and they provide a more satisfactory teleology. Burgos presents a semi-automatic process to assess the degree of ritual identity, reinforcing the hypothesis that rituals follow a similar pattern of structure and preparation according to a predetermined set of common elements, whether linked to religious or secular settings. Finally, Griffiths discusses the degree to which Gregory Bateson's concept of an ecology of mind can shed light on the capacities of AI, and in particular its ability to partake of the realm of the sacred.

DOI: 10.9781/ijimai.2021.08.012

This special issue discusses concepts that might, in principle, appear as two opposite poles or two layers of the same reality, but which can also be seen as two interwoven elements defining a single context or two aspects of a particular viewpoint: mutually dependent and difficult to understand separately. The contributors explore an emerging world in which the prospect of superintelligent systems raises both great hopes and great fears. Within this context, the spiritual life and transcendence acquire new functions and significance, and underlying the discussion here is the aspiration that they can still provide sources of meaning and hope.

Prof. Dr. Daniel Burgos[1]
Prof. Dr. Lluís Oviedo[2]
Prof. Dr. Dai Griffiths[1]
Prof. Dr. Andrea Vestrucci[3,4]

[1] Research Institute for Innovation & Technology in Education (UNIR iTED), Universidad Internacional de La Rioja (UNIR), Logroño (Spain)

[2] Pontifical University Antonianum, Rome (Italy)

[3] Graduate Theological Union, Berkeley, California (USA)

[4] University of Geneva (Switzerland)

References

[1] R. Kurzweil, The Singularity is Near, Viking, London, 2005.
[2] R. J. Sternberg, The Concept of Intelligence, Handbook of Intelligence, pp. 3-15. Cambridge University Press, 2020.
[3] H. Gardner, Frames of Mind: The Theory of Multiple Intelligences, Basic Books, New York, 1983.
[4] A. Newen, L. De Bruin, & S. Gallagher, The Oxford Handbook of 4e Cognition, Oxford University Press, 2018.
[5] S. Shwartz, Evil Robots, Killer Computers, and Other Myths: The Truth About AI and the Future of Humanity, Fast Company Press, 2021.
[6] Marshall McLuhan, Understanding Media: The Extensions of Man, Cambridge, MA: MIT Press, 1994 [first edition 1964].
[7] G. Marcus, Can Science Lead to Faith?, The New Yorker, April 26, 2013. Available at https://www.newyorker.com/tech/annals-of-technology/can-science-lead-to-faith
[8] J. Schmidhuber, In the beginning was the code: Juergen Schmidhuber at TEDxUHasselt, Belgium, November 10, 2012. Available at https://youtu.be/T1Ogwa76yQo
[9] E. Sutinen, A.-P. Cooper, Digital Theology: A Computer Science Perspective, Bingley: Emerald, 2021.
[10] P. M. Phillips, The Bible, Social Media, and Digital Culture, New York: Routledge, 2020.

TABLE OF CONTENTS

OPEN ACCESS JOURNAL

ISSN: 1989-1660

The International Journal of Interactive Multimedia and Artificial Intelligence is covered in Clarivate Analytics services and products. Specifically, this publication is indexed and abstracted in: *Science Citation Index Expanded, Journal Citation Reports/ Science Edition, Current Contents®/Engineering Computing and Technology.*

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.

"The Singularity is near!" Visions of Artificial Intelligence in Posthumanism and Transhumanism

Oliver Krüger*

University of Fribourg (Switzerland)

Received 16 January 2021 | Accepted 19 June 2021 | Published 20 July 2021

Abstract

Over the past 20 years, the idea of singularity has become increasingly important to the technological visions of posthumanism and transhumanism. The article first introduces key posthumanist authors such as Marvin Minsky, Ray Kurzweil, Hans Moravec, and Frank Tipler. In the following, the concept of singularity is reviewed from a cultural studies perspective, first with regard to the cosmological singularity and then to the technological singularity. According to posthumanist thinkers the singularity is marked by the emergence of a superhuman computer intelligence that will solve all of humanity's problems. At the same time, it heralds the end of the human era. Most authors refer to the British mathematician Irving John Good's 1965 essay *Speculations Concerning the First Ultraintelligent Machine* as the originator of the idea of superintelligence. Individual elements of the singularity idea such as the impenetrable event horizon, the *frontier* and the ongoing acceleration of progress are contextualized historically and culturally.

> *Singularity, The. The Techno-Rapture. A black hole in the Extropian worldview whose gravity is so intense that no light can be shed on what lies beyond it.*
>
> Godling's Glossary [1]

Keywords

Singularity, Artificial Intelligence, Ray Kurzweil, Transhumanism, Posthumanism.

DOI: 10.9781/ijimai.2021.07.004

I. Post- and Transhumanism

AMIDST the range of diverse thinkers advocating the overcoming of humanity with the help of new technologies, many are often called transhumanists. Yet despite this increasingly frequent usage, I would still like to emphatically call for a differentiation between technological posthumanism and transhumanism. Not only does the term posthumanism, which is commonly used in art and cultural studies research, itself need to be clarified, but in fact noticeable differences can be found between the purposes, contents and origins of transhumanism and technological posthumanism.

Transhumanism primarily originated in California during the 1960s, and was decisively influenced by the visions of the futurist Fereidoun M. Esfandiary (FM-2030), by the commitment of Timothy Leary, the pioneer of the psychedelic movement, and by the cryonics expert Robert Ettinger. In the late 1980s this movement gave rise to the "Extropians" around Max More and, as European involvement increased, the *World Transhumanist Association* founded by Nick Bostrom, David Pearce and Anders Sandberg in 1998.

Technological posthumanism, on the other hand, unites a number of authors who have been propagating the replacement of humans by artificial intelligences since the mid-1980s. Its main four proponents, Hans Moravec, Frank Tipler, Marvin Minsky, and Ray Kurzweil, argue on the basis of cybernetic theory. Before the early 2000s these authors did not refer to the transhumanist movement and its themes in any way.

The second argument for a separation between post- and transhumanism is based on the different emphases in terms of content. Transhumanists deal practically with the issues of prolonging life and enhancement of mental performance, such as through the use of smart drugs, life-prolonging diets, advances in prosthetic technology, the potential for a renewed form of eugenics, or even the prospects of cryonics, while these applications are rarely mentioned in posthumanist writings.

Whereas in transhumanism the subject of development is humankind and what becomes of human beings with the help of technological upgrades and enhancements, in posthumanism robots and artificial intelligence are the future carriers of evolution and progress. In a virtual habitat the immortal existence of humans is a welcome side effect of the autonomous progress of artificially intelligent, post-human beings [19].

Now, who are the most significant thinkers in posthumanism? In his work *Mind Children. The Future of Robot and Human Intelligence* (1988) the roboticist Hans Moravec (born 1948) offered a vision of a *post-biological* and *supernatural* future for humankind. The preface reads like a preamble to posthumanism:

> Engaged for billions of years in a relentless, spiraling arms race with one another, our genes have finally outsmarted themselves ... What awaits us is not oblivion but rather a future which, from our present vantage point, is best described by the words "postbiological" or even "supernatural". It is a world in which the human race has been swept away by the tide of cultural change, usurped by its own artificial progeny ... [2:1]

* Corresponding author.
E-mail addresses: oliver.krueger@unifr.ch

Moravec's outstanding importance for posthumanist philosophy stems primarily from the fact that, he was the first scientist to formulate the technical possibilities of virtual immortality. Not as a science fiction author, but as a scientific visionary, Moravec portrays the technical procedure of this possible "transmigration" in precise detail as a scanning process of the brain. He thus develops his vision of humans as virtual simulation within a computer's memory, which will ensure his infinite existence while biological humanity slowly dies out [2:108-109].

Frank Jennings Tipler (born 1947) serves as professor for mathematical physics at Tulane University in New Orleans since 1981. His research mainly focuses on questions of general relativity, quantum theory, and cosmology related to his interest in the genesis and future development of the cosmos. Tipler became famous overnight with his 1994 book *The Physics of Immortality. Modern Cosmology, God and the Resurrection of the Dead* [3]. His position differs from that of other posthumanists in many regards – whether it be the cosmological emphasis, his euphoric images of virtual paradise, or his scientific inclusivism, which does not seek to overcome religion but to integrate it. According to Tipler, when the sun has burned all of fuel, in many billions of years, the only chance of survival for humans will become a virtual existence in gigantic computers. Tipler determines the goal of these cosmological developments as the Omega Point, which he identifies with God.

Marvin Minsky's (1927-2016) significance for posthumanism lies above all in the formulation of the cybernetic understanding of humankind, that means to define the human being as a particular type of information processing machines. Even in an inconspicuous textbook on computer science, he places the evolution of humans in relation to that of machines: "One has found himself sharing the world with a strange new species within a single generation: the computers and computer-like machines." [4:VII]. As a co-founder of the Massachusetts Institute of Technology's Media Lab, Minsky was the teacher and mentor of a number of the contemporary representatives of posthumanism and transhumanism such as Luc Steels, Eric Drexler, and Ray Kurzweil (born 1948).

The latter is certainly Marvin Minsky's most famous former student. He has founded no less than six companies in the information technology industry since graduating from MIT in 1970. Another career high point was certainly his 2012 appointment as a Director of Engineering at Google, where he focuses upon machine learning and language processing. In various interviews, Kurzweil always emphasizes that he is doing his utmost to achieve the singularity [5:14-17].

His early work *The Age of Intelligent Machines* [6], published in 1990, was the best-selling book in computer science at the time. It provides a technical overview of the development of artificial intelligence. The book contains a short future scenario depicting potential consequences of the increasing use of machines in the working world, as well as some predictions for future leisure activities [6:401-416]. In 1990 Kurzweil's grandest prophecy was that a computer will have developed its own consciousness sometime between 2020 and 2070 [6:483]. However, Kurzweil wants to introduce the beginning of the end of humankind in his next book *The Age of Spiritual Machines* of 1999: According to him, by the year 2099 humans and machines will have merged, and humankind will have overcome its biological condition [7:277-280]. In his most radical work, *The Singularity is Near. When Humans Transcend Biology* of 2005, the prospect of salvation is accelerated by half a century to the year 2045, and Kurzweil promises a universal solution to all of humanity's problems [8]. Since the 1990s, Kurzweil has also been writing life-help books such as *Fantastic Voyage: Live Long Enough to Live Forever* [9] and *Transcend: Nine Steps to Living Well Forever* [10], both co-authored with Terry Grossman. In 2009, a documentary film about Kurzweil called *Transcendent Man. The Life and Ideas of Ray Kurzweil* was even screened [11].

II. Singularities

A. Introduction

The idea of the dawning of a new age of artificial intelligence has gained recognition far beyond the transhumanist milieu, primarily through Ray Kurzweil's book *The Singularity is near: When Humans Transcend Biology* (2005), numerous films and the founding of the *Singularity University* (*SU*) in 2008. Strictly speaking, the SU is not a university at all, it provides no curriculum, qualifying degrees and research facilities. It offers mainly marketing and network-working events for "disruptive" technological visions [12:63-76].

From a cultural studies perspective, this essay examines the cultural, religious, and philosophical elements of the singularity idea. This is not a scientific evaluation of the singularity or its technological feasibility. But with this analysis, cultural values and ideas can be uncovered that are also present in the further technological and political discourse on artificial intelligence. On closer inspection, the singularity proves to be a cultural rather than a technological idea.

Cultural studies scholars have previously attempted to arrange and analyze different approaches to singularity, yet these sometimes remain undifferentiated and polemical: Selmer and Alexander Bringsjord and Paul Bello see the entire singularity theory as a matter of faith without scientific basis [13]. The idea of singularity encompasses scientific concepts within mathematical function and system theory, geometry, solid-state physics, cosmology and cybernetics. The latter two areas particularly hold special significance for posthumanism. Even when merely scratching at the surface of the history of ideas, it quickly becomes clear that these two areas are closely interwoven. They contain numerous references, especially to the work of the Jesuit Pierre Teilhard de Chardin and his concept of the Omega Point. Alongside Reinhard Heil, I advocate considering each semantic layer individually, in order to elaborate the complex interdependencies between religion and science in this posthumanist utopia [14:44-46]. We will therefore examine the concept of singularity in three steps: The first two sections on cosmological and technological singularity will be followed by a cultural-historical contextualization of the concept itself.

B. Black Holes and Cosmological Singularities

The term singularity has been widely used in English since the 1980s, as well as being creatively applied in literature and television series for popular audiences. According to the cosmologists Roger Penrose and Stephen Hawking, singularities (in the plural) denote the special conditions of space and time, such as those created by black holes. These are moments when matter or its precursors are concentrated at a single point and space and light become infinitely curved. The beginning of the universe – the Big Bang – was marked by a singularity [15]. The common understanding of singularity as well as the popular reception of the term in literature and television series usually refer to the fantastic space and time effects of black holes, to which the Penrose-Hawking singularity theorem is applied.

Together with cosmologist John D. Barrow, Frank Tipler steered the concept of cosmological singularities into philosophical realms encompassing questions of life and humanity's place in the universe. The two cosmologists reflect on the initial and final singularity within a closed universe model, i.e. the beginning and the end of the universe, which at this moment has no spatial-temporal extension. Here, Barrow and Tipler identify analogies with Teilhard de Chardin's work and equate the final singularity with the divine Omega Point. These

two approaches can in fact be combined, since according to the *Final Anthropic Principle*, the end of the universe requires a final observer, which for Tipler is identical with God-Omega [16: 201-204, 470-471]. In his later works, *Physics of Immortality* of 1994 [3] and *Physics of Christianity* of 2007 [17], Tipler builds on these considerations and embeds the cosmological singularities in a theological framework, i.e. not only that God is the final goal of the universe, but that God is also its original cause, which was not yet subject to any physical laws.

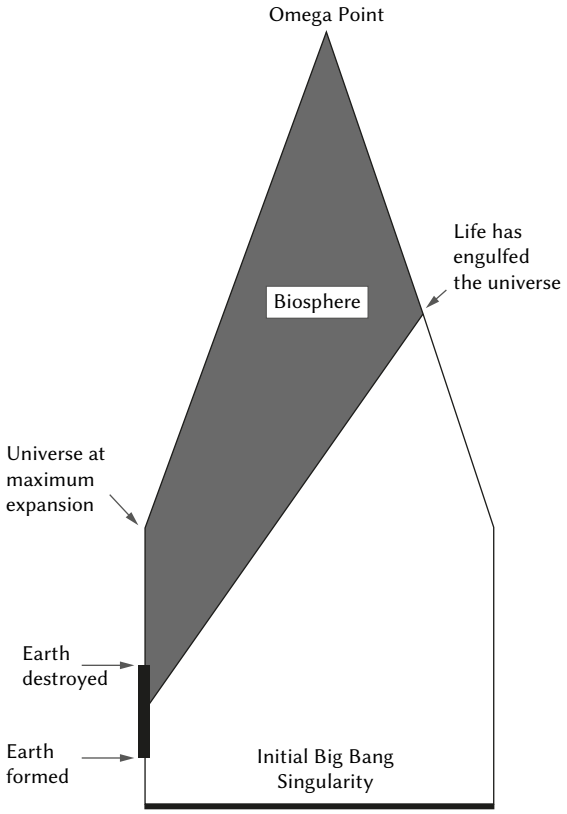


Fig 1. Penrose diagram of the future of life in the universe by Frank Tipler [3:145], Figure IV.9.

This image [Fig. 1] illustrates the temporal dimensions of Tipler's design. The earth's biosphere first begins to expand into space during our present age, in order to save the universe as it is colonized. In a 2013 interview with "Socrates" from *Singularity Weblog,* Tipler describes the properties of the final cosmological singularity as follows:

> The singularity is outside the natural world, it is beyond the natural world, and it is transcendent to the natural world. So, approaching the singularity … the amount of information, the amount of knowledge is approaching infinity as you are going into the final state. The processing rate is increasing to infinity. So, the total amount of information processing will be infinite [18].

Tipler takes an inclusive approach to the concept of technological singularity propagated by Kurzweil and other thinkers. He considers the technological singularity as merely a philosophical concept, while the cosmological singularity is presented as a proven mathematical theorem. According to Tipler, the technological singularity is only a small event in human history, caused by inevitable movement towards the cosmic singularity.

> The cosmological singularity is determining, requiring the existence of the computer science singularity. And I agree with various people as Hans Moravec and Ray Kurzweil. And I think the singularity in computer science will occur in this century. I think we are very close. I think we already have the necessary hardware [18].

When he calls himself a "fundamentalist physicist", it finally becomes obvious that there is not a hint of irony in Tipler's statements. Under the conditions that the universe is closed and that humanity is the only intelligent life form in the cosmos (both of which are mathematically proven, according to Tipler), earthly life forms *must* find a new vehicle:

> Namely, that eventually human meat, rational beings will be replaced by human downloads and our artificial intelligence of reason at least at the human level. I am convinced that's true. I am convinced it must be true because as you are going into the final singularity, necessarily … life can no longer exist, it has to move on another substrate. And, well, that's just human downloads [18].

C. The Technological Singularity

Post- and transhumanists collectively identify the mathematician and cyberneticist John von Neumann as the originator of the concept of technological singularity [19]. His detailed obituary was written by his long-time friend and scientific companion Stanisław Ulam in 1958 and he recounts an exchange with von Neumann on the idea of an "ever accelerating progress of technology and changes in the mode of human life, which gives the appearance of approaching some essential singularity in the history of the race beyond which human affairs, as we know them, could not continue." [20:5]

A quarter of a century later, the American mathematician and science fiction author Vernor Vinge explicitly bridges the gap between the cosmological and technological concepts of singularity for the first time, in a one-page article for the technology magazine *Omni* in 1983:

> We will soon create intelligences greater than our own. When this happens, human history will have reached a kind of singularity, an intellectual transition as impenetrable as the knotted space-time at the center of a black hole, and the world will pass far beyond our understanding [21:10].

Over the next years, Vinge applied the singularity merely as a running theme in the background of several of his science fiction novels. At NASA's *Vision 21* symposium in 1993 Vinge then confidently announced: "Within thirty years, we will have the technological means to create superhuman intelligence. Shortly after, the human era will be ended." [22] Vinge sketches four ways this technological singularity could appear: first, through computers; second through computer networks that develop consciousness and a superhuman intelligence; third, through human-computer interfaces that make humans super intelligent; or fourth, through the biological improvements of humans. Since the first three possibilities depend heavily on computer hardware, Vinge predicts the arrival of the singularity for the period between 2005 and 2030. He clearly states what his expectations are: "For me, the superhumanity is the essence of the Singularity. Without that we would get a glut of technical riches, never properly absorbed." [23:366].

According to Vinge the singularity will revolutionize all previous structures of human life and will instigate enormous changes in a very short period of time. To date, there is only one corresponding analogy in the history of evolution: "The rise of humankind. We will be in the Post-Human era." [23:367] Everything that will occur after the singularity are completely unknowable. Vinge therefore turns to the concept of the event horizon, as mentioned in his early article from 1983. In astrophysics observations of black holes are not possible beyond this point [23:367].

Vinge as well as most other authors refer to the British mathematician Irving John Good's 1965 essay *Speculations Concerning the First Ultraintelligent Machine* as the originator of the idea of superintelligence [46]. Good studied mathematics at Cambridge and served at Bletchley Park from 1941, where he was involved in the development of the first electronic computer, *Colossus*, under the direction of Alan Turing. Later he was professor of statistics at Virginia

Tech University in the United States. Good introduces his famous essay with a prophetic confession: "The survival of man depends on the early construction of an ultra-intelligent machine." [46:31] This computer, which Good anticipated would have been built by the end of the 20th century, would be far superior to humans in the storage and processing of information:

> Let an ultraintelligent machine be defined as a machine that can far surpass all the intellectual activities of any man however clever. Since the design of machines is one of these intellectual activities, an ultra-intelligent machine could design even better machines; there would then unquestionably be an "intelligence explosion," and the intelligence of man would be left far behind ... Thus the first ultraintelligent machine is the last invention that man need ever make ... [46:33]

As early as 1962 – at the height of the Cuba crisis – Good expected that future Russian and American ultra-intelligent machines (UIM) could merge into a single world government and guarantee a lasting peace: "Oracles of the world unite!" [47:195]

The American AI researcher Eliezer Yudkowsky first sought to transform the idea of technological singularity into a far-reaching philosophical concept through the formulation of the *Singularitarian Principles* in 1999. He served as co-founder of the *Singularity Institute for Artificial Intelligence* (today MIRI), which propelled the singularity debate through its *Singularity Summits*. Yudkowsky identifies as an atheist, transhumanist and cryonics expert, and pleads in his principles for a sharp distinction between the technological singularity and religious concepts.

The large and often rambling document contains many ambitious statements on "ultra-technology", globalization, the deification of the human being (*apotheosis*) and solidarity, as well as some minor aspects. *Singularitarians* are in his view "partisans" who consider technological singularity as superhuman intelligence to be a highly desirable goal to work towards.

> The Singularity holds out the possibility of winning the Grand Prize, the true Utopia, the best-of-all-possible-worlds – not just freedom from pain and stress or a sterile round of endless physical pleasures), but the prospect of endless growth for every human being – growth in mind, in intelligence, in strength of personality; life without bound, without end; experiencing everything we've dreamed of experiencing, *becoming* everything we've ever dreamed of *being* ... [24]

In the late 1990s, Yudkowsky was one of the few activists to introduce a moment of solidarity into the transhumanist debate. Those who advocate deification must also agree that everyone receives divinity. Those who accept the extermination of humanity by AI must therefore also accept their own extermination. The young Yudkowsky was characterized by a messianic optimism and a belief in the technological solution to all problems of existence: "I'm working to save *everyone*, heal the planet, solve *all* the problems of the world." [25].

How does Ray Kurzweil, currently the most influential posthumanist, fit into this debate in comparism to other thinkers? Vernor Vinge legitimizes his own prognosis tautologically: "But if the technological singularity can happen, it will." [22] Frank Tipler justifies technology's future development from the perspective of a cosmological teleology. For Yudkowsky, singularity appears as a given fact. But Ray Kurzweil and Hans Moravec with him choose a different path, one that is apparently oriented towards more verifiable criteria. Both thinkers extrapolate future technological progress by observing previous trends, and Kurzweil alone introduces the concept of singularity in his more recent publications from 2005 [26:95-110], [7:189-252]. It would therefore be prudent to review the development of these forecasts over the past three decades.

If information processing becomes the benchmark for measuring life's perfection, then the past and future will also be interpreted according to this paradigm. Moravec and Kurzweil dedicate large portions of their publications to presenting data on the growth of computers' processing and storage capacities, in addition to detailed questions regarding the possibility of artificial intelligence [2:37-51], [8:14-110]. Both authors attached their hopes for an exponentially accelerated further development and distribution of computers and robots to a quantified law of progress: *Moore's Law* [2:68], [7:13-25]. The assumption that computer development constantly accelerates can be traced to Intel co-founder Gordon Moore, who in the mid-1960s claimed that the size of an integrated circuit halves every 24 months, in other words, it becomes twice as powerful. This prediction, now known as *Moore's Law*, implies an indefinite exponential increase in computer performance [7:17-39].

As they aged, Moravec and also Marvin Minsky both became increasingly reserved and also sometimes more skeptical about the imminent realization of artificial intelligence on a human level. In their latest estimates they expect the emergence of a superhuman AI not before the year 2050 [27] – [28]. Unlike the other posthumanist theorists and transhumanist activists, Ray Kurzweil has not become more cautious or restrained in his statements over the last two decades. His three key books *The Age of Intelligent Machines* (1990) [6], *The Age of Spiritual Machines* (1999) [7], and *The Singularity is Near* (2005) [8] offer a dramatic choreography with a steady increase in futuristic statements. As his trilogy concludes, however, he crosses the boundary between technical prophecy and a spiritual philosophy that is more akin to Christianity or New Age beliefs.

As early as 1999, Kurzweil planned what he called the *Law of Accelerating Returns*. This was intended to replace *Moore's Law* around 2020 and establish an even higher acceleration rate amongst the future generations of self-designing machines. At this point not only would growth continue exponentially, but in fact the exponent itself would grow exponentially. Therefore – according to Kurzweil's 1999 book – around the year 2023 affordable PCs with the computing power of the human brain would become available, while in 2030 they would contain the mental power of an entire village. By 2029, about 99% of the thinking power on our planet would be provided by computers. According to Kurzweil, hardly anyone will continue to work in industrial production, agriculture, or the transportation industry [7:17-39] – [8:24-29].

Kurzweil identifies five stages in the history of evolution leading up to the realization of the singularity: 1. the origin of matter; 2. the origin of life; 3. the origin of brains/mind; 4. the origin of technology; and 5. the fusion of human and machine intelligence. In a sixth phase, superhuman intelligence will begin to colonize the entire universe [8:14-111]. The singularity, which, like the Big Bang, entails creating the entire cosmos anew, marks the absolute climax of this technological prophecy.

Kurzweil only defines this concept briefly: "It's a future period during which the pace of technological change will be so rapid, its impact so deep, that human life will be irreversibly transformed ... " [8:7]. A more precise description is not possible for humans: "So how do we contemplate the Singularity? As with the sun, it's hard to look at directly; it's better to squint at it out of the corner of our eyes." [8:371] Diane Proudfoot points out that this metaphor echoes the doctrine of God's indescribability, which was common in Christian mysticism. Thus Anselm of Canterbury proclaims in the 11th century: "I cannot look directly into [the light in which God dwells], it is too great for me ... it is too bright ... the eye of my soul cannot bear to turn towards it for too long." [29:368]

Kurzweil accentuates the prophetic meaning of his statements with the exact date of the singularity (published in oversized font in the original):

> *I set the date for the Singularity — representing a profound and disruptive trans-formation in human capability — as 2045. The nonbiological intelligence created in that year will be one billion times more powerful than all human intelligence today.* [8:136]

While Kurzweil's criteria for constituting the realization of the singularity remain rather vague, the promised prospects are boundless. In the opening lines of his book Kurzweil announces that all the magic described in the *Harry Potter* novels will soon be technologically available [8:4].

> The Singularity will allow us to transcend these limitations of our biological bodies and brains. We will gain power over our fates. Our mortality will be in our own hands. We will be able to live as long as we want (a subtly different statement from saying we will live forever). We will fully understand human thinking and will vastly extend and expand its reach. By the end of this century, the nonbiological portion of our intelligence will be trillions of trillions of times more powerful than unaided human intelligence. [8:9]

Kurzweil's book *The Singularity is Near* includes new reflections on the cosmological significance of earthly events and the ultimate goal of life in the universe. He also adopts Vinge's analogy of the event horizon of black holes: "Just as we find it hard to see beyond the event horizon of a black hole, we also find it difficult to see beyond the event horizon of the historical Singularity." [8:487]

The Russian Internet billionaire Dmitry Itskov's *2045 initiative* is also strongly influenced by Kurzweil's futurology. Its research program, launched in 2011, seeks to transfer a human personality into computer memory by the year of singularity. Itskov has named the intermediate stages avatars A-D, in reference to Hindu mythology [12:77-94].

Vernor Vinge and Ray Kurzweil use their understanding of singularity to canonically define its various qualitative elements.

- John von Neumann and Irving Good are the designated authors.
- The singularity entails a radical and rapid change.
- It is a consequence of the evolutionary development of life.
- It is determinate, it will occur in any case.
- It is connected with the development of super-intelligent computer systems.
- Humanity can participate via merging with computers.
- Predictions regarding what happens after the moment of the singularity are not possible.
- The singularity enables human immortality.
- The cosmological and technological concepts of singularity complement one another.

In this context, Frank Tipler's and Eliezer Yudkowsky's designs offer the extreme opposite poles of the techno-prophetic spectrum: Tipler at the Christian end, Yudkowsky at the atheistic – with Vinge and Kurzweil oscillating somewhere in between.

III. The Cultural Context of the Singularity Idea

How can one analyze a temporal concept like the singularity? Is it even a technological fact based on legitimate calculations? Many in the technophile scene have their doubts [48]. Social, psychological and cultural factors play a central role in the proclamation of a coming technological revolution. Nick Bostrom acknowledges that since the 1940s, the prognoses for the realization of artificial intelligence have slid backwards year after year, usually remaining about twenty years away: "Two decades is a sweet spot for prognosticators of radical change: near enough to be attention-grabbing and relevant, yet far enough to make it possible to suppose that a string of breakthroughs, currently only vaguely imaginable, might by then have occurred." [30:4]

At the beginning of the 1990s, MIT professor Pattie Maes noticed that most of her male colleagues were fascinated by the idea of soon being able to upload their brains into computer memory, thus overcoming death. Indeed, they believed that the advent of the first superhuman intelligence would immediately solve the problem of immortality – if only one could survive until this decisive moment. In 1993 Maes spoke about her systematized observations on her colleagues' predictions at the *Ars Electronica* meeting in Linz (Austria), in a presentation titled "Why Immortality is a Dead Idea". Astonishingly, what she found was that almost all futurists predicted the arrival of immortality within their expected lifetimes. No matter when the predictions were made or how old the actors were, the anticipated salvation would conveniently arrive around age 70 [31:206].

Stuart Armstrong and Kaj Sotala from MIRI have studied the systematics of AI prediction with scientific precision. They analyzed 257 temporal predictions for the arrival of a universal AI (the scope of the question was broader than in Pattie Maes' work, which only focused on predictions of AI in terms of immortality). Armstrong and Sotala's research found significant uncertainty in predictions about AI. This concerns both prediction methods (including apparent regularities, philosophical arguments, perceived status of the expert) and targets, which ranged from as little as six to more than 75 years. Particularly enlightening was the result that estimates by AI experts had exactly the same variance as those of non-experts (journalists, publicists, or prognosticators from outside the field). In both groups, the majority target a period 15 to 20 years in the future (which confirms Bostrom's impression). Researchers can thusly benefit from their own predictions, receiving research funding or appreciation as renowned experts [32:3-19].

If a revolutionary event is generally expected to occur in about two decades, regardless of when or by whom this prognosis was made, then it becomes important to consider the social dynamics and legitimacy of futurology more closely. What elements make up the singularity as a temporal concept? Firstly, it is justified by laws of progress and acceleration. The singularity also obviously constructs a threshold or boundary – which echoes the idea of the frontier that is so present in American cultural history (including its adaptations in the science fiction genre). As Armstrong and Sotala explain, the status of being a futurologist often serves to legitimize the predictions made. This "charisma of an eschatological prophet", as the sociologist Max Weber would put it, needs to be examined in greater detail.

Not all post- and transhumanists justify the appearance of the singularity – or of AI generally – by revelations or prophecies; they tend to refer to a mathematical theory of progress (e.g. *Moore's Law*). The assumption that progress is subject to a particular law rather than random chance is often attributed to the 17th-century English philosopher Francis Bacon. However, a general doctrine of progress was actually formulated during the Late Enlightenment through positivism. On the one hand, this philosophy considers scientific and technological developments to be bound by the law of progress. Yet on the other, it also identifies this progress as inherently linked to that of morality and politics. Within this framework, history – like the history of religious salvation before it – was understood as the universal history of all humanity [33:21-22]. On the threshold of the 18th century the French philosophers Fontenelle and Abbé de Saint Pierre first devised the general doctrine of progress. Fontenelle believes that progress was necessary and guaranteed, since following generations would always benefit from the knowledge and mistakes of their predecessors. Abbé de Saint-Pierre, in his vision of social and moral advancement, combined the progress of knowledge with the idea of increasing human happiness [34:98-143].

In 1795 the French philosopher Antoine Marquis de Condorcet published his *Esquisse d'un Tableau historique des progrès de l'esprit*

humain. This significantly impacted the English utilitarians, for whom the progress of the human race and the individual was attributed to the law of nature. History – as David Hume and Adam Ferguson agreed – was now to be pursued as a branch of mathematics. It would investigate the causal chain of historical progress, which Turgot and Auguste Comte conceived of mechanistically, in order to better shape the future [35]. At the same time, individual actions became interpreted as part of a larger historical process. A view became widespread that the progress of past ages not only ensured future progress but would also gradually accelerate. As Edward Gibbon predicted in his *History of the Decline and Fall of the Roman Empire*, it would be "infinitely slow in the beginning, and increasing by degrees with redoubled velocity" [36:169]. According to Francis Bacon, Adam Smith, Immanuel Kant and many other thinkers of that time, the fact of accelerating progress was undeniable for technical and scientific fields. In this way, they deduced the law of progress from both the observation of the past and their hopes for the future.

The inclusion of the utopian perspective as legitimation for the incessant acceleration of progress is a characteristic feature of every such ideology. 200 years before Kurzweil, the assumption that progress would accelerate enormously in the future already served two crucial purposes. Not only were benefits expected to materialize during one's lifetime, but also everyone who was fully committed to the process could count on taking part. A double motivation to believe and support therefore surrounds today's expectations of the singularity just as it did Enlightenment utopias [33:381-383]. The idea of ever-increasing acceleration is also due to another cultural source. The German scholar of religion Ernst Benz points out that such incessant acceleration was a characteristic of Christian salvation history. The discovery and Christianization of America was also shaped by these eschatological expectations. Columbus – convinced of the approaching end of the world – saw India (i.e. America) as Satan's last empire to be proselytized. According to Benz, the fundamental idea of accelerating progress is contextualized by the subjective expectation of salvation – that ultimate Christian goal. This is further nourished by New Testament reports and the visions recorded in the Book of Revelation or by the apostle Stephen. This longing for acceleration is particularly associated with the American theory of progress, which has often understood the unfolding history as part of God's plan for the coming of the promised land [37:18-21].

In addition to this idea of increasing acceleration, another crucial allusion to American cultural history is found in the understanding of the singularity as the last frontier. Since Puritans settled Massachusetts in the 17th century, the *frontier* has marked the border of the civilized and moral world against the wilderness, represented by the disordered chaos of the indigenous tribes of North America. The Christian-colonial sense of missionary purpose was further reinforced in the 1840s, when expansionist tendencies in American politics (particularly the annexation of Texas) were merged with the project of spreading freedom and democracy. They believed it be the manifest destiny of God's chosen American people to sow progress, civilization and freedom in the wild and untamed vastness of the continent. [38:69-77]

After the geographical frontier disintegrated with the settlement of the West and the extermination of most indigenous peoples, the frontier's metaphorical significance grew in other areas of society, especially science. Francis Bacon had already portrayed the researcher as a pioneer who ventured into undiscovered worlds. However, it was Vannevar Bush, the scientific advisor to President Franklin D. Roosevelt, who immortalized the metaphor for American academia in 1945 with his report *Science – the Endless Frontier*. In this document, Bush proposes guidelines for promoting science in the United States, which led, among other things, to the establishment of the *National Science Foundation*.

> It has been basic United States policy that Government should foster the opening of new frontiers. It opened the seas to clipper ships and furnished land for pioneers. Although these frontiers have more or less disappeared, the frontier of science remains. It is in keeping with the American tradition – one which has made the United States great – that new frontiers shall be made accessible for development by all American citizens. [39:46]

From John F. Kennedy to George W. Bush and Barack Obama, the metaphor of the intellectual frontier has continued to play an important role in American scientific policy [39:29-155].

As conceptualized by Vinge, Yudkowsky and Kurzweil, the singularity is based on this important metaphor of the *endless frontier*. The singularity in the sense of an event horizon of black holes remains impenetrable and insurmountable for humans. But for artificial intelligence, the singularity would be the beginning of an unlimited expansion into the universe, in which humans are also allowed to participate.

As already indicated, this perception of singularity as the last boundary to be overcome has been popularized by numerous adaptations in science fiction stories and films. This genre establishes the connection between the spatial and the scientific metaphors – i.e. human civilization finally surpasses the last frontier of human knowledge as it moves into space. One particular catalyst for such ideas was the scientific work of the Princeton physicist Gerard O'Neill (1927-1992), who from the 1970s onwards presented numerous technical designs for colonizing space, the *High Frontier* [40:168-208].

In the fifth *Star Trek* movie, *The Final Frontier* (1989), Captain Kirk is forced to overcome the "Great Barrier" in the center of the Milky Way on his spaceship, in order to seek God on a mythical planet. The first two *Star Trek* television series (1966-1969, 1987-1994) always prefaced their opening credits with the magic words: "Space, the final frontier." Less fantastically, in *The Black Hole* (1979) Maximilian Schell, playing the brilliant but unscrupulous scientist Dr. Hans Reinhardt, tried to convince a stranded spaceship crew that the ultimate truth, God, and eternal life in a world beyond physical laws waiting for them on the other side of a black hole. The scientist then transforms the recalcitrant members of his own crew into mindless cyborgs. At the end of the film, the surviving heroes actually fly through a Dante-inspired, hellish inferno and then glide behind an angel into a paradise flooded with light. In the 20th century Western heroes thus seamlessly transform into space heroes. The overcoming of the *final frontier* – the singularity of black holes – becomes the heroic enterprise of white men, whether these come equipped with heterogeneous accents like a fist-swinging macho (James Tiberius Kirk) or as possessed geniuses (Max Reinhardt) [40:139-167].

There is also no question that the temporal aspect of the singularity is influenced by the Christian end of days. The overcoming of old age, illness and death corresponds to the Christian expectation of salvation (especially in Tipler's vision of a resurrection of the dead). However, the essential analogy to the Christian apocalypse remains ambiguous: the singularity is neither the result of a continuous and positive development of progress nor of total annihilation. Like the Christian history of salvation, the concept connects the downfall of human beings with the certainty of a post-singular promise: death followed by resurrection.

Christian and singularity prophecies share another important structural feature: that signs reveal the imminence of this end. The Revelation of John lists many apocalyptic elements (prophecies, destructions, sacrifices, testimonies) that occur before the final battle against Satan and the Last Judgment (Rev 4-20). In his three futuristic books, Kurzweil in particular develops an increasingly precise description of milestones that will precede the singularity, including an evaluation of his own earlier predictions.

Unlike in Vinge and Kurzweil's version of the singularity, Christian writings do in fact provide a precise description of the post-apocalyptic

period: The New Jerusalem is described in great detail (Rev 21-22). In the Christian and Jewish traditions, salvation is dependent on God's judgment of one's moral conduct. According to all posthumanist authors, the singularity makes immortality available to every living human being. This idea of universal salvation for all human beings is only found explicitly amidst the Unitarian Universalists, who in their 1803 Winchester Profession proclaimed that the one Holy Spirit of Grace " … will finally restore the whole family of mankind to holiness and happiness." [41] The fact that Kurzweil grew up a Unitarian should not be overestimated at this point, as other advocates of singularity reach the same conclusion.

One final relevant aspect for this analysis lies in the role of the heralds of the singularity. Here Ray Kurzweil stands out, both in terms of his claims and the colorfulness of his autobiographical self-representation. Although he can reflect on a number of inventions and awards accomplished during the 1970s and 1980s, he has not yet been able to utilize the Internet and digitalization to achieve any technological breakthroughs. Compared to today's Internet entrepreneurs, he is only a lightweight with his estimated assets of $27 million. Naturally, the question then arises as to why Kurzweil in particular is called upon to praise singularity and system-changing technologies when he himself apparently has been alrgely unable to benefit at all from those trends. For him, the construction of a charismatic genius was even more important.

While in kindergarten Kurzweil was already aware of his own destiny: "At the age of five, I had the idea that I would become an inventor. I had the notion that inventions could change the world." [8:1] He built his first robots at the age of eight. He believes not only that he foresaw technological innovations, but in his 1990 book *The Age of Spiritual Machines* he also claims to have predicted the demise of the Soviet Union (1990/91) due to decentralized communication networks [6:446-447]. The documentary film *Transcendent Man. The Life and Ideas of Ray Kurzweil* from 2009 is a brilliant example of modern hagiography: a "legend of saints". In it, Kurzweil is accompanied by the film crew on his worldwide lectures. His followers, such as actor William Shatner, singer Stevie Wonder or former Secretary of State Colin Powell, praise him hyperbolically on camera. One immediately notices Kurzweil's trauma at losing his father, as well as his obsession with reaching the age of singularity through taking 150 vitamin pills daily. Apart from the mantra that the singularity is near and will change everything, the film does not contain much substance, and actually offers no in-depth discussion of the concept [12:100]. The films *The Singularity* (2012) by Doug Wolens and the film *The Singularity is Near* (2010), produced by Kurzweil himself, did not focus on the figure of Kurzweil. However, they were able to further popularize this futuristic scenario. The continual acceleration that Kurzweil promotes in his three futurological monographs offers a recognizable parallel to religious prophecy. This phenomenon is uncannily familiar in the history of religion, especially regarding the lack of fulfilled predictions. This feature is particularly striking in Kurzweil's work, since all other post- and transhumanist thinkers of recent decades relativize or tone down their predictions, or else broaden their temporal horizons. One might even be tempted to suggest a new *Law of Increasing Disappointment*, whereby the only things growing exponentially in transhumanism are the predictions themselves.

As a prophetic figure, Kurzweil also claims a special position: Vannevar Bush declared the endless frontier of the sciences in 1945. Kurzweil proclaims the end of this period of searching for knowledge will occur precisely one century later in the year 2045. He thus situates himself as the last prophet of the end times, the seal of the prophets. No further advances in prophecy could surpass Kurzweil's visions: when the singularity arrives humankind's time will be finished, and the fate of the universe will be decided.

IV. Conclusion

In the movies *Terminator* (1984) and the *Matrix* trilogy (1999-2003), a powerful artificial intelligence seeks to exterminate or enslave the (last) humans. Similarly, in marked contrast to the naive futurologies of Kurzweil and the transhumanists, postsingularity science fiction predominantly follows the tradition of dystopian cyberpunk literature [42:124-125]. Elaine Graham notes that more recent science fiction is increasingly blatant in dissolving the boundary between religion and science. The secular and the sacred; the human being and God; faith and knowledge; these all appear increasingly less as polar opposites, but rather now merge and blur in a post-secular era [43:362]. Dystopian visions no longer propagate the overcoming of a religious superstition by a rationalist techno-culture, but rather now celebrate the fusion of these two spheres.

For example, in Rudy Rucker's novel *Postsingular* of 2007 [44], a Christian fundamentalist US president seeks to transform the entire Earth into a *virtual earth (Vearth*) with the help of a computer scientist using nano-robots. He sees this transformation as the realization of biblical prophecy via restoration of the Garden of Eden, where suffering, war, and death are banished, and life is completely coordinated. Rucker reveals that this desire stems from trauma experienced by the computer scientist during his youth, when he lost his friend in an accident. In *Postsingular*, the interests of Christian and cybernetic fundamentalism overlap in their hatred of both women and creation in general [44] – [42:40-45].

It seems obvious that the prophecy of singularity is strongly influenced by cultural and religious ideas. The assumption of laws of progress, as well as the steady acceleration of progress claiming universal validity for the entire history of humankind, can all be traced back to an Enlightenment striving for perfection. But what is new in the singularity is Vinge and Kurzweil's idea introduced of an absolute and impenetrable limit to this progress: the singularity as the *last frontier*. The term repeats semantics from the physics of black holes, as well as their popularized representations in literature and film. Even more astonishing is that the concept of singularity allows a religious teleology to creep into post- and transhumanism, which 15 years ago was dismissed as exotic. This occurs first and foremost structurally, as the entire history of earthly life heads towards a moment of salvation. In concrete terms this happened when Ray Kurzweil bluntly adopted Frank Tipler's notion of the complete colonization of the universe, culminating in the realization of God [19].

Actually, the British science fiction author Charles Stross had already anticipated the conclusion of my analyses in his short story *Accelerando* (2005) with a few words. After extensive debates about the nature of the singularity, one of the two characters sums up laconically:

> "Is not happening yet," contributes Boris. "Singularity implies infinite rate of change achieved momentarily. Future not amenable thereafter to prediction by presingularity beings, right? So has not happened … Singularity is load of religious junk. Christian mystic rapture recycled for atheist nerds." [45:184]

Acknowledgment

I would like to thank Ralf Hoffmann, Ali Jones and Paul Knight for their kind support and advice.

References

[1] *Godling's Glossary* 1998. Quoted by Anders Sandberg. Accessed: 01/15/2021 (Online). Available: aleph.se/Trans/Global/Singularity/index.html.

[2] H. Moravec, *Mind Children. The Future of Robot and Human Intelligence*, Harvard: Harvard UP, 1988.

[3] F. J. Tipler, *The Physics of Immortality. Modern Cosmology, God and the Ressurection of the Dead*, New York: Anchor, 1995.
[4] M. L. Minsky, *Computation. Finite and Infinite Machines*, Prentice Hall: Englewood Cliffs, 1967.
[5] C. Keller, *Buildung Bodies. Der Mensch im biotechnischen Zeitalter. Reportagen und Essays*, Zürich: Limmat, 2003.
[6] R. Kurzweil, *The Age of Intelligent Machines*, Cambridge: MIT Press, 1990.
[7] R. Kurzweil, *The Age of Spiritual Machines. When Computers Exceed Human Intelligence*, New York: Viking Press, 1999.
[8] R. Kurzweil, *The Singularity is Near. When Humans transcend Biology*, New York: Penguin Books, 2005.
[9] R. Kurzweil, T. Grossman, *Fantastic Voyage. Live Long Enough to Live Forever*, New York: Rodale Books, 2004.
[10] R. Kurzweil, T. Grossman, *Transcend: Nine Steps to Living Well Forever*, New York: Rodale Books, 2010.
[11] B. Ptolemy, *Transcendent Man*, USA 2009 (Movie).
[12] T. Wagner, *Robokratie. Google, das Silicon Valley und der Mensch als Auslaufmodell*, Köln: PapyRossa Verlag, 2015.
[13] S. Bringsjord, A. Bringsjord, P. Bello, "Belief in the Singularity is Fideistic," in *Singularity Hypothesis. A Scientific and Philosophical Assessment*, A. H. Eden et al. Ed. Berlin: Springer, 2012, pp. 394-408.
[14] R. Heil, "Transhumanismus, Nanotechnologie und der Traum von Unsterblichkeit," in *Visionen der Nanotechnologie*, A. Ferrari, S. Gammel Eds. Heidelberg: Akademische Verlagsgesellschaft, 2010, pp. 25-49.
[15] S. Hawking, R. Penrose, "The Singularities of Gravitational Collapse and Cosmology," in *Proceedings of the Royal Society*, vol. 314, pp. 529-548, 1970.
[16] J. D. Barrow, F. J. Tipler, *The Anthropic Cosmological Principle*, Oxford / New York: Oxford University Press, 1986.
[17] F. J. Tipler, *The Physics of Christianity*, New York: Penguin Books, 2007.
[18] F. J. Tipler, "The Laws of Physics Say The Singularity is Inevitable!" Interview (Video) with Socrates / Nikola Danaylov of the SingularityWeblog 10/29/2013. Accessed: 01/15/2021 (Online). Available: https://www.singularityweblog.com/frank-j-tipler-the-singularity-is-inevitable
[19] O. Krüger, *Virtual Immortality. God, Evolution and the Singularity in Post- and Transhumanism*. New York: Columbia University Press, 2021.
[20] S. Ulam, "John von Neumann 1903–1957," in *Bulletin of the American Mathematical Society*, vol. 64, pp. 1-49, 1958.
[21] V. Vinge, "First Word," in *Omni*, January, p. 10, 1983.
[22] V. Vinge, "The Coming Technological Singularity. How to Survive in the Post-Human Era." 1993. Accessed: 01/15/2021 (Online). Available: http://mindstalk.net/vinge/vinge-sing.html.
[23] V. Vinge, "Technological Singularity," in *The Transhumanist Reader*, M. More, N. Vita-More Eds. Chichester: Wiley-Blackwell, 2013, pp. 365-375.
[24] E. S. Yudkowsky, "Singularitarian Principles. Version 1.0.2. Extended Edition." 2000. Accessed 01/15/2021 (Online). Available: https://web.archive.org/web/20070613190005/http://yudkowsky.net/sing/principles.ext.html#preface.
[25] E. S. Yudkowsky, "Singularitarian Principles. Version 1.0.2. Extended Edition." 2000. Accessed 01/15/2021 (Online). Available: http://yudkowsky.net/obsolete/principles.html.
[26] H. Moravec, *Robot. Mere Machine to Transcendent Mind*, Oxford / New York: Oxford University Press, 1999.
[27] M. L. Minsky, "Marvin Minsky on Singularity." Interview with Socrates / Nikola Danaylov of the SingularityWeblog, 07/12/2013. Accessed 01/15/2021 (Online). Available: https://www.youtube.com/watch?v=3PdxQbOvAlI.
[28] H. Moravec, "Rise of the Robots – The Future of Artificial Intelligence," in *Scientific American*, 03/23/2009. Accessed 01/15/2021 (Online). Available: https://www.scientificamerican.com/article/ rise-of-the-robots.
[29] D. Proudfoot, "Software Immortals: Science or Faith?" in *Singularity Hypothesis. A Scientific and Philosophical Assessment*, A. H. Eden et al. Eds. Berlin: Springer, 2012, pp. 367-394.
[30] N. Bostrom, *Superintelligence. Paths, Dangers, Strategies*. Oxford: Oxford University Press, 2014.
[31] R. Brooks, *Flesh and Machines. How Robots will Change us*, New York: Vintage, 2002.
[32] S. Armstrong, K. Sotala, "How We're Predicting AI – or Failing To." 2012. Accessed 01/15/2021 (Online). Available: https://intelligence.org/files/PredictingAI.pdf.
[33] D. Spadafora, *The Idea of Progress in Eighteenth-Century Britain*, New Haven: Yale University Press, 1990.
[34] J. B. Bury, *The Idea of Progress. An Inquiry into Its Origin and Growth*. New York: Dover Publications, 1955.
[35] J.-A.-N. Marquis de Condorcet, *Esquisse d'un Tableau historique des progrès de l'esprit humain*, Paris: Vrin, 1970.
[36] E. Gibbon, *The History of the Decline and the Fall of the Roman Empire*, vol. 4, London: Methuen, 1925.
[37] E. Benz, "Akzeleration der Zeit als geschichtliches und heilsgeschichtliches Problem," in *Abhandlungen der Geistes- und Sozialwissenschaftlichen Klasse der Akademie der Wissenschaften und der Literatur*, vol. 2, pp. 1-53, 1977.
[38] J. D. Torr, *The American Frontier*, San Diego: Greenhaven Press, 2002.
[39] L. Ceccarelli, *On the Frontier of Science. An American Rhetoric of Exploration and Exploitation*, East Lansing: Michigan State University Press, 2013.
[40] M. W. Kapell, *Exploring the Next Frontier. Vietnam, NASA, Star Trek and Utopia in 1960s und 1970s American Myth and History*, New York: Routledge, 2016.
[41] Unitarian Winchester Profession, 1803/1899. Accessed 01/15/2021 (Online). Available: https://uudb.org/articles/winchester.html.
[42] J. Raulerson, *Singularities. Technoculture, Transhumanism, and Science Fiction in the Twenty-first Century*, Liverpool: Liverpool University Press, 2013.
[43] E. L. Graham, "The Final Frontier? Religion and Posthumanism in Film and Television," in *The Palgrave Handbook of Posthumanism in Film and Television*, M. Hauskeller, T. D. Philbeck, C. Carbonell Eds. New York: Palgrave, 2015, pp. 361-370.
[44] R. Rucker, *Postsingular*. New York: Tor Books, 2007.
[45] C. Stross, *Accelerando*. London: Orbit, 2005.
[46] I. J. Good, "Speculations Concerning the First Ultraintelligent Machine," in *Advances in Computers*, vol. 6, pp. 31-88, 1965.
[47] I. J. Good, "The Social Implications of Artificial Intelligence," in *The Scientist Speculates: An Anthology of Partly-Baked Ideas*, I. J. Good Ed. London: Basic Books, 1962, pp. 192-198.
[48] B. Cantwell Smith, The Promise of Artificial Intelligence. Reckoning and Judgement. Cambridge: MIT Press 2019.

Oliver Krüger

Oliver Krüger (*1973) studied Sociology, Classical Archeology and the Religious Studies at the University of Bonn (Germany) from 1994-1999, graduated with a Master in Sociology. In 2003 he received his Ph.D. in Religious Studies from Bonn University. From 2002-2005 he was a research fellow and lecturer at the collaborative research center *Dynamics of Rituals* at the University of Heidelberg. From 2005 to 2007 he was a Visiting Fellow and Visiting Lecturer at the *Center for the Study of Religion* at Princeton University. In 2007 he became Professor for Religious Studies Fribourg University (Switzerland). He served as President of the Swiss Society for the Study of Religion from 2011-2014 and in 2012/13 he was fellow at the Center for Advanced Studies *Morphomata* (University of Cologne). Major publications are: *Virtual Immortality. God, Evolution and the Singularity in Post- and Transhumanism*. New York: Columbia University Press 2021; *Die mediale Religion*, transcript: Bielefeld: transcript 2012; "Gaia, God, and the Internet – revisited. The History of Evolution and the Utopia of Community in Media Society," in: *Online – Heidelberg Journal for Religions on the Internet*, vol. 8, 2015.

Can AI Help Us to Understand Belief? Sources, Advances, Limits, and Future Directions

Andrea Vestrucci[1,2]*, Sara Lumbreras[3], Lluis Oviedo[4]

[1] Graduate Theological Union (USA)
[2] University of Geneva (Switzerland)
[3] Pontifical University of Comillas (Spain)
[4] Pontifical University Antonianum (Italy)

Received 17 April 2021 | Accepted 19 June 2021 | Published 3 August 2021

Abstract

The study of belief is expanding and involves a growing set of disciplines and research areas. These research programs attempt to shed light on the process of believing, understood as a central human cognitive function. Computational systems and, in particular, what we commonly understand as Artificial Intelligence (AI), can provide some insights on how beliefs work as either a linear process or as a complex system. However, the computational approach has undergone some scrutiny, in particular about the differences between what is distinctively human and what can be inferred from AI systems. The present article investigates to what extent recent developments in AI provide new elements to the debate and clarify the process of belief acquisition, consolidation, and recalibration. The article analyses and debates current issues and topics of investigation such as: different models to understand belief, the exploration of belief in an automated reasoning environment, the case of religious beliefs, and future directions of research.

Keywords

Artificial Intelligence, Beliefs, Religion, Computational Model, AI Limits.

DOI: 10.9781/ijimai.2021.08.003

I. Introduction

COGNITIVE science has tried since its inception to offer reliable models on how human mind works. One of these models is built on a representational approach. In this approach, ideas and mental accesses to reality are viewed as representations, and the operations and processes that result in thoughts and decisions are interpreted in computational/algorithmic terms.

Other models present alternative explanations for mind's operations different from a computational model. These other models are broadly labelled as "externalists". They assume that the computational model is not able to fully describe the complexity of mental phenomena. According to this perspective, mental phenomena involve factors that are irreducible to computational/algorithmic terms, and there are better candidates to explain some specific aspects of human cognition – for instance, the so-called embodied, embedded, enacted, and extended ("4e") theories of mind [1].

Some recent developments in robotics and information theory have increased the richness of these perspectives, but also the plurality of interpretations. An interesting example are theories that try to define consciousness as a sort of integration measure in the context of information theory. In the work of Tononi [2], several metrics are proposed to compute consciousness as a measure of how integrated information is in a system. Integrated information theory has provided formulae for *Phi* or this integration of information, which measures the level of feedback and interaction between the components of a system. This purely mathematical view has been criticized for implying panpsychism (all systems would be conscious to some degree) and for being non-functionalist (the theory does not address the functions of consciousness or their implications).

Another example has involved robots with peripherals such as sensors that receive inputs from the world and actuators that have an impact on it. In some cases, these robots have been able to derive models of themselves from the information received from outside. This has been described as *embodied* and *embedded* systems [3]. However, it is unclear whether embodied cognition or even embodied AI might provide a convincing model to represent human mind and cognition [4].

The process of believing is a good test for computational modelling, and it invites to develop more sophistication models. Believing is a common human experience: everybody holds beliefs of different kind, related to several life contexts, with distinct ranges and applications, from simple ones – such as "I believe the weather today will be good" – to the more engaging and abstract ones – such as "I believe that my life is meaningful".

The process of believing increasingly occupies a central role in the research. This might appear as a change from previous approaches. In fact, although the believing process has long been an object of interest for epistemology, cognitive sciences, and philosophy of mind, it has often been given a secondary, or "lower", status, in contrast to "higher" or "stronger" cognitive attitudes or faculties such as "knowledge", "reason", or "intellect" [5]. This peculiar status of belief is connected with the *probability* of the truth of beliefs. Beliefs are not necessarily

* Corresponding author.
E-mail address: avestrucci@gtu.edu

true: they can be true (or false), or they can be more true than false (or more false than true), i.e., they come in various degrees of strength, certainty, and confidence about their truth [6], [7]. In other words, a belief has a truth value that is – more, or less – probable. Thus, since belief is not necessarily veridical, but only probably veridical, this aspect of human cognition has been considered to be secondary to other epistemological notions (e.g., knowledge, especially when knowledge is equated with what justifies belief [8, ch. 9]), or belief is required to satisfy some specific conditions in order to enjoy the same status of other cognitive outputs [9].

However, the probabilistic status of the truth value of belief can be an important resource, because it offers a good environment to test to what extent which computational modelling better helps to understand how beliefs arise, stabilize, and even vanish. Several proposals for computational models of belief present probabilistic estimates for belief truth value. From its very beginnings, Artificial Intelligence (AI) has provided such models. Inference engines apply logical rules to an existing knowledge to deduce new facts. Bayesian networks are able to incorporate the probabilities of facts in order to derive the probabilities of other facts related to them [10]. The present article outlines some analyses and explorations of the computation of the probability of beliefs via the application of automated reasoning programs (section IV); it also discusses whether all beliefs can be object of computation and translated into algorithmic structures, or if there are subsets of beliefs that might not be interpreted in probabilistic terms, for instance believing in values or believing in a transcendent, divine being. Fodor [11] offers some interesting arguments against a computational approach to beliefs in values, due to the character of "generality" of such beliefs in contrast with other processes which are much easier to compute. Concerning religious beliefs, they could be presented in probabilistic terms, but their transcendent status might question such codification [12].

Recent developments in AI, such as deep learning (a type of machine learning based on multiple layered artificial neural networks), have increased the expectations that AI could help us to better understand the functioning of our own minds, and thus to fill some gaps that seems to affect current computational models for cognitive processes. This might include the formation and constitution of beliefs. For instance, AI systems built on pattern recognition and machine self-learning manage to achieve tasks that could come closer to some aspects of the believing process, such as believing in something as consequence of a recurrent pattern of events, or believing as the result of learning from new data. This might purport the idea that AI systems could work in a way that is close to human mental processes, and hence they can assist in discerning about belief formation, development, confirmation, or negation.

The present article aims to explore the hypothesis that such advances in AI could help to discern in a more accurate way to what extent develop the research on the computability of mental activities such as believing process. To do so, the article analyzes recent literature on how, and to what extent, AI systems might contribute to our modelling and understanding of belief. The study of recent computational approaches to believing processes might cast a new light on the philosophical, logical, psychological, and cognitivist perspectives on belief. Section II analyzes different models of belief and believing process. In section III we seek to detect the challenges of the computational approach and to discern the extent of its heuristic potentials. Section IV deepens some recent analyses and explorations of beliefs in automated reasoning environments, with specific reference to the formalization of belief in doxastic logic, the applications of automated theorem provers, and the assessment of the skepticism about the translatability of beliefs into machine syntax. Section V deepens how a specific set of beliefs, religious beliefs, may benefit from a computational approach, and to what extent these beliefs are irreducible to such approach. Finally, section VI outlines and discusses three directions for the future of the research on the intersections between computational modelling and the extent and improvement of our understanding of belief.

II. Modelling the Process of Believing

In this section we analyze models of belief and believing process that consider developments in computational study. Such models might foster a positive interaction between philosophy of mind and cognitive psychology, limit the risk that beliefs are studied in separate compartments, and increase the communication between fields – for instance, the integration of recent advances in the epistemology of beliefs into cognitive psychology research [13], [14], [15], [16].

At least six models have been proposed in recent years to describe the structure and the dynamic of beliefs. These six models are: the credition or "functions model" by Angel and Seitz [17] [18]; the "stages model" by Connors and Halligan [19]; the "network model" by Castillo *et al.* [20]; the "complex system model" by Lumbreras and Oviedo [21]; the "conversion model" by Smith [22]; and the "dimensions of faith model" by Donaldson [23]. We present a summary of these six models to allow for a brief comparison and to assess their computational features.

A. Creditions as Processes of Believing

The first model proposes a self-organizing system, with strong neurological roots, based on four functions: "enclosure" or integration of some basic units – perceptions, ideas – into an existing network in which they are accommodated; "converter function", which establishes belief's application range or its influences on an action course; "stabilizer function", able to keep some constant and reliable appearance despite environment changes; "modulator function", regulating the interplay between cognition and emotions.

The credition function [24] allows the individual to trust her inner probabilistic representations, and acts at two different dimensions: cognition and emotion. Credition has the key function of guiding action by means of reciprocating feedback, which involves exploration and learning. Credition is therefore an essential cognitive process to understand the human mind and behavior. It is important to underline that credition belongs not only to the realm of cognitive processing but also to the domain of subjective experience. Thus, credition is not only calculated but also experienced, as will be discussed later.

Further developments in the original pattern [25] have proposed an integrated model that considers, first, the broadly assumed dual schema that distinguishes between perceptual (or immediate) and evaluative (or analytic) processes; and, second, the "hierarchical structure of belief representation" that distinguishes between a physical, an interpersonal, and a social level. These two dimensions can be integrated to provide a structure where perception gives place to representations through action and evaluation or selection; this happens at the three mentioned levels: physical, interpersonal, and social, in a nested hierarchical process.

The described model is complex as it integrates a former and a latter schema, or reads the former four functions through two dimensions: the first describes the dynamics that link perception to selection in forming beliefs; the second records different levels influencing that configuration. In this last version, the four basic functions can be followed through some cognitive processes taking place along two distinct axes: one axis moves from perception to belief formation as somewhat an internal process of cognitive elaboration; the second axis accounting for more external influences or interaction. In this way, general functions are represented or accomplished in these two dynamic sets or components.

B. Stages of Belief Formation

Based on studies on delusion, Connors and Halligan [19] assume a functional stance to explain beliefs in terms of representations that help a subject to regulate her own behavior. They establish a five stages model of belief formation:

1. Precursor or proto-belief, which may be triggered by sensation or event, by external information or social communication, and by introspection on pre-existing beliefs.
2. Search for meaning, which "involves explaining or accounting for the experienced precursor and accommodating it within the existing web of beliefs" and avoiding cognitive inconsistency [19, p. 7].
3. Candidate belief evaluation, a process that scrutinizes new possible beliefs, after testing their explanatory power and their congruence with formerly acquired beliefs; such process is often subjected to the influence of many biases and affective states.
4. Accepting or holding the belief, or conscious assumption of the new belief as true after all the required tests, thus giving the belief a relative stability.
5. Consequential effects of holding the belief, on how new beliefs influence the world observation and judgments; this happens through a re-configuration of the existing "web of beliefs" and their fixation in memory, and determines decisions and an action course.

This schema appears as quite lineal and consistent with psycho-cognitive observation, together with neurological data. A logical process is followed from the first step, when a "proto-belief" is forming in one's mind, through its filtering and confirmation, after unconscious or conscious tests and influenced – or even biased – by cognitive mechanisms present in human mind.

C. Beliefs as Self-Sustaining Networks

Castillo *et al.* [20] describe a belief as "a network of perceptual experiences that have something in common." Inspired by complex networks, in a similar way as how ecological systems transfer energy, they intend that a belief links experiences through a transfer of meaning. In that sense, such experiences can be elaborated as perceived changes that influence or help to solve a task. Their functions are recognized to be adaptive: beliefs help to make predictions, to constrain attention, and to bridge interruptions due to variability.

Also in this model, relevant experiences need to be coupled to former ones giving rise to beliefs. The ecosystems analogy is further explored to find similar dynamics in beliefs processes: autocatalysis, or positive feedback between different system levels; circular causality, or mutual influence between single elements and final outcome, reinforcing all the system; and centripetalism, "the idea that a network will attract resources into its circuit to sustain itself" [20], amplifying its relevance to confirm its acquired positions and to exclude conflicting ones.

This model allows to explain how the system reaches some stability despite fluctuations in the environment and some aspects characteristic of beliefs, such as storage, retrieval, and apparent agency. At the same time, this schema helps to understand how beliefs emerge and change. In analogy to how new systems emerge from dissipating gradients and reaching new equilibria, beliefs emerge to gain new meaning after different experiences interact and become coupled. Change is the result of perceived mistaken beliefs, often too locally linked. However, due to the self-enforcing nature of belief networks, conflicting evidence is not enough to explain change, which happens only when the old belief stops to be perceived as beneficial for the entire system, giving rise to a new network and more efficient coupling. This process points to a degree of order present in such systems, which can be conceptualized, as in other similar networks, always as something taking place spontaneously, i.e., as order emerging from chaos in living systems.

This new model takes advantage of cybernetic dynamics that count with a consistent tradition. As such, a belief is always seen as an ordered set able to link or couple experiences, and, by the same token, broader beliefs can be conceived as sets that are coupled in a systemic or coherent way, addressing some adaptive tasks, or covering some function in relationship with their environment.

D. Beliefs as Complex Systems

Recently, beliefs have also been likened to complex systems [21]. A complex system is a self-organized ensemble of reiterated, uncontrolled multiple interactions between a plurality of components. Examples of complex systems include Earth climate system, ant colonies, and the Web [26].

This analogy integrates some of the principles of belief systems as networks, and expends them to make sense of the dynamics of belief within the wider scope of complex systems. As mentioned, a complex system is an entity composed of many interacting components. Even if the components are relatively simple, the behavior of the system is difficult to predict due to the emergence of new phenomena in the system. Examples of complex systems can be found in a wide variety of context, from engineering to biology or finance, but, regardless of their specific context, all complex systems share the same properties, such as nonlinearity, emergence, spontaneous order, adaptation, and feedback loops [27]. The power of complex system studies lies in how general these properties are, and how they help explain very different phenomena. For instance, from a complex system perspective, the behavior of a flock of birds trying to advance and protect itself while avoiding internal collisions could be linked to that of a group of firms trying to develop their companies according to established business models while avoiding excessive competition.

From the complex system perspective, the main properties of belief are: being goal-oriented, openness, complexity, spontaneous order, and adaptation. The goals of belief have been studied in the literature under differing lights [28]. Belief serves at least three different purposes: it provides a model of the world and anticipates the consequences of action; it filters new evidence and establishes priorities for the decisions; and it defines what is important, what is a priority, what should or should not be done. Beliefs are open because they receive inputs from the interaction with the environments or with other individuals ([29], [30], [31]). They are complex because of the number of factors that influence them: personal features such as analytic cognitive style [32], feelings of superiority [33] or even parenting styles [34] have been shown to influence belief formation and change. Belief networks, as complex systems, are subject to nonlinear phenomena. Nonlinearity means that the same stimuli does not lead always to the same response. For instance, it takes more information to change beliefs than to confirm them. In belief networks, change is generally difficult but, in times of crisis, the change in one belief can spread to a large number of them. This is true for personal and for social beliefs, where the crisis dynamic could be explained as a paradigm shift [35]. Emergence means that new properties and structure originate from the system. In addition, beliefs organize in more or less consistent and related spheres of influence. Consistent belief systems have been described by some as "attractors", states towards which a system tends to evolve [36]. Finally, the adaptation property is displayed by belief systems as they evolve to fulfil their objectives of providing a model of the world, filtering experience and guiding action.

This model can be useful to integrate many of the properties of belief and to anticipate some of the phenomena that might emerge in a manner that can be subject to empirical testing. Still, this model is in its early steps and its development could lead to interesting insights.

E. Beliefs as Functional Maps

Aaron Smith [22] proposes a model more akin to religious beliefs and inspired on sequential models of religious conversion. His foundation lies in a functional or adaptive view of beliefs attuned to social connection, risk detection, and life navigation. He devises five components in that process: "concepts" including relevant representations and forms of recall; "computation" that works like an "engine of belief" or mechanism that process information to produce inferences; seven "iterative mechanisms", from "personal identification" through repetition and practice to reasoning and confirmation; "commitment" or personal assimilation, which includes unconscious reinforcement and rewards by promises and expectations; and, finally, "consequences" or positive effects in terms of social benefits or cultural innovation by trust.

This model is complex and not just sequential, since some loops and mutually enforcing means are present, especially at the model's core, i.e., the "computation" component, which includes several cognitive mechanisms working together to prompt beliefs, including religious beliefs.

F. Dimensions of Faith

For Steve Donaldson [23], faith is a general psychological attitude that consists in attributing a probability to the existence of something or to an event or expectation. As mentioned, believing means considering the probability for something to be true. This is a central cognitive function that contributes importantly to give sense to our world and to make decisions. Clearly, beliefs aim to be true. In order to confirm the truth of belief or to present evidence for it, several mechanisms may contribute (e.g., personal introspections, emotions of fear or pleasure, rational claims), and several factors may endanger the process, such as the lack of rational interest.

Religious beliefs are viewed as a type of belief that shares common elements or similar structures with other cognitive systems [37]. As such, by studying beliefs we acquire a heuristic tool that applies to a broader spectrum, from economy to emotions, from science to religion.

Donaldson classifies beliefs according to three levels of observation: primary or immediate; secondary or mediated by other means; and tertiary or resulting from reflection. He then establishes a scale to determine the levels of certainty and how beliefs that are "known for sure" are different from those that represent values.

The general idea in the six models that we mentioned above is that beliefs follow their specific logic and can be modelled based on some basic characteristics and relationships. All six proposals assume a functional stance: beliefs help us engage with the world and with our relationships in and with it. All six models apply a cognitive psychological framework, and in some cases the computational aspects emerge as a part of a complex system that includes many other dimensions or components. For instance, in both "network model" and "conversion model", cybernetics and computation appear as one stage of a global process. In these models, computation can play a role in beliefs formation, but believing cannot be reduced to just computational means or processes, or to algorithmic elaboration of collected information. The critical point is to what extent AI developments could help to shed more light on this aspect or could reveal some aspects still hidden in our ability to form and hold beliefs. This topic is analyzed in the next section.

III. Making Place for AI in the Believing Process

In this section we attempt to see if advances in the field of AI and its applications may help the research on belief and provide insight into the believing process, and whether AI is in competition, or in consonance, with the previously-analyzed models of belief. The first step is to distinguish for what AI might be relevant, and for what it might not play any role. This might beg the question about what lies beyond the realm of computation. Nevertheless, it might be useful to advance a proposal built on available views that clarify the affinities and distinctions AI and human cognition.

A good starting point is the recent book by Brian Cantwell Smith [38]. The book attempts to discern, through an in-depth knowledge of AI systems, what is specifically human in our way to know and to deal with the world. The central point is the distinction between the concepts of reckoning and judging. Reckoning is the type of calculation provided by current AI systems; judging is the human kind of decision making, based on the evaluation of circumstances, events, knowledge, and also beliefs. This "judging" is built on ethical values, existential insights and projects, wisdom, and the distinction between what is actual and what is possible, what is real and what is apparent.

According to Cantwell Smith, the cognitive capabilities of humans and AI systems are different, and they can hardly overlap. AI systems deal with discrete representations of the world that are processed through algorithms. On the other hand, humans engage directly with the world through non-linear representations and projects, and by formulating propositions that have non-dualistic truth value – that is, propositions that are neither 100% true nor 100% false. Amongst those propositions or statements there are the so-called "beliefs", statements that have a truth-value that is probabilistic.

Although some have proposed that embedded AI bridges the gap between what could be associated to Cantwell Smith's concepts of "reckoning" and "judging" [39], the mere inclusion of sensor inputs on a process might not be equivalent to cover the full spectrum of elements that constitute human believing process [40]. Could increasing the number of sensors lead to anything qualitatively different? Along the same lines, fuzzy logic has been used to formalize the probabilistic truth value of belief [41]. However, fuzzy logic might be considered to be not fully able to capture all traits of human believing process. In such process, the probability of a belief might also be connected to emotional states or prior beliefs. It would be needed a much more sophisticated fuzzy logic than, for instance, the one currently used in engineering contexts.

What lessons can we learn from this "cognitive" distinction between human and machine? Let us imagine a "believing machine", an AI system able to generate beliefs from a large number of data or information.

This hypothetical believing machine can be helpful to better assess some hidden processes in human belief formation. If we ask how this machine would work, we can formulate the following points:

- The believing machine would be a system able to collect all relevant information or inputs through pattern recognition. The machine would be able to distinguish between what is relevant and what is not for a belief formation. It would filter inputs and prior beliefs based on rules of contiguity and causality.
- The believing machine would decide through the right algorithms an output from the totality or a selection of the collected information. This task would need specific statistic tools, such as probability calculation. Hence, the system would act as a predictor, by using prior information to predict the outcomes of actions. As such, the machine would be a support for decision-making.
- The process of decision-making could make use of machine learning systems guided through positive and negative feedbacks from the application of acquired beliefs. Thus, the believing machine would filter new evidence and would evolve the belief system to fitting with prior beliefs (including the emotional investment of the subject) and new information.

All these processes could also be constitutive parts of human believing process.

However, contrary to the human situation, this machine would not engage with the beliefs it generates; it would be indifferent for it to believe in a belief or in a different one. On the other hands, we humans put much at stake in our beliefs, we are committed to (some of) them and such commitment influences the way we think, judge, and act in the world. Believing in the righteousness of a value *x* (for instance, eco-responsibility, or gender equality) affects our life in a very different way than not believing in the righteousness of this value *x*.

Moreover, our beliefs are constantly changing: they are confirmed or dismissed, enforced or discredited by ideas, experiences, relationships, and introspection. They demand reformulation and adaptation to new life situations. The process of recalibration of beliefs involves impressions, emotions, cultural and social circumstances, personal commitments, etc. In other words, beliefs are experienced by a subject and have an object – believing is a subjective experience, analogously to "qualia" (the subjective and conscious experiences, such as the sensation of cold or heat). As such, the believing process is deemed to be simultaneously a cognitive and an emotional process – as the credition model underlines. This makes difficult to compare humanly experienced belief to mechanical processes undertaken by an AI. Furthermore, the case made by Cantwell Smith regarding the distinction between machine and human cognitive capabilities might even introduce a second level of belief: believing that we humans and machines are radically different precisely as far as beliefs are concerned, and that this difference is positive and good for human flourishing.

Probably the believing machine would also need external assistance. For instance, pattern recognition requires a previous work of tagging by an operator who identifies and labels relevant objects or information. Even if statistical methods could be applied in its most sophisticated way, interpretation of the results would demand a further consideration and judgment. Moreover, feedbacks can be ambiguous and complex: some beliefs could result in double effect actions, and again some discernment based in judgment and a broader view would be needed – something close to what Cantwell-Smith calls "wisdom".

A machine able to generate beliefs could become a good heuristic tool. This machine could even improve at the point to incorporate a number of functions to the point of coming close to human believing process. However, it might be hard to imagine an algorithmic translation of the aspects of interpretation, judgment, and commitment that qualify human beliefs.

IV. Computation of Belief Via Doxastic Logic

In section II we analyzed some theories of belief that are conceived by using terminology borrowed from Computer Science such as network, modelling, system, etc. In section III we outlined some distinctions between human and machine cognitive capabilities by focusing on the formulation of beliefs and on how those beliefs impacts human life.

Now, it is time to analyze current attempts of positive interactions between belief and computer programs. Those attempts concern the application of automated theorem provers to assess the epistemic value of beliefs, that is, to calculate the probability for a belief to be true (or false), and thus to modify the belief in order to improve its probability of being true. The application of such programs to beliefs requires an intermediate step: the translation of a belief into a formal language that can be understood by the syntax of the machine. This formal language is provided by doxastic logic, the logic that deals with opinions and beliefs.

Doxastic logic is a subset of modal logic, the logic that formalizes possibility and necessity. Possibility and necessity are the "modes" of the truth-values of a proposition: a proposition can be necessarily true (or false), for instance once the proposition is demonstrated, or a proposition can be possibly true (or false), for instance before that the proposition is demonstrated. The link between modal logic and belief is the following: as stated in sections I and III, beliefs are expressed in propositions that have a truth-value that is probabilistic. Moreover, as already mentioned, the probabilism of beliefs' truth-value is precisely what distinguishes beliefs from other human cognitive attitudes and faculties. Given that a belief is expressed by a proposition that is probably true (or false), then it is possible for this proposition to be true (or false). Therefore, the proposition expressing a belief can be translated into the formal language of modal logic, since this is the logic that studies the possibility or necessity of the truth-values of propositions. "Doxastic logic" is the name of the field of modal logic that studies the formalization of beliefs.

The study of the logical formalization of beliefs dates back to the 1950s. One of the early most famous work on this is [42]. In this seminal work Hintikka applied possible world semantics to the logical study of knowledge and belief. Possible world semantics interprets possibility and necessity (the two operators of modal logic) as quantifiers over possible worlds: necessary is a proposition that is true in all possible worlds, and possible is a proposition that is true in some possible worlds. Thus, the application of possible world semantics to knowledge and beliefs is another way of saying that epistemic logic (the logic of knowledges) and doxastic logic (the logic of beliefs) are subsets of modal logic.

Since then, the scientific community has witnessed an exponential growth of the research on the extent and the limits of the logical study of beliefs [43]. The objects of this research are multiple: how to formalize the connection between beliefs and their premises or presuppositions; if, and how, the statements about beliefs can be axiomatized; the extension of the logical investigation from the mere content of belief to other connected topics, such as the purpose of belief, the consequences of believing something, and the justification of beliefs; (connected to the previous point) the way to recalibrate and correct beliefs via the interaction with other agents (believers) or the acquisition of new information; the logical treatment of the lack of belief, for instance in the sentence "there is something that I neither believe nor disbelieve" [44].

The logical study to beliefs has the worth of refining our understanding and insight on beliefs. This includes the clarification of some logical issues affecting belief or defining its epistemic specificity, for instance if belief is compared to knowledge. One of those logical issues concerns the possibility (or even the necessity) for beliefs to be inconsistent, that is, to entail a contradiction with its premises. For an overview of the varieties of inconsistency that occur in beliefs, see [45]. Moreover, the research has deepened the logic at the basis of the improvement, awareness, or progressive resolution of such logical issues of belief from the standpoint of the *believer* itself; this is a kind of dynamic epistemic (or doxastic) logic [46].

The development of the logical study of belief and of the processes of producing, enriching, and modifying beliefs provides a formalization of beliefs and believing processes. This formalization uses the operator *B* for "belief" and variables for subjects and objects of belief. A standard formalization is the following formula: $B a\gamma$, which reads: "The subject *a* believes γ". This formalization is in first-order doxastic logic, because the operator *B* ("belief") is applied to an object. However, doxastic logic can also be of higher-order, in case the operator *B* is applied to itself, as in formulas that are built on a nested doxastic operator *B*, for instance the formula $Ba\gamma \rightarrow BaBa\gamma$, which reads: "If the subject *a* believes γ, then *a* believes to believe γ". Thus, high-order doxastic

logic involves believing about beliefs. Higher-order doxastic logic is useful to study problems concerning the inconsistency of a belief, or to formalize processes of epistemic reflection over one's own belief (usually called "introspection").

The formalization of beliefs and believing processes in higher-order modal language can be used as input for automated reasoning systems in order to compute the probability of truth-value of such beliefs and believing processes. Recent attempts in this direction have used the following higher-order automate theorem provers: LEO-II, TPS, Stallax, and IsabelleP[1]. Such experiments have shown the different strengths of the four theorem provers, given that not all epistemic and doxastic problems were solved in the automated reasoning environment [47]. This would lead to a potential improvement of the theorem provers.

Another example is the computation of degrees of plausibility/possibility of beliefs via a ratio between sets of possible worlds [48]. In this case, the degree of plausibility of a belief was computationally checked via the application of the model checker Mc-COGWED on belief translated in the language of a specific logic, the COmputationally Grounded WEighted Doxastic logic (COGWED).

These experiments present examples of positive interaction between (formalization of) beliefs and computation. On the one hand, they show that beliefs and believing process – when correctly formalized in the language of high-order doxastic logic – can indeed be translated into algorithms and, thus, be computed. As such, these experiments provide useful insights on the logical consistency of beliefs, the degree of probability of their truth value, and the extents and mechanisms of modifying and improving the beliefs. On the other hand, the different degrees of solvability and complexity of doxastic problems involving beliefs invited to develop more effective and higher-performance theorem provers. Since higher-order theorem provers represent a fundamental field in AI research [49], the applications of theorems provers to formalized beliefs strongly invite to an interdisciplinary cross-fertilization between three areas of research: research on AI, philosophical research on beliefs, and research on the epistemology and logic of belief.

However, as already hinted in the previous sections, it is possible to detect an apparent limit in the computational approach to beliefs: the idea that the formalization of beliefs in doxastic logic is possible only via a simplification of the actual believing process. This simplification concerns the fact that beliefs are understood as propositions of which it is possible to calculate the degree of plausibility and to check the veracity, and, thus, to evaluate on which extent beliefs can count as knowledge.

This criticism harkens back to the distinction – hinted in the previous section – between the cognitive capabilities between human and AI systems as far as beliefs are concerned. In fact, the simplification of belief might invite to purport that the machine is confined to operate only upon a limited number of beliefs, the ones that can be expressed in formal language and translated into machine syntax [50]. This would imply the exclusion of all beliefs which content is not limited to events that can or cannot be, and therefore that cannot be reduced to a calculation of the probability of their truth value. These beliefs include, again, the belief in values, emotions, personal virtues and weaknesses. Religious beliefs pertain to this domain too: they are beliefs on beings that are transcendent, uncaused, independent from our mind, and that nevertheless play an important role in our existences, affecting our choices, actions, and lives.

In sum, the skepticism about the translatability of beliefs into machine syntax focuses on the idea that formalizations of beliefs might not distinguish between specific contents of belief – since formalization treats such contents as variables. However, in our life some beliefs are more important than others, last longer than others, or have more important consequences or a stronger impact than others, precisely for their specific content. For instance, a belief such as "John is sick" can be different depending on the relationship between John and the subject. Such criticism would invite to disregard all attempts to translate beliefs into computational language since such attempts seem to consider only a simplification of the complexity of beliefs.

However, it is useful to consider at least two counterarguments to these criticisms.

The first counterargument concerns the fact that doxastic logic deals not only with contents referring to events, but also with self-awareness. This is the above-mentioned case of introspection, formalized by formulas built on nested doxastic operators ($B\alpha\gamma \rightarrow BaB\alpha\gamma$). As such, (higher-order) doxastic logic studies beliefs that have mental states – and not only events – as their objects. Higher-order doxastic logic could represent the difference in importance or complexity of beliefs in terms of the degree of "nestedness" of the doxastic operator. Moreover, dynamic doxastic logic deals with complex forms of believing processes such as belief change, belief revision, and complex forms of belief such as collective belief. The complexity of beliefs is precisely the material for the current advances in doxastic logic [46].

The second counterargument refers to the specificity of religious beliefs. We deepen it in the next session.

V. The Case of Religious Beliefs

As stated in the previous section, a criticism that questions the relevance of the application of automated reasoning programs to the understanding of belief concerns the risk of losing the specificity of religious beliefs as beliefs in transcendent entities, i.e., in entities that are abstract, uncaused, and whose existence is independent from human mind.

There is a counterargument against this criticism: religious beliefs are indeed beliefs in something, that is, they have an object as much as beliefs in events. As such, nothing impedes to present a formalization also of such beliefs.

In fact, the γ (the content of belief) in the formula $Ba\gamma$ can easily be interpreted as a religious content, "Vishnu's existence", "God's omnipotence", "deity *x*" or "property *y* of the deity *x*". Nothing impedes the computation of such belief. This includes the computation of arguments in support of religious beliefs – i.e., arguments that claim to prove the validity of the attribution of the property *y* to the deity *x* object of a religious belief.

The computational translation of arguments in support of a religious belief (based upon a formalization of such arguments) can be important from the point of view of the epistemic introspection of the religious believer because it might help to distinguish between what is strictly necessary and what is not necessary in the logical structure of the argument. In other terms, the computational translation of an argument supporting a religious belief might help to detect what is redundant in the non-computational version of the argument, thus clarifying the belief itself. The consequence of this operation is the increase in self-awareness of a belief, and, thus, the possibility to improve the consistency of the argument supporting the belief.

[1] Concerning IsabelleP, see [47, p. 122]: "The higher-order proof assistant Isabelle/HOL is normally used interactively. In this mode it is possible to apply various automated tactics that attempt to solve the current goal without further user interaction. Examples of these tactics are blast, auto, and metis. It is also possible to run Isabelle from the command line, passing in a theory file containing a lemma to prove. Finally, Isabelle theory files can include ML code to be executed when the file is processed. While it was probably never intended to use Isabelle as a fully automatic system, these three features have been combined to implement a fully automatic Isabelle/HOL, called IsabelleP".

We present two experiments. The first experiment is proposed by Oppenheimer and Zalta, on the wake of the program in computational metaphysics [51]. Oppenheimer and Zalta applied the theorem prover Prover9 to their axiomatization and formalization of Anselm's ontological argument for the existence of God [52], [53]; the result of this application is the discovery that Prover9 needs less lemmas and premises to prove the argument than the ones required in the humanly-formalized version of the argument. This discovery is fundamental to assess the logic of the ontological argument, and, thus, to deepen the extent and limits of the soundness, plausibility, and justification of the belief of such existence.

The second experiment focuses on the work led by Christoph Benzmüller: it consists in the application of high-order theorem provers to a formalization of Gödel's ontological argument [54], [55]. Also in this case, the application of automated reasoning programs led to a simplification of the logical structure of the argument. This discovery provides an incomparable help to deepen the soundness, plausibility, and meaning of believing in an entity (called "God") that possesses all positive qualities at the highest degree.

However, it is possible to question whether these experiments truly address the issue of the specific content of religious belief. As stated, this specific content are entities that are transcendent, i.e., that are abstract, uncaused, and that exist independently on human brains, and that nevertheless affect human lives. It might seem that the aspect of "affecting human life" is completely missed in the two experiments mentioned above. In light of this impact that religious beliefs have on the life of the believers, religious beliefs can be considered part of the big family of "existential beliefs", i.e., beliefs that provide meaning and purpose of human existence. The specificity of religious belief is precisely to have as object a transcendent entity that is source of existential meaning.

Let us harken back to the believing machine of section III. This machine would be able to generate religious or spiritual beliefs, in the same way in would generate beliefs about events, people, politics, economy. The difference would be that the formulation of religious beliefs would imply the distinction between transcendence and immanence. Is this distinction just a minor issue, something that could be easily programed, or is it something unassailable for a machine?

It seems to be hard to conceive a system which tags an event or information as "transcendent" or "supernatural". But it seems to be even harder to conceive a system which recognizes the existential value of transcendence, in the same way as a religious or a spiritual mind is able to do. In sum, what seems to be difficult is to build a *self*-transcending machine. In fact, according to Cantwell Smith, AI systems cannot refer to something external, even less if this "externality" is radically external, i.e., beyond the physical world, "transcendent". Such machine could only assist discerning when something moves to this transcendent level, thus requesting more information about what this transcendent level is about. And anyway, in no case the machine would be able to grasp the existential meaning of this transcendent thing, i.e., the connection between this transcendent with the existence of the machine itself. It would seem that the capacity of transcendence marks a limit for AI systems, and adds a new entry to the list of specific human cognitive features described by Cantwell Smith.

However, it is important to underline a possible ambiguity with the term "transcendence". We can understand it in two ways: as a term that refers to something that lies beyond the physical realm, i.e., something abstract; or we can understand "transcendence" as referring to something that lies beyond the limits of human intelligence, and, thus, beyond the limits of human language.

In the former case, it is worth mentioning again the program in "computational metaphysics". Computational metaphysics is "the implementation and investigation of formal, axiomatic metaphysics [...] in an automated reasoning environment" (http://mally.stanford.edu/cm/). Axiomatic metaphysics is an axiomatic theory of abstract objects [56], [57]. Thus, if we understand "transcendence" as "set of abstract objects", then our understanding of these abstract objects can indeed be computed, and the experiment by Oppenheimer and Zalta supports this.

On the other hand, if "transcendence" refers to something that lies beyond the limits of language, then there are two options: either there is no possible linguistic formulation of this transcendence, or this transcendence shows the limit of language. In the first case, the object of belief cannot be expressed by language, then our belief in such transcendence is void because it cannot be formulated. In the second case, the limit of language is still stated by language, e.g. in the sentence "The transcendent object *x* marks the limit of language". The linguistic formulation of the limit of language implies the distinction between object language and metalanguage: a metalanguage is a language that speak about another language called "object language". Now, to be coherent with the definition, transcendence shows the limit not only of a given object language, but of every possible metalanguage. Therefore, the discourse on this transcendence (a discourse called "theology") is a discourse on the structure of the relationship between object language and metalanguage – a relationship that is at the basis of any possible logical endeavor. Given that this discourse is in principle formalizable [58], nothing impedes that the "belief" in this metalanguage-limiting transcendence is formalizable in the syntax of a machine.

There is also an alternative way to conceive a positive interaction between religious belief and computation. This approach conceives religious beliefs from a decision-making perspective. Rather than focusing on the epistemic aspect of belief, this approach concerns the practical aspect of belief: the modifications and improvements of one's course of action in light of the influx that a specific belief has on the determination of future actions. In this practical approach, the focus switches from "believing what" or "how/why believing what" to "believing, and then doing what". In other words, this approach defines the specificity of religious belief not by referring to a specific (more or less satisfactory) connection with the epistemic requirements of belief, but by referring to the aspects of commitment, decision, choice of action that are the manifestations, outputs, or expressions of one's faith [59], [60]. This approach would contribute to the interaction between machine and belief by connecting the believing process to the research on the computation of decision-making processes [61].

VI. Directions of Future Research

In light of what analyzed, we see at least three directions of future research:

1. The first direction concerns fostering the exploration of the complexity of beliefs in an automated reasoning environment. This includes several points: 1.1. Applying automated reasoning programs to different forms of belief might encourage the dialogue between, on one hand, the research in doxastic logic and dynamic epistemic logic and, on the other hand, philosophy of mind and cognitivist psychology: this interdisciplinary dialogue would better assess what aspects and types of belief have yet to be formalized in doxastic terms. 1.2. (connected to the previous point) Developing the investigation of beliefs in an automated reasoning environment helps to better clarifying what precisely is the "existential" aspect of belief, e.g., what are its epistemological specificity, and what is the specific practical impact of "existential beliefs" on our decision-making processes. 1.3. This first direction of research would also improve the understanding of the distinction of different types

and subtypes of beliefs, for instance as a development of what was presented by Hadley [62]; this would contribute to intersect cognitive science and AI on the topic of belief, and it would be a good starting point for integrating computational modelling in the research on the epistemology of specific kinds of belief – such as religious beliefs [63].

2. The second direction of future research focuses precisely on religious beliefs and the interactions between religious statements (as expressions of religious beliefs) and automated reasoning programs. This includes presenting other applications of theorem provers to other arguments issued from religious beliefs (e.g. *a posteriori* arguments, theological paradoxes, deontic arguments on divine justice, etc.). This direction is simultaneously close and distinct from some recent contributions in (and on) analytic theology [64]: analytic theology aims to "press philosophical tools into theological service" [65, p. 475], while this direction of research aims to apply computational tools for theological service. Such "theological service" consists in detecting redundancies, improving coherency, and reassessing the validity of theological arguments within an axiomatic framework. On the wake of the program in computational metaphysics, this direction of research is called "computational theology" [66]. Moreover, recent research focuses on the relationship between magic and technology [67]; it will be useful to deepen the use of AI in the sociological context of magic as a way to clarify the distinction and analogies between religion and magic.
3. The third direction of research focuses on how the study on the extents and limits of interactions between AI systems and beliefs can contribute to the current debate on the definition of belief systems understood as collections of beliefs with different contents. One example of belief system is religion. The limits that affect all competing definitions of religion – substantive/ontological, functionalist [68], [69], etc. – can be better framed via the deepening of the specificity of logical and computational aspects of belief, including the computational understanding and clarification of the arguments in support of such beliefs. This might have positive applications to the recent discussion on the consonances between religious belief and mathematical realism [70].

These three directions of research might even open to advances in AI developments. The challenge to apply automated reasoning programs to doxastic problems might encourage the development and improvement of these programs themselves [47]. Moreover, the three directions of research might provide elements for fostering the question of the place of belief in scientific research, and the research on the relationship between religion and science.

Will it be possible to write algorithms able to express the complexity of our believing activities and processes? Or will the richness of the spectrum of beliefs, and in specific religious beliefs, prove to be a limit to computability? Whatever the answer might be, as far as it is not tested, it is only a matter of opinion – better, it is only a matter of *belief*. Thus, the best course of action is to foster the multidisciplinary interactions and consonances between the research in AI and the investigation on believing processes, so to provide strategies to test our hypotheses, and to come up with conclusions that are at least provisionary.

References

[1] A. Newen, L. de Bruin, and S. Gallagher, Eds., *The Oxford Handbook of 4e Cognition*, Oxford, U.K.: Oxford University Press, 2018, doi: 10.1093/oxfordhb/9780198735410.013.45.

[2] G. Tononi *et al.* "Integrated Information Theory: From Consciousness to Its Physical Substrate," *Nature Reviews Neuroscience*, vol. 17, pp. 450-461, 2016, doi: 10.1038/nrn.2016.44.

[3] M. Hoffmann and R. Pfeifer, "Robots as Powerful Allies for the Study of Embodied Cognition from the Bottom Up," in *The Oxford Handbook of 4e Cognition*, A. Newen, L. de Bruin, and S. Gallagher, Eds., Oxford, U.K.: Oxford University Press, 2018, pp. 841-862, doi: 10.1093/oxfordhb/9780198735410.013.45.

[4] R. Manzotti, "Embodied AI beyond Embodied Cognition and Enactivism," *Philosophies*, vol. 4, no. 3, 2019. Accessed: Aug. 1, 2021. [Online]. Available: https://www.mdpi.com/2409-9287/4/3/39.

[5] M. Ayers and M. R. Antognazza, "Knowledge and Belief from Plato to Locke," in M. Ayers, *Knowing and Seeing: Groundwork for a New Empiricism*, Oxford, U.K.: Oxford University Press, 2019, ch. 1, pp. 3-33, doi: 10.1093/oso/9780198833567.003.0001.

[6] L Eriksson and A Háyek,"What Are Degrees of Belief?," *Studia Logica*, vol. 86, pp. 183-213, 2007, doi: 10.1007/s11225-007-9059-4.

[7] F. Huber, "Belief and Degrees of Belief," in *Degrees of Belief*, F. Huber and C. Smith-Petri, Eds.,Berlin/Heidelberg, Germany: Springer, 2009, ch. 1, pp. 1-33, doi: 10.1007/978-1-4020-9198-8.

[8] T. Williamson, *Knowledge and Its Limits*, Oxford, U.K.: Oxford University Press, 2002, doi: 10.1093/019925656X.001.0001.

[9] M. Schulz, "Strong Knowledge, Weak Belief?," *Synthese*, 2021. Accessed: Aug. 1, 2021. [Online]. Available: https://link.springer.com/article/10.1007/s11229-021-03180-x.

[10] F. V. Jensen, *An Introduction to Bayesian Networks*, London, U.K.: UCL Press, 1996.

[11] J. Fodor, *The Mind Doesn't Work that Way*, Cambridge, MA, USA: MIT Press, 2000.

[12] R. Swinburne, *The Existence of God*, 2nd ed., Oxford, U.K.: Oxford University Press, 2004, doi: 10.1093/acprof:oso/9780199271672.001.0001.

[13] W. Spohn, *The Laws of Belief: Ranking Theory and Its Philosophical Applications*, Oxford, U.K.: Oxford University Press, 2014, doi: 10.1093/acprof:oso/9780199697502.001.0001.

[14] M. Smith, *Between Probability and Certainty: What Justifies Belief*. Oxford, U.K.: Oxford University Press, 2016, doi: 10.1093/acprof:oso/9780198755333.001.0001.

[15] H. Leitgeb, *The Stability of Belief: How Rational Belief Coheres with Probability*. Oxford, U.K.: Oxford University Press, 2017, doi: 10.1093/acprof:oso/9780198732631.001.0001.

[16] L. Moretti, *Seemings and Epistemic Justification: How Appearances Justify Beliefs*, Cham, Switzerland: Springer, 2020, doi: 10.1007/978-3-030-43392-5.

[17] H. F. Angel, "Religiosität als menschliches Potential. Ein anthropologisches Modell der Religiosität im neurowissenschaftlichen Horizont," in *Religiosität: Anthropologische, theologische und sozialwissenschaftliche Klärungen*, H. F. Angel *et al.*, Eds., Stuttgart, Germany: Kohlhammer, 2006, ch. 5, pp. 62-89.

[18] H. F. Angel, L. Oviedo, R. F. Paloutzian, A. L. Runehov, and R. J. Seitz, *Processes of Believing: The Acquisition, Maintenance, and Change in Creditions*, Cham, Switzerland: Springer, 2017, doi: 10.1007/978-3-319-50924-2.

[19] M. H. Connors, P. W. Halligan, "A Cognitive Account of Belief: A Tentative Roadmap," *Frontiers in Psychology*, vol. 5, 2015. Accessed: Aug. 1, 2021. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpsyg.2014.01588/full.

[20] R. D. Castillo, H. Kloos, M. J. Richardson, and T. Waltzer, "Beliefs as Self-Sustaining Networks: Drawing Parallels Between Networks of Ecosystems and Adults' Predictions," *Frontiers in Psychology*, vol. 6, 2015. Accessed: Aug. 1, 2021. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpsyg.2015.01723/full.

[21] S. Lumbreras and L. Oviedo, "Belief Networks as Complex Systems," *Limina: Grazer theologische Perspektiven*, vol. 3, no. 2, pp. 92-108, 2020, doi: 10.25364/17.3:2020.2.5.

[22] A. Smith, *Thinking about Religion: Extending the Cognitive Sciences of Religion*, Basingstoke, U.K.: Palgrave MacMillan, 2014, doi: 10.1057/9781137324757.

[23] S. Donaldson, *Dimensions of Faith: Understanding Faith Through the Lens of Science and Religion*, Eugene, OR, USA: Wipf & Stock, 2015, doi: 10.1111/heyj.13085.

[24] H. F. Angel and R. J. Steiz, "Process of Believing as Fundamental Brain Function: The Concept of Credition," *Research Bulletin of the Sigmund Freud PrivatUniversität Wien*, vol. 3 no. 1, 2016, doi: 10.15135/2016.4.1.1-20.

[25] M. Sugiura, R. J. Seitz, H. F. Angel, "Models and Neural Bases of the Believing Process," *Journal of Behavioral and Brain Science*, vol. 5, no. 1, pp. 12-23, 2015, doi: 10.4236/jbbs.2015.51002.

[26] J. Ladyman, J. Lambert, K. Wiesner, "What is a Complex System?," *European Journal for Philosophy of Science*, vol. 3, pp. 33–67, 2013. doi:10.1007/s13194-012-0056-8.

[27] M. Mitchell, *Complexity: A Guided Tour*, Oxford, U.K.: Oxford University Press, 2011.

[28] J. D. Frank, "Nature and Functions of Belief Systems: Humanism and Transcendental Religion," *American Psychologist*, vol. 32, no. 7, pp. 555–559, 1977, doi: 10.1037/0003-066X.32.7.555.

[29] J. Crocker, S. T. Fiske, and S. E. Taylor, "Schematic Bases of Belief Change," in *Attitudinal Judgment*, J. R. Eiser, Ed., Springer Series in Social Psychology, New York, NY, USA: Springer, 1984 pp. 197-226, doi: 10.1007/978-1-4613-8251-5_10.

[30] N. Rodriguez, J. Bollen, and Y. Y. Ahn, "Collective Dynamics of Belief Evolution Under Cognitive Coherence and Social Conformity," *PloS One*, vol. 11, no. 11, 2016. Accessed: Aug. 1, 2021. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0165910.

[31] B. Sodian, D. Zaitchik, and S. Carey, "Young Children's Differentiation of Hypothetical Beliefs from Evidence," *Child Development*, vol. 62, no. 4, pp. 753-766, 1991, doi: 10.1111/j.1467-8624.1991.tb01567.x.

[32] G. Pennycook, J. A. Cheyne, P. Seli, D. J. Koehler, and J. A. Fugelsang, "Analytic Cognitive Style Predicts Religious and Paranormal Belief," *Cognition*, vol. 123, no. 3, pp. 335-346, 2012, doi: 10.1016/j.cognition.2012.03.003.

[33] K. Toner, M. R. Leary, M. W. Asher, and K. P. Jongman-Sereno, "Feeling Superior is a Bipartisan Issue: Extremity (Not Direction) of Political Views Predicts Perceived Belief Superiority," *Psychological Science*, vol. 24, no. 12, pp. 2454-2462, 2013, doi: 10.1177/0956797613494848.

[34] T. Ruffman, J. Perner, and L. Parkin, "How Parenting Style Affects False Belief Understanding," *Social Development*, vol. 8, no. 3, pp. 395-411, 1999, doi: 10.1111/1467-9507.00103.

[35] K. Jones, "Some Epistemological Considerations of Paradigm Shifts: Basic Steps Towards a Formulated Model of Alternation," *The Sociological Review*, vol. 25, no. 2, pp. 253-272, 1977, doi: 10.1111/j.1467-954X.1977.tb00289.x.

[36] B. Goertzel, "Belief Systems as Attractors," in *Chaos Theory in Psychology and the Life Sciences*, R. Robertson and A. Combs, Eds., New York, NY, USA: Psychology Press, 1995, ch. 9, pp. 123-134, doi: 10.4324/9781315806280.

[37] A. Fuentes, *Why We Believe: Evolution and the Human Way of Being*, New Haven, CT, USA: Yale University Press, 2019.

[38] B. Cantwell Smith, *The Promise of Artificial Intelligence: Reckoning and Judgement*, Cambridge, MA, USA: MIT Press, 2019, doi: 10.7551/mitpress/12385.001.0001.

[39] D. McDermott, "Artificial Intelligence and Consciousness," in *The Cambridge Handbook of Consciousness*, P. D. Zelazo, M. Moscovitch, and E. Thompson, Eds., Cambridge, U.K.: Cambridge University Press, 2007, ch. 6, pp. 117-150, doi: 10.1017/CBO9780511816789.007.

[40] S. Lumbreras, *Respuestas al transhumanismo: cuerpo, autenticidad y sentido*, Madrid: Digital Reason, 2020.

[41] M. Daňková and L Běhounek, "Fuzzy Neighborhood Semantics for Multi-agent Probabilistic Reasoning in Games," in *Information Processing and Management of Uncertainty in Knowledge-Based Systems*, IPMU 2020, Communications in Computer and Information Science, vol 1239, M.-J. Lesot *et al.*, Eds., Cham, Switzerland: Springer, 2020, pp. 680-693, doi: 10.1007/978-3-030-50153-2_50.

[42] J. Hintikka, *Knowledge and Belief: An Introduction to the Logic of the Two Notions*, 2nd ed., V. F. Hendriks and J. Symons, Eds., London, U.K.: College Publications, 1962.

[43] R. Rendsvig, J. Symons, "Epistemic Logic," *Stanford Encyclopedia of Phlosophy*, E. N. Zalta, Ed., 2021. Accessed: Aug. 1, 2021. [Online]. Available: https://plato.stanford.edu/archives/sum2021/entries/logic-epistemic/.

[44] Y. Ding, "On the Logic of Belief and Propositional Quantification," *Journal of Philosophical Logic*, 2021, doi: 10.1007/s10992-021-09595-8.

[45] D. Perlis, "The Role(s) of Belief in AI," in *Logic-Based Artificial Intelligence*, J. Minker, Ed., Boston, MA, USA: Springer, 2000, pp. 361-374, doi: 10.1007/978-1-4615-1567-8_16.

[46] A. Baltag, B. Renne, "Dynamic Epistemic Logic," *Stanford Encyclopedia of Philosophy*, E. N. Zalta, Ed., 2016. Accessed: Aug. 1, 2021. [Online]. Available: https://plato.stanford.edu/archives/win2016/entries/dynamic-epistemic/.

[47] C. Benzmüller, "Combining and Automating Classical and Non-Classical Logics in Classical Higher-Order Logics," *Annals of Mathematics and Artificial Intelligence*, vol. 62, no. 1, pp. 103-128, 2011, doi: 10.1007/s10472-011-9249-7.

[48] T. Chen, G. Primiero, F. Raimondi, and N. Rungta, "A Computationally Grounded, Weighted Doxastic Logic," *Studia Logica*, vol. 104, pp. 679–703, 2016, doi: 10.1007/s11225-015-9621-4.

[49] K. Bansal, S. M. Loos, M. N. Rabe, C. Szegedy, and S. Wilcox, "HOList: An Environment for Machine Learning of Higher-Order Theorem Proving," in *Proceedings of the 36th International Conference on Machine Learning*, Long Beach, CA, 2019. Accessed: Aug.1, 2021. [Online]. Available: http://proceedings.mlr.press/v97/bansal19a/bansal19a.pdf.

[50] J. Leach, *Mathematics and Religion: Our Languages of Sign and Symbol*, West Conshohocken, PA, USA: Templeton Press, 2010.

[51] B. Fitelson and E. N. Zalta, "Steps Towards a Computational Metaphysics," *Journal of Philosophical Logic*, vol. 36, pp. 227-247, 2007, doi: 10.1007/s10992-006-9038-7.

[52] P. Oppenheimer and E. N. Zalta, "A Computationally-Discovered Simplification of the Ontological Argument," *Australasian Journal of Philosophy*, vol. 89, no. 2, pp. 333–350, 2011, doi: 10.1080/00048401003674482.

[53] P. Oppenheimer and E. N. Zalta, "On Anselm's Ontological Argument in *Proslogion* II," submitted for publication.

[54] C. Benzmüller and D. Fuenmayor, "Computer-Supported Analysis of Positive Properties, Ultrafilters and Modal Collapse in Variants of Gödel's Ontological Argument," *Bulletin of the Section of Logic*, vol. 49, no. 2, 2020, doi:10.18778/0138-0680.2020.08.

[55] C. Benzmüller and D. Fuenmayor, "Can Computers Help to Sharpen Our Understanding of Ontological Arguments?" *Mathematics and Reality, Proceedings of the 11th All India Students' Conference on Science Spiritual Quest*, IIT Bhubaneswar, Bhubaneswar, India, 6-7 October, 2018, pp. 195-226, doi: 10.13140/RG.2.2.31921.84323.

[56] E. N. Zalta, *Abstract Objects: An Introduction to Axiomatic Metaphysics*, Dordrecht: D. Reidel, 1983. Available: http://mally.stanford.edu/abstract-objects.pdf.

[57] E. N. Zalta, *Principia Logico-Metaphysica*, 2021, unpublished. Accessed: Aug. 1, 2021. [Online]. Available: http://mally.stanford.edu/principia.pdf.

[58] A. Vestrucci, "Metalanguage and Revelation: Rethinking Theology's Language and Relevance," *Logica Universalis*, special issue "Theological Discourse and Logic", S. Krajewski and M. Trepczyński, Eds., vol. 13, no. 4, pp. 551–575, 2019, doi: 10.1007/s11787-019-00236-y.

[59] L. Buchak, "Faith and Steadfastness in the Face of Counter-Evidence," *International Journal of Philosophy of Religion*, vol. 81, pp. 113–133, 2017, doi: 10.1007/s11153-016-9609-7.

[60] L. Buchak, "Can It Be Rational to Have Faith?," in *Probability in the Philosophy of Religion*, J. Chandler and V. S. Harrison, Eds., Oxford, U.K.: Oxford University Press, 2012, ch. 12, pp. 225-246, doi: 10.1093/acprof:oso/9780199604760.003.0012.

[61] M. Calder *et al.*, "Computational Modelling for Decision-Making: Where, Why, What, Who and How," *Royal Society Open Science*, 2018. Accessed: Aug. 1, 2021. [Online]. Available: https://royalsocietypublishing.org/doi/10.1098/rsos.172096.

[62] R. F. Hadley, "The Many Uses of 'Belief' in AI," *Minds and Machines*, vol. 1, pp. 55-73, 1991.

[63] L. Oviedo and K. Szocik, "Religious–And Other Beliefs: How Much Specificity?" SAGE Open, January 2020. Accessed: Aug. 1, 2021. [Online]. Available: https://journals.sagepub.com/doi/full/10.1177/2158244019898849.

[64] O. D. Crisp and M. C. Rea, Eds., *Analytic Theology: New Essays in the Philosophy of Theology*. Oxford, U.K.: Oxford University Press, 2009, doi: 10.1093/acprof:oso/9780199203567.001.0001.

[65] O. D. Crisp, "Analytic Theology," *The Expository Times*, vol. 122, no. 10, pp. 469–477, 2011, doi: 10.1177/0014524611408533.

[66] A. Vestrucci, "Computational Theology and Natural Theology," presented at the 2021 conference of the Ian Ramsey Center *Natural Theology in the 21st Century*, University of Oxford, Oxford, U.K., 15-17 July 2021.

[67] L. Obadia, "Moral and Financial Economics of 'Digital Magic':

Explorations of an Opening Field," *Social Compass*, vol. 67, no. 4, pp. 534-552, 2020, doi:10.1177/0037768620950237.

[68] V. S. Harrison, "Internal Realism and the Problem of Religious Diversity," *Philosophia*, vol. 34, no. 3, pp. 287-301, 2006, doi: 10.1007/s11406-006-9029-5.

[69] K. Schilbrack, *Philosophy and the Study of Religions: A Manifesto*, New York, NY, USA: Wiley Blackwell, 2014.

[70] V. S. Harrison, "Mathematical Objects and the Object of Theology," *Religious Studies*, vol. 53, no. 4, pp. 479-496, 2017, doi: 10.1017/S0034412516000238.

Andrea Vestrucci

Ph.D. in ethics from the University of Lille SHS and the University of Milan, and Th.D. in systematic theology from the University of Geneva. He was a professor of ethics and logic at the Federal University in Fortaleza, Brazil, and currently he is a researcher in systematic theology and ethics for the Graduate Theological Union, Berkeley. He is also a privatdozent for the University of Geneva. He is a laureate of the Academic Society of Geneva. He was the recipient of the Australian Award by the Australian Federal Government. He also taught and was researcher at Monash University, the University of Münster, and the University of California Berkeley. His current research focuses on the ethics of Artificial Intelligence, the interactions between religion, logic, and technology, and the ethical and methodological aspects of interreligious dialogue. Amongst his recent publications: the book Theology as Freedom: On Martin Luther's De servo arbitrio (Mohr Siebeck, 2019), the paper "Recalibrating the Logic of Free Will" (Theology and Science, 2020), the edition Religions and Languages: A Polyphony of Faiths (Sophia, Springer, 2021).

Sara Lumbreras

Sara Lumbreras holds a PhD and a MSc Eng from Universidad Pontificia Comillas. She is a professor at the Institute for Research in Technology and teaches at the Industrial Management Department at the ICAI School of Engineering and the Financial department at the ICADE School of Business and Law. She is currently Deputy Director of Research Results at the Institute of Research in Technology. Her research focuses on the development and application of decision support techniques for complex problems, mainly in the energy sector and in particular in grid design, but also in finance and the health sector, where she applies Artificial Intelligence to the prediction of disease evolution. She works with classical optimization techniques (such as Benders' decomposition), metaheuristics and Machine Learning. She has five years of experience in the private sector (JPMorgan London). In addition to her technical research, she studies transhumanism and the implications of Artificial Intelligence in anthropology.

Lluis Ovideo

Lluis Oviedo is a full Professor for Theological Anthropology at the Pontifical University Antonianum of Rome, and Fundamental Theology at the Theological Institute of Murcia (Spain). He is team member of research group on Creditions, based in Graz University (Austria). He edits a book series on "New approaches to the scientific study of religion" (Springer) and the bibliographic bulletin ESSSAT News & Reviews. Research interests focus on the dialogue between theology and sciences, including the new scientific study of religion, and more recently to the interaction between religious belief and AI, and to how religious beliefs and behaviors are related to personal and social wellbeing.

Artificial Intelligence and Spirituality

José Fernando Calderero Hernández*

Universidad Internacional de La Rioja, Logroño (Spain)

Received 18 December 2020 | Accepted 19 June 2021 | Published 7 July 2021

Abstract

Drawing from a conceptual review of the terms 'mind', 'intelligence', 'spirit', 'spirituality', 'spiritual intelligence' and their possible interrelations, an approach to the concept 'human nature' is made in relation to transhumanism and post-humanism. In addition, through a reflection on the nature and meaning of the terms 'datum', 'coding', 'language', 'energy', 'concrete', and 'abstract', some dimensions of 'artificial intelligence' (AI) and their analogies and differences with 'the spiritual' are shown. After a brief foray into the concept of 'reality' and its probable 'fuzziness', we discuss their intrinsic and inherent mutability, and the possible existential dependence of some of their parts on the intentional activity of personal beings. We point out the dangers, for intellectual rigor and therefore for life in general, and human life in particular, of reductionist interpretations of reality that, arguing at having been scientifically proven, are intended to provide a closed and indisputable explanation of facts and phenomena of diverse aetiology, ignoring the need for 'management of the unknown'. Consequently, an open, synergetic, harmonious vision of the role of technology and the humanities, especially those most focused on the study of the intangible, is necessary for the progress of knowledge and, therefore, for the mutually beneficial care of humanity and nature.

Keywords

Intelligence, Spirituality, Artificial Intelligence (AI).

DOI: 10.9781/ijimai.2021.07.001

I. Introduction

If we were to imagine humanity as a single family trying to thrive by exploiting available natural resources in an intelligent manner, in alignment with the goals of the United Nations 2030 Agenda for Sustainable Development (United Nations, 2015) [1], to achieve the objectives of 'caring for our common home', in the words of Pope Francis (2015) [2], it would seem reasonable to stockpile all available goods and resources for efficacy and efficiency. It would be advisable to use all available resources and avoid any reductionism that might, due to narrow-mindedness, compromise the success of the endeavour; in this respect, the huge importance of abundant intangible resources available to humans, such as talent, creativity, imagination, etc. – many of which, if not all, are located within what we might call the 'human spirit' – must be underscored. On the other hand, these very talents have prompted the emergence of tangible artefacts, subject to procedures and methods inherent to experimental sciences but with a 'modus operandi' quite similar to human intelligence, resulting in their being called artificial intelligence in a broad sense.

Following the logic of the benefits of pooling resources, it seems advisable and necessary to contribute to it, in an attempt to provide potentially valuable perspectives here by exploring different aspects of both worlds.

The thesis of the article consists in pointing out that AI can help to minimize negative consequences of the important dose of ambiguity, polysemy and synonymy existing in some of the relevant terms necessary both in the field of spirituality and in that of AI.

* Corresponding author.

E-mail address: josefernando.calderero@unir.net

We will argue how the difficulty of precisely defining the relevant keywords for the study of both topics, and the concepts of blurred reality and interdimensional unity of reality can be useful to show the relevant role that AI can, and should, play in the advance of those areas traditionally considered as remote from experimental science and technique and contribute to disseminating the need for effective synergy between technical advances in AI and the conceptual and methodological needs of the disciplines that study the intangible.

We consider it relevant to reflect on the contributions that AI can make to the study of spirituality since we understand that this is seriously harmed by the limitations imposed by language, which force both to dispense with relevant and significant nuances and dimensions and to only consider suitable those research results that can be expressed with absolute precision. We understand that the great information processing capacity of AI can allow, to a certain degree, the exceeding of the aforementioned limits.

II. A Conceptual Review

Do the words we use describe unequivocally and with absolute precision the external, internal, ontological characteristics of what is, or is considered, real (or whose possible existence is, at least, not discarded) and could be deemed pre-existing, or must we settle for using them as mere approximations, we might say asymptotic, of the 'realities' they attempt to describe?

If we expect to radically answer this question, we must, in each case, determine exactly all the applicable dimensions of the object under consideration, a presumably unattainable endeavour even within the limits of the dimensional spectrum recordable by human senses and technology, at least at the current degree of development of both of those 'resources'. If, in order to consider an object perfectly described,

we demanded that the descriptor element, word, code, symbol, etc. include all the object's dimensions, we would find ourselves faced with huge, presumably insurmountable, operational difficulties. Imagine, for example, that we needed to convey the 'perfect' description of a simple piece of bread across a channel that only used written text as code. With certain limitations, we could describe its colour (omitting logically the nuances of its different areas and using only the words describing the main colours), its size (renouncing a detailed description of the outline of its edges, supposing the word 'edge' made sense on a subatomic scale), its weight (up to a reasonable number of decimals), its location (evidently not that of each of its parts, but possibly an approximate reference to its geometric centre), its temperature (assigning to it an average temperature calculated by the measures at different points), the date and time it was baked, its chemical composition (again using averaged data if the dough was not perfectly uniform), and possibly some other dimensions whose identification and measurement were reasonably possible. It would be much more difficult, even impossible, to describe its smell, the traceability of its components, its commercial appeal, its radioactivity... If such an apparently simple task does not seem feasible, what might we say about dimensions that remain possibly undetectable, whether due to limitations of the technical or biological recording instrument or because they have not yet been discovered? It does not seem very rigorous to deny their existence, or at least the possibility of their existence, and their potential eventual influence, due to the simple fact of being unable to assert their existence.

In view of the above, which must be considered despite falling more in the realm of the philosophy of language, it seems reasonable not to be overly optimistic when attempting to find definitions for these concepts that are, if not irrefutable, at least accepted by a reasonable number of scholars: 'mind', 'intelligence', 'spirit', 'spirituality', 'spiritual intelligence', 'human nature', 'datum', 'coding', 'language', 'energy', 'concrete', 'abstract', and 'reality'. However, accepting this impossibility, we will try to ultimately offer at least one explanation for each of these words that is sufficiently accepted by the academic community, while prudently keeping in mind the conceptual background of Caeiro's input (2018) [3]:

> All text is relative. Text is intertext, linking various texts, quotes, ideas… which do not belong to the author; there is only a confluence of stories coming for different cultures. The matrix (screen, fabric, panel) where writing and languages (visual, alphabetical, oral…) are located is an organ with its own entity and will, constructed with threads and scraps taken from different spaces and times of rendering, preventing us from knowing what the true and originating fact is.
>
> CAEIRO, 2018, p. 164.

In an area of knowledge as rich in intangible elements as psychology, it does not seem possible to imitate institutions such as the International Union of Pure and Applied Chemistry [IUPAC] in their successful paths towards the design and implementation of an specific international nomenclature; nevertheless, it would be very beneficial for an agency to be eventually created that could, aided by current developments in semantics engineering, big data analytics and artificial intelligence, provide the artificial sciences in general, and the social sciences and educational sciences in particular, with terminology support and a specific glossary explaining the various senses and meanings that different authors, schools of thought, organizations, etc., assign to the most commonly used terms internationally. Nor would it be unreasonable for said desired agency to coordinate the possible creation of neologisms that would enable understanding in the numerous, rather conflicting, situations generated by the abundant polysemy and synonymy of some languages, especially in the use and comprehension of relevant terms.

In any case, the design, creation and implementation of the aforementioned glossary should allow the processing of keywords without sacrificing semantic richness; Rather, it would be about taking advantage of the current possibilities of AI to enrich languages, trying to adapt them to the complexity of reality instead of trying to adapt it, in vain, to a comfortable simplification of the language.

When trying to decipher the meaning of the word 'mind', we find the following definition in the Dictionary of Psychology of the American Psychological Association (APA) [4]:

> 1. Broadly, all intellectual and psychological phenomena of an organism, encompassing motivational, affective, behavioural, perceptual, and cognitive systems; that is, the organized totality of an organism's mental and psychic processes and the structural and functional cognitive components on which they depend. The term, however, is also used more narrowly to denote only cognitive activities and functions, such as perceiving, attending, thinking, problem solving, language, learning, and memory. The nature of the relationship between the mind and the body, including the brain and its mechanisms or activities, has been, and continues to be, the subject of much debate.
>
> …/…
>
> 5. Human consciousness regarded as an immaterial entity distinct from the brain.
>
> 6. The brain itself and its activities. In this view, the mind essentially is both the anatomical organ and what it does.
>
> APA DICTIONARY OF PSYCHOLOGY Mind

There is no need to dwell on the term's evident polysemy, with its resulting lack of rigour and its negative consequences in practice.

The concept of 'intelligence' fares no better, continuing to hide its essence from researchers across centuries. Currently, two great 'classical' perspectives basically continue to be in force: Spearman's (1927) [5], which defended the existence of a g factor, as general mental energy, and Thorndike's (1920) multifactor theory [6] about the existence of many different intellectual capacities, developed many decades prior to the popular 'theory of multiple intelligences' now well-known through the media. This dichotomous approach is being overtaken by the more global concept of 'unidiverse intelligence', according to which 'intelligence is singular and multiple at the same time' (Martínez-Otero, 2016, p.119) [7].

Zubiri (1982) offers a broader concept of 'sentient intelligence', stating that 'There is no sensing "and" intellection, but merely sentient intellect, an intellect impressing as real what is real' (p.15) [8], thus leading to the consideration of intelligence as part of a whole that intrinsically and essentially includes affectivity. According to this concept, the human bond with reality would cease to be considered exclusively, or fundamentally, an intellective issue. This very interesting approach opens the door to considering intelligence as one simple element in the process of 'communion' (common union) with what is real, thus unleashing possibilities of new, broader relationships than those that must be expressed through codes.

It seems pertinent to mention at this point that 'to educate is to help each human being establish and maintain valuable bonds with reality' (Calderero, Aguirre, Castellanos, Peris, Perochena, 2014, p. 144) [9]. Now we add 'especially with people' (Calderero. Perochena, Peris, 2015, p. 123) [10], where 'each significant word in the proposed statement would be an exponent of a profound semantic load like a meristem generating new concepts and practical didactic applications' (Calderero, Aguirre, Castellanos, Peris, Perochena, 2014, p. 40) [9]. That there is no conceptual restriction regarding those 'valuable bonds with reality', which could be, and are, very diverse and have different characteristics, is understood; unconscious or unknowable bonds cannot be discarded.

Despite the widespread publication of Gardner's quasi-definition of intelligence, according to which 'human intellectual competence must dominate a set of problem-solving skills, enabling the individual to resolve genuine problems, or the difficulties they face and, when appropriate, create an effective product' (Gardner, 1983, p. 66) [11], we can consider a more 'official' definition of 'intelligence': 'the ability to derive information, learn from experience, adapt to the environment, understand, and correctly utilize thought and reason' (American Psychological Association) [4].

For an understanding of the complexity of the task, we recommend reading the entry for 'intelligence' offered by Ferrater Mora (1965) [12].

There is similar ambiguity regarding the term 'spirit'; such is the volume of synonymy and polysemy for this word that Ferrater Mora (1965) states that 'in view of it all, one might wonder if it would not be better to banish the words "spirit" and "spiritual" from philosophy, primarily if we keep in mind that in some modern languages there is confusion between what is "spiritual" and what is "mental"' (Ferrater Mora, 1965, p. 572) [12]. The same author states the 'in numerous cases, spirit (under that same name or others) is understood as something opposed to matter' (Ferrater Mora, 1965, p. 572) [12].

Although the word 'something' might seem to be barely rigorous, its use is reasonable given that the different sciences involved were unsuccessful in being more precise; let us then accept it as being 'something immaterial', while recognising that if we wish to arrive at its ultimate meaning, this is not very illuminating either. Elucidating what 'immaterial' means would be easier if we knew what matter is, but this is not the case because the concept vanishes if we descend, or ascend, to quantum levels. However, as it enjoys a few behavioural features that are somewhat predictable on a human scale and allow us to use it, obtaining previously imagined results, 'matter' is considered something more real, 'objective'. For its 'resonance' with this state of affairs, might we assign to that 'something immaterial' an 'existential status', calling it 'spirit' by observing some of the workings which may be attributed to it?

Even accepting its existence as 'something' distinct from matter, we remain far from having resolved the multiple unknown aspects that arise. For example, if it does not have matter, its nature is not 'suited' to the categories of space or time. Not having matter, it would not have borders, edges. Thus, following a physical analogy, it could not be in a place. Would that mean that 'my spirit' might not necessarily follow me wherever I go? Can my 'spirit' grow old?

At the risk of being daring, we ask ourselves a question whose answer would require a reassessment of certain tenets held very firmly by our rationalistic 'Western' culture, an adjective that clearly lacks sufficient intellectual rigour. Said question would be: Does 'the radical distinction between the sciences of nature and the sciences of the spirit' posed by Dilthey (1883, as cited in Rizo, 2015, p.276) make sense? [13].

In principle, 'spirituality' presents fewer interpretation issues, as its meanings invoke styles of thought, social movements, vital approaches 'per se' more open to diverse interpretation without greater specification apparently being necessary, which is why we shall be content with the definitions provided by the Royal Spanish Academy dictionary:

> 1. Fem. Nature and condition of spiritual.
>
> 2. Fem. Quality of what is spiritualised or reduced to the ecclesiastical condition.
>
> 3. Fem. Spiritual deed or thing.
>
> 4. Fem. Set of ideas referring to spiritual life.
>
> DICCIONARIO DE LA REAL ACADEMIA ESPAÑOLA (R.A.E.). Espiritualidad. [Spirituality] [4]

Although it could be said that the concept of 'spiritual intelligence' is an ancient construct, this explicit phrasing is fairly recent and gaining strength in academic literature.

Torralba's (2010) [14] position seems acceptable when he states that there exists in human beings

> ...a complex series of abilities not present in other vertebrates which allow for the elaboration, with compelling reason, of the hypothesis of a form of intelligence that could be called spiritual.
>
> TORRALBA, 2010, p. 14.

Despite the evident weakness of including in the definition the very term being defined, we cite the following definition for spiritual intelligence: 'the ability to build a healthy (or adaptive) system of spiritual values or beliefs and adopting it as a lifestyle (i.e. adhering to those values)' (Arias & Lemos, 2015, p. 96) [15]. We do so without, of course, omitting that 'its characterisation, development, and education constitute a very open subject worthy of exploration' (Torralba, 2015, p. 15) [14].

The currently named 'transhumanist' and 'post-humanist' streams of thought have once again brought to the fore the as-yet-insufficient explanations of what the human being is, and what is proper to him. As witnessed by prior generations, it remains imperative to delve further into understanding human nature and its similarities and differences with other living beings and artefacts that can imitate, and even exceed, human behaviours, as well as thoughts or very similar processes, through artificial intelligence.

Bostrom (2014) [16] defines transhumanism as

> The intellectual and cultural movement that affirms the possibility and desirability of fundamentally improving the human condition through applied reason, especially by developing and making widely available technologies to eliminate aging and to greatly enhance human intellectual, physical, and psychological capacities.
>
> BOSTROM, 2014, p. 1.

This is a definition that, except for the reference to 'eliminat[ing] aging', could practically be applied to any sphere of knowledge; is there any discipline whose aims are not about 'fundamentally improving the human condition'? In light of the controversy and criticism, and the reluctance in various academic and social areas regarding the presumable risk of dehumanisation, the somewhat absurd view that improving humans would be counterproductive could be held. However, it makes sense to be concerned, as there is always danger of interpreting humanity's 'improvement' as an attempt to 'manufacture' a 'superman', an endeavour which has historically ended quite tragically.

Taking another step in the processes of human transformation, we find in Valera and Marambio (2019) [17] that:

> The focal point of post humanism consists not so much in the uncritical acceptance of the possibilities offered by technology —as occurs with transhumanism— but rather in a total contamination and hybridisation of human beings with other living beings and with machines.
>
> VALERA & MARAMBIO, 2019, p. 308.

In this regard, we might ask ourselves whether humans, speaking both individually and collectively, are free or not from interaction with other inanimate, animate, or 'pseudo animate' beings, this latter group being the devices, such as robots, which can imitate typical actions of living beings themselves. We can at least draft some response in the obvious sense that the existence and survival of concrete populations and humans have historically been connected to a greater or lesser extent with available resources and with other populations or individuals. It is not infrequent to find humans who can move, or live, due to the occasional or permanent help of technological artefacts, some of them capable of 'making decisions'. That said, does the existential need for interdependence with others necessarily imply negating an essential humanness?

Without purporting to define human nature, it seems necessary to accept its existence despite it being a concept currently denied in wide areas of the philosophical arena. We agree with Marcos (2010) [18] when stating that 'Human nature, seen as a limit and restriction, could at the same time be a condition allowing any improvement, at least as its axiological principle' (p. 200); that 'Ultimately, if human nature is completely natural, then it is technically available' (p. 201); and that 'The technical assault on human beings is advocated from these philosophical foundations' (p.201).

Everything seems to indicate the need, the urgency, of moving towards a profound knowledge of what a human being is, because, as stated by Marcos (2010) [18]:

> Never, therefore, has the philosophical task of considering human nature itself been so pressing. This will be what will enable us to judiciously apply technologies to therapy and farming. The error is not in technology, rather in its Utopian-futuristic confluence and direction. Never has the consideration of human nature been more pressing, to avoid its very loss along the way to action. And to avoid as well fear-induced Luddite attitudes which paralyse science and technology, and consequently the possibility of the effective improvement of human life.
>
> MARCOS, 2010, p. 203.

Before reflecting on the nature and meanings of a series of other terms that are related to the subject at hand such as "data", "encoding", "language", "energy", "concrete", "abstract", it must be noted that 'it seems unworthy of the lowest intellectual rigor to consider that any researched reality must be fully explained through the exclusive application of known methodologies' (Calderero & Calderero, 2017, p. 52) [19], and that 'we understand that when making the meaning of a term or concept explicit, the use of a form such as "we name … as X" fits reality better than the usual "X is…"'. (Calderero & Calderero, 2017, p. 39) [19]. Assuming both tenets are valid, we intend to leave open the possibility of discovering new unknown realities, new dimensions of known realities or new interpretations regarding realities already studied from other points of view.

A. Datum

Even in academia, the social custom of considering data as objective elements, devoid of any connotation, and consequently awarding them full credibility, especially if they are of a quantitative nature, seems quite prevalent; however, given that facts and data are realities of a different nature, this custom generates a lack of intellectual rigour and, depending on the social, cultural or intellectual area in which it is applied, can even cause serious damage.

Domínguez (2001) [20] states that:

> Data are a cultural product; they cannot therefore be grasped in an aseptic manner. Furthermore, data do not appear, but are rather constructed by us during the research process.
>
> DOMÍNGUEZ, 2001, p. 114.

and that:

> The view that all data are "altered" three times: by those who produced them (cultural alteration), by their historical process (post depositional alteration), and by those who interpret them (interpretive alteration). This last alteration is what leads us to conclude that data are always inserted into an interpretative discourse.
>
> DOMÍNGUEZ, 2001, p. 118.

In order to indicate at least one difference between a reality and its translation into data, we can use the analogy of the sound possibilities of a piano and those made by a slide trombone or a violin, for example. In the former, only sounds predetermined by the position of the keys may be made, while in the case of the trombone and violin, or similar instruments, it is possible to obtain what could be called a continuous frequency. The formal expression of a datum is necessarily restricted to the expressive capabilities of the measurement instrument used, according to the definition of the construct whose variable value is being determined. It could be concluded that the use of data, despite being necessary to process information, is reductionism.

B. Coding

If, according to Alegre (2019) [21], coding is transforming information from one type of representation to another (p. 29), then coded information is necessarily different from the original – related to it, in the best of cases by applying a bijection, but different. The object of coding is to adapt information to the interpreting instrument, which necessarily means the impossibility of processing of any information which, by its nature, is not 'understood' by the system that must process it.

C. Language

Like any relevant concept, 'language' is not free from synonymy and polysemy, so we must once again recognise the impossibility of a perfect, universal definition of the word, and must settle for a generic meaning that is compatible with a broad range of different 'languages': non-verbal, iconic, musical language, etc., ultimately extending to any system that allows the communication of ideas, facts and feelings.

It is interesting to reflect on what Echeverría (2017) [22] presents as the First Tenet of the Ontology of Language, where:

> Language is, above all, what makes human beings the particular type of beings we are. We posit that human beings are linguistic beings, beings that live in language. Language, we postulate, is the key to understanding human phenomena.
>
> It is important to avoid a reductionist interpretation of this tenet that restricts the complexity of human phenomena to language and therefore disregards other non-linguistic dimensions of the human existence. We know that human beings are not just linguistic beings and that, therefore, language does not exhaust the multi-dimensionality of the human phenomenon.
>
> ECHEVERRÍA, 2017, p. 21.

If we accept that language does not exhaust the multi-dimensionality of the phenomenon, we must conclude that neither can it exhaust the complexity of everything real. That is, reality cannot be completely viewed as reflected by any language, any type of representation. What is described and its description are necessarily different, related, but different realities.

In support of this thesis, we cite LEOCATA (2003) [23]:

> The constant rethinking of scientific theories, the questioning of what previously seemed immutable, conspire for the logically constructed language to be considered as something both necessary .../... and yet hypothetical, in terms of its correspondence with the real world. We know how we can logically order our language, but that alone does not guarantee knowledge of what the world is like. Thus, for the philosophies of the analysis of language, the old Kantian theme of an unattainable "thing in itself" is reconsidered, what Davidson calls, from the point of view of the philosophy of language, "the inscrutability of reference".
>
> LEOCATA, 2003, p. 288.

D. Energy, Concrete, Abstract

By mentioning the concept 'energy', we risk having it considered out of place in this context; however, it is appropriate to deal with it given its possible position as intermediary construct between that which is 'material' and that which is 'spiritual'. Indeed, 'energy' is not 'matter', nor is it 'spirit', but it occupies a substantial role in the lives of humans and, although like those of the other relevant concepts its definitions have weak aspects, it can help us approach our goal of finding analogies and differences between the field of data and the spiritual.

In attempting to approach, perhaps naively, the concept of energy, we must mention Bunge (1999) [24], who states:

> Every science that deals with concrete (or material) things, from physics to social sciences, uses one or more concepts of energy. For example, a psychobiologist wishes to measure the metabolic cost (in calories) of one bit of information transmitted by a synapse; an anthropologist, a sociologist, and an economist are interested in knowing what a community's per capita energy consumption is; they also wish to know whether the members of a given society work in a way that optimises their energy efficiency.
>
> BUNGE, 1999, p. 54.

> Being ubiquitous, the general concept of energy must be philosophical and, particularly, metaphysical (or ontological); i.e., it is like the concepts of thing and property, space and time, causality and chance, law and trend, and so many others.
>
> BUNGE, 1999, p. 54.

> What complicates the problem and sometimes misleads the specialist is that (a) there are as many types of energy as there are large process genres; (b) there are as many concepts of energy as there are general physics theories; (c) the general concept of energy, or just energy, is so general, that it belongs to metaphysics or ontology; and (d) consequently, the general principle of energy conservation is also philosophical, although it has multiple physical roots.
>
> BUNGE, 1999, p. 56.

Accepting the thought-provoking perspective of considering the concept of energy, in one of its meanings, as a metaphysical construct, and therefore worthy of being studied, also, by specialists in philosophy, we can use it as a basis for other concepts such as 'concrete' and 'abstract', so highly relevant to any attempt to understand 'reality'.

Continuing with Bunge (1999) [24], who suggests 'identifying energy with mutability' (p. 54), we quote his postulate 1, comment 1, theorem and corollary 1:

> POSTULATE 1: All concrete (material) objects, and only those, are changeable.
>
> In other words: "For every x: x is concrete (material) if and only if x is changeable."
>
> In logical symbols:
>
> Comment 1
>
> We have identified "material" with "concrete." This convention is more customary in philosophy than in physics. According to it, the fields are as material as the stones.
>
> For example, photons are material in the philosophical sense of the word, though they do not have mass, solidity, or their own shape (attributes of matter before the advent of field physics).
>
> From Postulate 1, together with the Definition, it follows that:
>
> THEOREM For every x: if x is a material object, then x possesses energy, and vice versa.
>
> In summary:
>
> Here are two immediate consequences of this theorem. The first is:
>
> COROLLARY 1: Abstract (not concrete) objects lack energy.]
>
> BUNGE, 1999, pp. 54-55.

It seems necessary to recognize that such an approach may be debatable, even if to defend it he has resorted to the use of formal logic, since, for example, the fact that a person changes his mind does not imply that his mind is something concrete.

In this context, we think that we could ask ourselves whether abstract objects have an existential entity beyond human thought, or, in other words, it might be asked whether an abstract "object" can be properly said to "exist".

As an example we show an approach to the concept of geometric point, according to which 'it is an "entity of reason' – i.e. an intersubjective perception with broad consensus – without physical existence', (Calderero 2019) [25]

According to de la Pienda (1992) [26]:

> It could be reasoned that: if everything that exists is material, is my concept of the geometric point also material? If everything material is temporal space and, therefore, three-dimensional and changing, how many dimensions does my concept of the geometric point have and what does it change into? If it has three dimensions, then the concept of the geometric point is a contradiction, an absurdity. However, it is a key concept in Euclidian geometry, whose services to science do not seem very questionable.]
>
> DE LA PIENDA, 1992, p. 5.

Despite renouncing the effort to find a satisfactory definition of the basic concept 'geometric point' after an intense study of various sources, we cannot, without seriously violating intellectual rigour, deny that, in some way, 'the point exists', although it seems that its existence is ineffable, and it therefore resists all description.

We have arrived at a critical point in our article: the bonds we humans establish with different real beings (in the broadest sense of what is 'real', not in a reductionist reference to what is only material) are not necessarily linked to a full understanding of their nature. Therefore, we can establish valuable relationships with the different types of realities, inanimate beings, living beings, artifacts, material constructions, concepts, systems, people, etc., without the need to fully understand them or define them with absolute precision. Which does not prevent us from experiencing the need for the, always partial and incomplete, representations of reality to adjust as much as possible to whatever the essence, nature and properties of things are.

After the conceptual revision carried out, and given that it is not possible to define the "spiritual" unambiguously, we are going to make an approach in the sense of trying to clarify the differences between the artifacts moved by AI and the beings that according to different cultural traditions could be considered, at less in part, spiritual and that we dare to call "people". In this sense, in a propaedeutic way, we can mention as characteristic notes of a spiritual being that, at least, is immaterial, irreplaceable, timeless, indelimitable, capable of initiative, responsible, capable of unpredictability. Table I shows us some of these differences.

III. The "Fuzziness" and Mutability of Reality

Up to this point, we have repeatedly used the words 'real' and 'reality' without having used any precision to mitigate the quite probable risk that they might be interpreted in different ways.

Without trying to define both words by reverting to reductionism and anticipating our affirmative response to the question 'Is there anything that can be called "reality"?', (Calderero & Calderero, 2017, p. 99) [19], a few approximations may be outlined in the form of a postulate, which may be useful in communicating somewhat effectively:

- There is something susceptible to being called 'reality'.
- There is 'reality' and there are 'realities'. It is absurd to deny the 'existence' of everything and of something unknown or unknowable.
- The reality of something and its personal or consensual interpretation are different realities.
- The perception and description of a reality are themselves realities, although no coincidence exists between the two or with the reality that is being perceived or described.
- There are subjective realities, the images, concepts, or descriptions (narrative, coded, depicted graphically, and audible, etc.) that one constructs independently, or because of, from existing stimuli.
- There are objective realities whose existence and properties do not depend on human manipulation or interpretation; as an

TABLE I. Some Analogies and Differences Between Devices Driven by AI and Beings with Spirituality

	...a device driven by artificial intelligence...	**...a being, totally or partially spiritual (which may be called a 'person')...**
When communicating with their surroundings...	...uses codes or predetermined symbols that are recognisable through its programming.	...can use open procedures with the ability to improvise and intuit.
In their relationship with polysemy and synonymy...	...cannot process them beyond their most obvious senses unless it has a huge amount of data and possible combinations and a very high level of programming.	...can fluidly utilise connotative and metaphorical language to proceed, obviously according to the degree of the person's intellectual and cultural learning.
Regarding its responsibility for its operations (actions) we can say that...	...has no responsibility.	...has responsibility as an essential feature, although it may be greater or lesser according to certain conditions.
The 'intelligence' of...	...is cognitive, only for recognition and comparison with stored information.	...is cognitive and affective, with no clear distinction between those dimensions.
The behaviour of...	...is predictable.	...is not always predictable.
When processing, storing, and using information...	...does so at great speed and with much precision.	...must exert certain effort.
Decision-making for...	...is fast and neutral, fully restricted to programming instructions and closed criteria.	...is creative and usually occurs after weighing repercussions of the decision in other areas, such as the morality of the actions.
Regarding moving, acting...	...can only imitate living beings within the parameters of its manufacturing and design.	...can be original, unprecedented.
The work performance of...	...is delivered in complete alignment with requirements and specifications.	...tends to depend on the degree of fulfilment of certain conditions.
Faced with concepts like 'compassion', 'affection', or such...	...cannot react, as it 'ignores' the concept and the practice.	...can feel referenced or affected.
Regarding the care of people and things, contributing to improving humanity and nature...	...can be very effective if its design and maintenance are focused in that direction.	...will undertake it as far as their education, capabilities, beliefs, ideologies, etc. allow.
Regarding the construction, use, and interpretation of data...	...can only move within the limits allowed by its design and the nature of the information.	...has options that enable them to question all the related dimensions and imagine new ones.
Faced with concrete occurrences...	...cannot act until they have been transformed into 'understandable' data.	...can act without needing all the information; in fact, the interpretation, response, and assessment of consequences can be taken into consideration at a glance, even if there is difficulty in describing what happened.
The language of...	...must be exclusively denotative.	...may be connotative, metaphorical.

example, any animate or inanimate being whose existence, for millennia, was prior to the development of the human capacity for communication, will suffice.

- It is highly likely that new realities or dimensions, undetected at a given time due to the permanently limited scope of technology and methods to conceive and represent reality, may be discovered, as has always occurred to date.
- The postulate 'In every rigorous scientific investigation, the possibility of the existence of unknown variables, dimensions, realities must be considered' may be deemed legitimate.
- Scientific and investigative work, whose goal consists in advancing towards the discovery of unknown, empirical, conceptual, or any other type of realities, is important. The very nature of research seems to demand the need for institutions to allocate proportionate resources to the study of 'the unknown'.

An inherent difficulty in understanding reality is the paradox that it is reality itself which impedes our progress towards understanding it, because:

> in each science the ideal is objectivity, but from day to day reality forces us to take into consideration accessible information that is less secure, but employable in our reasoning and in our computers.
>
> KAUFMAN & ALUJA, 1987, p. 35. [27]

Along these same lines, it is interesting to consider Popper (1976, as cited in Velarde, 1991) [28]:

> It is always undesirable to strive to increase precision itself – especially linguistic precision – as it leads, in general, to a lack of clarity and a waste of time and effort in the preliminary stages, which often prove useless as they are overtaken by the actual progress of the matter: one must never try to be more precise than what the situation of the problem demands.
>
> VELARDE, 1991, p. 10.

Furthering this idea, which we might call 'the proportionality of the perceptive-cognitive effort', let us consider what the appropriate distance from which to 'see', or to thoroughly grasp, a painting would be. If we place ourselves at a great distance, we might not even see it physically; as we approach, we will see it increasingly better, until we pass a certain point of inflexion when we will be too near and we

no longer see it, as our view can only encompass a fragment. This metaphor can probably be applied to the discussion of the degree of specialisation, specification or precision with which one approaches the object of research or investigation. Could it not be that by one's trying to gain precision in the exhaustive control of the process, the object of study was made 'fuzzy' and the loss of perspective made the knowledge sought difficult? The quality of perception of the object of research is not only influenced by the physical or conceptual distance at which the researcher places himself; the excess of data can also impair discernment and the possibility of relevant understanding. In this respect, it is convenient to remember that 'to counter communicationism or an excess of information we propose educating to practice fuzziness' (Caeiro, 2018, p. 170) [3]. This concept of 'fuzziness' is gaining popularity in decision-making in business. Could this concept be an interpretative key that helps to overcome pseudo-dichotomies of usage in the areas of artificial intelligence and spirituality?

In non-specialist fields, there is an excessive tendency to associate artificial intelligence with discrete data, without considering that it is possible to manage relatively diffuse information in AI, since 'fuzzy logic is a branch of artificial intelligence founded on the concept 'it's all a matter of degrees', which allows the management of vague or hard-to-specify information'], (Jerez, Jofré, & Burgos, 2006, p.11) [29]. In any case, we must remember that, regardless, 'fuzzy logic' continues to move within limited fields that allow handling 'fuzzy sets' whose edges are less defined than those of conventional sets, but that exist; they continue to require mathematical formalisms.

Uncertainty, "approximateness", are an essential part of scientific knowledge; no one doubts that the results of scientific investigations are merely statistical, despite which it is frequent in science teaching to present theories to students as closed, indisputable postulates; today, for example, it is still taught in classrooms that 'in the absence of application of a non-balanced force (Fnet = 0), an object at rest remains at rest, and an object in motion remains in motion with constant velocity (constant speed and direction)'] (Wilson & Buffa, 2003, p. 106) [30]. This principle enjoys huge credibility among students and professors, given its 'unequivocal verifiability'; the statement can be proven true sufficiently reliably in any teaching laboratory, since the experimental data can be considered 'evidence'. However, when the assertion is taken to its ultimate end, there exists precisely the problem that nobody, ever, has been able to prove it more than apparently, and only by using relative terms, always referring to systems considered theoretically immobile. The concept 'at rest' is a theoretical construct, as is the situation of an 'absence of application of a non-balanced force'; it rather seems that any material particle, no matter how minute, is never at rest and cannot not be subjected to any force. If it is impossible to even consider that the components of solid material substances are at rest, or for it to be appropriate to do so with extensive objects, since as they necessarily turn with the Earth and it around the Sun and the Sun also moves, it seems legitimate to conclude that physical reality is continuously changing. We could say that mutability is an essential property of reality. Additionally, chemical transformations are continuous processes which, by their very nature, cannot cease to occur at any time.

On the other hand, it would not be rigorous to omit from any study of reality the influence that intangible 'human spirit' elements have on its configuration. Without delving too deeply, one can deduce that all large and small human works owe their existence to the fact of having been conceived, at least broadly, in a human mind; it seems the creative process goes from the idea, an intangible, to a practical realisation. Everything seems to indicate that there is a strong bond, which is difficult to detect and make concrete operationally, between immaterial realities (without physical substance) and concrete material realities.

We resist accepting as indisputable the postulates of those scientific currents that maintain that human mental activity is completely explained by biochemical or bioelectric processes, closing the possibility of existing to other dimensions that cannot be processed by the instruments and methods of experimental science. We agree with Artigas (1984) [31] when he states that:

> The experimental science approach assumes a point of view in which the kinds of things that can be said, and therefore the kinds of entities that can be found, are predetermined. Concretely, experimental science does not extend, in principle, to spiritual realities; therefore, denying the spirit on the basis of these sciences is unsustainable scientism.
>
> ARTIGAS, 1984; quoted in ÁLVAREZ, 2019, [32], p. 61.

Perhaps human creativity is something more than just the original, unprecedented, unique response capacity to different stimuli and we should consider the need to accept the existence of "fields of consciousness". In this sense, assuming the possibility that they may be considered questionable, it is appropriate not to discard, and reflect on, the contributions of Grof (1999) [33]:

> Newtonian science is responsible for having offered us a very limited vision of human beings and their true potential. For about two hundred years it has dictated the criteria of what constitutes an acceptable experience and what is an unacceptable experience of reality. From its viewpoint, a 'normal' person is that which is capable of reproducing exactly the external objective world described by Newtonian science. Consequently, from that perspective, our mental functions are limited to receiving information provided by our sensory organs, storing it in the 'memory banks of our mental computer', and recombining the sensory data to create something new. Any significant departure from that perception of 'objective reality' – a consensual reality that the general population considers the only truth – is interpreted as the product of a runaway imagination or a mental disorder.]
>
> GROF, 1999, p. 18.

> Instead of speaking of discrete objects and empty spaces between them, today the universe is considered to be a continuous field of variable density. According to modern physics, matter is interchangeable with energy, and conscience – which is not limited to activities taking place inside our skull – forms a part of the same fabric of the universe.
>
> As the British astronomer James Jeans said over sixty years ago, the universe of modern physics resembles more a great thought than a giant super machine.
>
> GROF, 1999, p. 20.

From these ideas, we could deduce that:

> There seems to be no possibility of creating a better world through the mere outside intervention that does not include a deep transformation of human conscience.
>
> GROF, 1999, p. 308.

Taking as a reference the last two quotes, it does not seem necessary to demonstrate that the activity of personal beings, persons, such as human beings is significant when studying the transformations that physical reality experiences, because for over two million years humanity has had tools, coarse and fledgling yet useful, to accomplish diverse tasks. Since the dawn of prehistory, this tendency to transform reality has only grown. The current challenge is to delve into the psychic influence exerted by the human mind on other persons or animate beings. In the extremely vast field of human interactions one can observe, even at a glance, the influence that the mere presence of a person exerts on the thoughts, words and behaviour of others.

Assuming the aforementioned disparity exists between reality and its various representations, and considering that all life depends in large measure on the intellectual development of human beings, it stands to reason that it is of great interest that relevant decision-making should avoid the many reductionisms abounding in intellectual

circles, and whose consequences are suffered in all spheres – personal, family, social, political and economic. In this sense, we venture to propose the creation, and their use in research and instruction, of indicators, or failing that, clarifiers, to be used as correction factors when introducing in theories inherent in different sciences sufficient elements of uncertainty that 'invite' the consideration of the possible presence of unknown variables and the subsequent in fieri nature of any knowledge or scientific discovery. We deem it would be highly profitable, in all senses of the word, to promote a questioning of the many blindly assumed and disseminated topics, distancing ourselves as far as possible from the sense of writings published by Cervera University in 1827 to King Ferdinand VII, of Spain:

> Far from us this dangerous novelty of thinking (reflecting), which has caused harm for a long time, finally rupturing, with undeniable effects, tainting custom, totally disrupting empires and religion in every part of the world.
>
> GACETA DE MADRID, núm. 53, p. 211. [34].

This is an explicit statement which, despite being in the distant past, continues to inform certain areas of academics, albeit more subtly; proof of this might be found in the solemn declaration that is pronounced by the rector in several Spanish universities when awarding the academic cap to new doctors:

> Receive the Book of Science which it is your duty to teach and advance, and let it be a sign and a warning to you that, however great your ingenuity, you must render obedience and reverence to the doctrine of your teachers and predecessors.

At the risk that it might be considered a 'contradiction in terminis', we believe it would be very convenient, and maybe necessary, to lay the foundation of a possible branch of epistemology, 'administration of the unknown' without which it seems unlikely that 'the unknown' could become known. The attitudes, strategies and protocols that could lead to discovering, exploring, formalising and in that case utilising unknown but maybe intuited dimensions, beings or relationships, are worthy of consideration as a relevant element of research, even in those cases when results are not foreseeable in the short and medium term.

IV. The Interdimensional Unity of Reality

Despite accepting the huge difficulty, even impossibility, of associating unequivocal meaning to the term 'reality' and being aware of the 'fuzziness' of the concept and of 'conceptualised objects', we will reflect on two of its possible qualities: it is multifaceted (multidimensional) and has intrinsic and essential unity.

According to the Royal Spanish Academy dictionary [35], 'dimension' is:

> 1. Fem. Aspect or facet of something.
>
> 2. Fem. Measure of magnitude in one direction.
>
> 3. Fem. Physics. Each magnitude that fixes the position of a point within a space.
>
> 4. Fem. Physics. Each of the fundamental magnitudes: time, length, mass, and electrical charge, which express a physical variable.
>
> DICCIONARIO DE LA REAL ACADEMIA ESPAÑOLA (R.A.E.). Dimensión

We once again perceive the lexical inadequacy we must necessarily work with in our efforts to reach significant progress in detecting reality, in its processing, and in the development of valuable links that might overcome the limitations of any representation system.

In the arena of the previously poorly-named 'exact sciences', there are also similar ambiguities to be found:

> The concept of dimension can be considered of great importance in mathematics, because it is a source for understanding other concepts in that discipline, but it is also difficult to conceptualise due to its complexity in being defined, as well as considering that inside mathematics it is used in various ways depending on the area where one is working.
>
> PÁEZ, ORJUELA, & ROJAS, 2008, p. 1. [36]

The matter becomes complicated when trying to define the meaning of 'dimension' in moral, linguistic, economic, political, artistic arenas, and others. We are faced with a term with a considerable degree of polysemy. Add to this complexity the fact that the different 'dimensions' an object 'has', be it physical or not, probably only exist in the collective mind of the scholars that have coined the corresponding constructs.

Without seeking to demonstrate it, we point out that the history of human knowledge seems to endorse the idea that every being or construct can be contemplated from different angles and analysed from such different views as the ones evaluating the existence or not, and the degrees, of its 'size', 'position', 'weight', 'colour', 'chemical composition', 'economic value', 'symbolic significance', 'deterioration', 'origin', 'evolution', 'morality', and so many others that it would be improper, and impossible, to try to list here. In accordance with this idea, it would seem reasonable to broaden the object of study of the different sciences in order to avoid excessive focus on the study of certain areas considered as belonging to a few disciplines and 'non-existent' in others. To that effect, it would behove us all to reflect on the eventual benefits of favouring, at least at high levels, interdisciplinary, interfacultative research, and consequently, teaching.

Conversely, is it acceptable that, in examining the same object of study, researchers of various disciplines arrive at incompatible conclusions, and that both are considered correct? Wouldn't it be more reasonable to consider that some of the research or the paradigms or the theories considered, or all of them, must be revised until the incompatibility is removed, or until it is proven that it was only apparent, or new theories were generated that could align said disparate results conceptually or operationally? If, as presumably will occur, we accept the reasonability of the preceding statement of non-contradiction, we are somehow assuming the unity of reality, which we might identify as 'unidiverse', echoing the statement by Martínez-Otero (2009) [37] with their 'theory of unidiverse intelligence', in which intelligence is presented as 'a unitary and multiple faculty' (p. 1)

Assuming the suitability of encouraging interdisciplinary synergy, it would be appropriate to consider the benefits of broadening the current widespread, and lauded, STEM paradigm, an acronym for science, technology, engineering, and mathematics, recommended for students as future 'professional outlets', to a more inclusive STEAM that incorporates an A for 'arts', in line with Maeda (2013) [38], who holds that:

> Design creates the innovative products and solutions that will propel our economy forward, and artists ask the deep questions about humanity that reveal which way forward actually is.
>
> Government agencies are beginning to acknowledge that art and science – once inextricably linked, both dedicated to finding truth and beauty – are better together than apart.
>
> MAEDA, 2013, p.1

There are some authors, far from settling for science and art being better together than apart, who suggestively propose to 'broaden the categories of science and technology to those of art' (Caeiro, 2018, p. 168) [3]. We feel that, dispensing with the rhetorical format of the statement, it offers great depth, understanding that it attempts to alleviate the reductionist effect of the limited scientific paradigm to the study of the section of reality detectable by human senses and by what we might consider their 'extension', technology. By broadening the categories of science and technology to those of art, we understand that they would be enriched by the deep, and mysterious, intangible,

and spiritual elements of art, not only with the aesthetic, harmonious elements of algebraic expressions used in a great portion of scientific content, but with aesthetic aspects that are not minor due to the importance of beauty in human life, and not reducible to pragmatic, utilitarian aspects without loss of dignity.

On the other hand, it should be pointed out that, since artistic practice is a human activity we could describe as essential, the term 'arts' can also be interpreted as 'humanities', which leads us to overcome the possibly artificial barrier that separates 'sciences' from 'humanities'. Therefore, we should cease to consider them as a kind of 'benevolent concession' to be necessarily incorporated into academic, political or economic life in order to 'humanise' it so it does not appear to be too cold or 'stark'. According to the classic concept, the constant progress towards knowledge requires the seven liberal arts, as expressed by De la Iglesia (2001) [39]:

> Mercury, fulfilling his duty as husband, presented Philology with the principle dowry of his divine wedding gift: seven wise servants to help his beloved to continue her constant progress towards knowledge. Three of them (Grammar, Rhetoric, and Dialectic) would attend to perfect her internal world; the other four (Arithmetic, Music, Geometry, and Astrology) would enable a wider understanding of the external world.
>
> DE LA IGLESIA, 2001, p.131

In line with reality's interdimensional unity, we understand that studies that lead to its illumination should be interdimensional 'per se', so that the conjunction of research instruments and methods can guarantee a minimum of interdisciplinary synergy, by which the risk of conceptual and methodological reductionism can be reduced as much as possible. To that end, we understand that from the onset of their studies the instruction of young students and researchers should have a 'fractal' nature, so that the desired interdisciplinary synergy for academic and research projects does not become a requirement implanted a fortiori, but a natural consequence of the global mentality personally acquired by each young person from the earliest age. We understand that in this presumably desirable process of interdisciplinary instruction, artificial intelligence will be called upon to occupy a relevant role by enabling the processing of enormous quantities of data, simultaneously showing the convergence and divergence among them, and enabling what we might call a 'macro hyper-textual and hyper-relational' language; thus, by considerably reducing the task of searching for information, encouraging intellectual activities that are less mechanical and more oriented towards deeper and intangible aspects of human beings.

Paraphrasing Moreno, Carrasco, and Herrera Viedma (2019) [40] when they state that '[t]he main objective of this work, therefore, is to define a formal framework that allows market orientation to be effective in the context of big data' (p. 7), we feel that a desirable challenge would be to define a framework that would allow human activity's spirit orientation to be effective within the context of big data.

Furthering the required synergy between the areas pertaining to data and those corresponding to the spirit and humanities, we agree with Lope Salvador, Mamaqi and Bordes (2020) [41] regarding the need to put into

> ...perspective three large matters: 1) the need to update the set of digital competencies for the efficient analysis of massive amounts of data as the basis for the professional profile of the cyber-analyst; 2) the assumption that AI is offering new epistemological opportunities in social sciences and humanities that must be leveraged; and 3) the implementation of procedures derived from AI for the effective analysis of the content of scientific publications when evaluating quality and innovation.]
>
> SALVADOR, MAMAQI, & BORDES, 2020, p. 85

V. CONCLUSIONS

After having carried out the planned conceptual review of some relevant terms, and having reflected both on the concept of "reality" and its "fuzziness" and on its interdimensionality and unity, and with the intent of contributing to a greater integration of human knowledge in the pursuit of a better quality of life, as understood in all possible senses and not only as relates to physical wellbeing, we summarise below some of the possible conclusions that could be drawn.

Regarding AI, we consider that:

- With its enormous power for data processing, it can be greatly helpful in alleviating, where possible, the proverbial lexical insufficiency related to spirituality, psychology, philosophy and the humanities in general.
- It can be enormously helpful in highlighting the eventual lack of foundation of all the principles, axioms, or postulates of a philosophical, moral and spiritual nature whose main, or only, value resides in their high degree of dissemination in the media, and which are uncritically assumed by many and often socially enforced. Semantic engineering could be a magnificent tool for discriminating the genuinely spiritual from bastard concepts likely attributed to the spirit or the spiritual despite possibly having a different, even fraudulent, origin.
- It can generate codes, labels, morphology and syntax capable of processing great volumes of information with many significant nuances that can be placed 'above' the human mind's level of understanding and handling. Thus, overcoming conventional languages, they can establish other valuable links with reality, interacting with it without the reductionisms that conventional languages might insert.
- It can help to design, or design directly, instruments for interpreting reality that are more 'empathetic' with its intangible aspects, or taken as such.

Regarding it being interdisciplinary, we believe it would be highly beneficial for humanity and nature, and therefore in mutual benefit of a desirable environmental balance, that:

- In areas related to spiritual instruction, religious or not, and in educational institutions, academia, especially in universities specialising in humanities, sufficient knowledge of the nature, properties, and scope of science, technology, and, concretely, artificial intelligence, should be promoted so that, particularly the new generations might perceive the eventual mutual benefits to be derived from working together, abandoning the paradigm of suspicion of the dehumanisation that many people attribute to technology. It would be beneficial if, in the study of scientific-experimental fields, and in university degrees, intellectual rigour were sufficiently encouraged, so that new generations being educated would avoid the reductionism of viewing as the only valid sources of knowledge those that exclusively use data.
- It would be very positive for human and environmental development to take significant steps towards the creation of interdisciplinary, interfacultative, interuniversity teams of research that would study in depth, and with the most advanced AI techniques, how to integrate the methodologies pertaining to generating knowledge (broadly) corresponding to experimental areas with those of non-experimental areas, especially those centred on the study of the intangible.

We understand that the scope of these conclusions is not limited to the eventual interest of specialists in AI and Sciences of the Spirit but may be useful for students of disciplines such as Linguistics, Psychology, Philosophy, Morals, Theology and Educational Sciences.

References

[1] Naciones Unidas. Transformar nuestro mundo: la Agenda 2030 para el Desarrollo Sostenible. Resolución de la Asamblea General A/71/1, aprobada el 25 de septiembre de 2015. http://www.un.org/ga/search/view_doc.asp?symbol=A/RES/70/1&re-ferer=/english/&Lang=S

[2] Francisco. Carta Encíclica Laudato Si' del Santo Padre Francisco. Sobre el cuidado de la casa común. Mayo, 24 de 2015.

[3] Caerio, M. "Ser persona en la sociedad del conocimiento y el espectáculo". *Arte y políticas de identidad*, 18, pp. 159-176, 2018, doi: https://doi.org/10.6018/reapi.336061.

[4] APA Dictionary of Psychology. Disponible en https://dictionary.apa.org/.

[5] Spearman, C. *The Abilities of Man: Their Nature and Measurement*. New York, USA: McMillan, 1927.

[6] Thorndike, E.L. "Intelligence and its Uses". *Harper´s Magazine*, 140, 227-235. 1920.

[7] Martínez-Otero Pérez, V. "Alcance socioeducativo de la teoría de la inteligencia unidiversa". *Holos*, vol. 5 (), pp. 116-126. 2016, doi: 10.15628/holos.2016.4731.

[8] Zubiri, X. *Inteligencia y logos* (Vol. 2). Madrid, España: Alianza Editorial. 1982

[9] Calderero, J. F., Aguirre, A. M., Castellanos, A., Peris, R, Perochena, P. "Una nueva aproximación al concepto de educación personalizada y su relación con las TIC." *Teoría de la Educación. Educación y Cultura en la Sociedad de la Información*, vol. 15, no. 2, pp. 131-151. 2014, doi: https://doi.org/10.14201/eks.11890.

[10] Calderero, J. F., Perochena, P. y Peris, R. "Estudio integrador de elementos significativos en la formación de maestros. Una propuesta para la autoevaluación docente". *Tendencias pedagógicas,* vol. 25, pp. 121-148. 2015.

[11] Gardner, H. *E Frames of Mind. The Theory of Multiple Intelligences.* Nueva York, Basic Books. 1983. Versión castellana: E*structuras de la Mente. La Teoría de las Inteligencias Múltiples.* México, Ed. Fondo de Cultura Económica, 2001.

[12] Ferrater Mora, J. *Diccionario de Filosofía* 5º edición Editorial Sudamericana. Buenos Aires, Argentina. 1965.

[13] Rizo-Patrón de Lerner, R. "Reconsiderando la relación entre la naturaleza y el espíritu". *Escritos de filosofía*, no. 3, pp. 265-287. p. 276. 2015.

[14] Torralba, F. *Inteligencia espiritual.* Barcelona, España: Plataforma Editorial. 2010.

[15] Arias, R., & Lemos, V. "Una aproximación teórica y empírica al constructo de inteligencia espiritual". *Enfoques*, vol. 27, no. 1, pp. 79-102, 2015.

[16] Bostrom, N. "Introduction—The Transhumanist FAQ: A General Introduction" in *Transhumanism and the Body* (pp. 1-17). New York: Palgrave Macmillan, 2014.

[17] Valera, L., & Marambio, J. T. A. Posthumanismo e hibridación. *Pensamiento. Revista de Investigación e Información Filosófica, 75(283* S. Esp), pp. 307-319, 2019, doi: https://doi.org/10.14422/pen.v75.i283.y2019.016.

[18] Marcos, A. "Filosofía de la naturaleza humana". *Eikasia. Revista de Filosofía*, vol. 6, no. 35, pp. 181-208, 2010.

[19] Calderero Hernández, J. F., Calderero de Aldecoa, A. *Filosofía y sentido común.* Madrid, España: Sekotia. 2017.

[20] Domínguez Berenjeno, E. L. "Arqueología y territorio: de la 'interpretación arqueológica' al 'dato histórico'". *SPAL, Revista de prehistoria y Arqueología de la Universidad de Sevilla vol. 10*, pp. 109-122, 2001, doi: http://dx.doi.org/10.12795/spal.2001.i10.05

[21] Alegre Ramos, M.P. *Sistemas operativos monopuesto* (2ª ed.). Madrid, España: Paraninfo. 2019.

[22] cheverría, R. (2017). *Ontología del lenguaje.* Buenos Aires, Argentina: Ediciones Granica SA.

[23] Leocata, F. (2003). Persona, lenguaje, realidad. EDUCA.

[24] Bunge, M. "La energía entre la física y la metafísica". *Revista de Enseñanza de la Física*, vol. 12, no. 1, pp. 53-56. 1999.

[25] Calderero J. F. jfcalderero.wordpress.com. (27 de abril de 2019). ¿Existes o eres invención humana? https://jfcalderero.wordpress.com/2019/04/27/existes-o-eres-invencion-humana/

[26] de la Pienda, J. A. "Cientifismo marxista". *Espíritu: cuadernos del Instituto Filosófico de Balmesiana,* vol. 41, no. 106, pp. 167-184. 1992.

[27] Kaufman, A., & i Aluja, J. G. *Técnicas operativas de gestión para el tratamiento de la incertidumbre.* Barcelona, España: Hispano Europea. 1987.

[28] Velarde, J. *Gnoseología de los sistemas difusos.* España: Servicio de Publicaciones de la Universidad de Oviedo. 1991.

[29] Jerez López, P., Jofré Nuñez, C. y Burgos Letelier, D. "Lógica borrosa aplicada en ADR de Europa, Asia y Latinoamerica". Ph.D. dissertation, Facultad de Economía y Negocios, Universidad de Chile, Santiago, Chile, 2006.

[30] Wilson, Jerry D., Buffa Anthony J., Lou, Bo. *Física.* (5ª ed.) México: Pearson educación. 2003.

[31] Artigas, M. (1984). Máquinas pensantes y conocimiento humano. En Actas del III Simposio de Teología Histórica (7-9 mayo 1984). Confrontación de la teología y la cultura. Valencia: Facultad de Teología San Vicente Ferrer. p. 392.

[32] Álvarez-Álvarez, J. J. (2019). Apuntes para el repensamiento de la enseñanza de la Arquitectura. La cuestión epistemológica y la necesidad de una razón ampliada. Revista de Arquitectura, vol. 21, no. 2, pp. 57-67.

[33] Grof, S. *La mente holotrópica.* Barcelona, España: Kairós. 1999.

[34] de Madrid, Gaceta. 3 de mayo de 1827. Gaceta de Madrid, núm. 53.

[35] REAL ACADEMIA ESPAÑOLA: Diccionario de la lengua española, 23ª ed., [versión 23.3 en línea]. https://dle.rae.es

[36] Páez, J., Orjuela, C., & Rojas, C. (2008). "El concepto de dimensión: errores y dificultades" in 9° Encuentro Colombiano de Matemática Educativa. Valledupar, Colombia, 2008.

[37] Martínez-Otero Pérez, V. "Propuestas educativas derivadas de la teoría de la inteligencia unidiversa". *Revista Iberoamericana de Educación*, vol. 50, no. 1. 2009, doi: http://dx.doi.org/10.35362/rie5011851.

[38] Maeda, J. "Stem + art= steam". *The STEAM Journal*, vol. 1, no. 1, Art. 34. 2013, doi: http://dx.doi.org/10.5642/steam.201301.34

[39] de la Iglesia, J., "Las artes liberales en la Biblioteca Real del Escorial: dos antecedentes iconográficos". In *El Monasterio del Escorial y la pintura: actas del Simposium*, 1/5-IX-2001 (pp. 119-164). Real Centro Universitario Escorial-María Cristina, El Escorial, España, 2001, pp. 119-164.

[40] Moreno, C., González, R. A. C., & Viedma, E. H. (2019). "Data and Artificial Intelligence Strategy: A Conceptual Enterprise Big Data Cloud Architecture to Enable Market-Oriented Organisations". *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 5, no. 6, pp. 7-14. 2019, doi: http://dx.doi.org/10.9781/ijimai.2019.06.003

[41] Salvador, V. L., Mamaqi, X., & Bordes, F. J. V. "La Inteligencia Artificial: desafíos teóricos, formativos y comunicativos de la datificación". *Icono14*, vol. 18, no. 1, pp. 58-88, 2020.

José Fernando Calderero Hernández

José Fernando Calderero Hernández is a Doctor of Philosophy and Educational Sciences (Complutense University of Madrid, Spain. 2003) and has a degree in chemistry (University of Salamanca, Salamanca, Spain. 1972). He is currently a professor of "Theory and Practice of Educational Research" and "Life Cycles and Communication in the Family" at the International University of La Rioja (UNIR), President of the Education Chapter of AEDOS and Vice President of the Foundation "Parents for Excellence, Padrex". He has been Dean of the Faculty of Education of the UNIR, Deputy Director of the Education Area of the Villanueva University of Madrid, professor of the University of Navarra and of the Bachelor of Education of the University of Wales, speaker in courses for teachers and managers in Spain and America, with a professional performance of 27 years as a manager and university professor and 24 years as a director and teacher of high schools. He has written several educational books and academic articles and supervised some doctoral theses. His lines of work and research are focused on personalized education, family education and the development of critical sense.

Why the Future Might Actually Need Us: A Theological Critique of the 'Humanity-As-Midwife-For-Artificial-Superintelligence' Proposal

Marius Dorobantu*

Vrije Universiteit Amsterdam (Netherlands)

Received 3 February 2021 | Accepted 19 June 2021 | Published 30 July 2021

Abstract

If machines could one day acquire superhuman intelligence, what role would still be left for humans to play in the world? The 'midwife proposal,' coming from futurists like Ray Kurzweil or James Lovelock, sees the invention of AI as a fulfillment of humanity's cosmic destiny. The universe 'strives' to be saturated with intelligence, and our cyborg descendants are much better equipped to advance this goal. By creating AI, humans play their humble, but instrumental, part in the grand scheme. The midwife proposal looks remarkably similar to modern Christian anthropology and cosmology, which regard humankind as "evolution becoming conscious of itself" (Pierre Teilhard de Chardin), and matter as having a predisposition to evolve toward spirit (Karl Rahner, Dumitru Stăniloae). This paper demonstrates that the similarity is only superficial. Compared to the midwife hypothesis, Christian theological accounts define the cosmic role of humanity quite differently, and they provide a more satisfactory teleology. In addition, the scientific and philosophical assumptions behind the midwife hypothesis – that the cosmos is fundamentally informational, that it intrinsically promotes higher intelligence, or that we are heading toward a technological singularity - are rather questionable, with potentially significant theological and ethical consequences.

Keywords

Artificial Superintelligence, James Lovelock, Novacene, Ray Kurzweil, Singularity, Theological Anthropology.

DOI: 10.9781/ijimai.2021.07.005

I. The Future Still Needs Us, but Only for a While

At the turn of the century Bill Joy, then Chief Scientist at Sun Microsystems, published in Wired his famous article, "Why the Future Doesn't Need Us" [1]. It represented a moment of chilling public realization that some of the most dystopian future scenarios were no longer mere sci-fi fantasies. In Joy's view, the convergence of robotics, genetic engineering, and nanotechnology could constitute an existential threat for humanity, rendering us an "endangered species."

Today such an article would hardly make the headlines. But what is perhaps most remarkable about Joy's piece is that it came at the end of a so-called AI winter, a period of declining hype, interest, and funding for AI, triggered by the field's failure to deliver on its naïve and grandiose initial promises. AI research had emphatically taken off in the 1950's, with resounding successes in mathematics and game-playing, things notoriously difficult for humans. Computer programs could solve problems, prove theorems, and play some strategy games at human-level performance. Since such highly intellectual tasks proved relatively easy to replicate, it was widely believed that more mundane abilities, like perception or locomotion, would not pose too many problems. This is well illustrated by how MIT scientists in the 1960's thought that they could collectively solve computer vision as a summer project [2]. Even more ambitiously, it was predicted that within just one generation we would witness the first machine endowed with human-level intelligence [3]. It is needless to say that that wasn't the case. The 'mundane' capabilities of human intelligence proved much more difficult to replicate than the 'intellectual' ones, something known as Moravec's paradox [4]. The failure to deliver on those big promises deeply affected the public perception of AI, and the possibility of truly intelligent machines was again relegated to the realm of fantasy and sci-fi. This mood was still in place in the late 1990's and early 2000's, around the time when Joy's article saw the light.

However, unbeknown to the public, new technologies were developing, and a new AI-revolution was brewing. The first wave of AI had operated from the assumption that the human mind processes information sequentially, like a computer algorithm, by manipulating a finite set of symbols by means of logical operations. Teach a computer program enough of these symbols, rigorously define the rules for how they should be combined, and a powerful enough computer should start to reason like a human. To the contrary, the approach behind second-wave AI, generally known as machine learning, was rather different. Instead of assuming any theory of human cognition, it tried to roughly imitate the organ of human cognition, that is, the brain, and see whether this could lead to intelligent behavior. The most successful branch of this approach, deep learning, stacks many layers of artificial neural networks together and trains them to recognize patterns, presumably in a somewhat similar fashion to how a human brain learns to recognize patterns in the world.

Deep learning lies behind the biggest AI successes of the new millennium. Our banks, our stock markets, our airports, our

* Corresponding author.
E-mail address: m.dorobantu@vu.nl

smartphones, and our social media feeds are largely powered by second-wave AI algorithms. AI programs today are capable of learning from scratch how to play strategy games. They can compose symphonies and coherent texts, and they can also make decisions in spite of incomplete information. All these arguably resemble learning, creativity, or intuition, capacities once considered to be uniquely human. They are also critical in another respect: they open the possibility for AI to re-design itself and augment its intelligence. Concerns have been formulated that if AI were to reach human-level competency at programming AI, this could trigger a positive feedback loop. Such an AI could build a more capable version of itself, that would, in turn, be more competent at building an even better version and so on. This scenario is known as an "intelligence explosion," a term introduced by I.J. Good [5]. AI would thus quickly reach superintelligence, a level way above all humans collectively.

Nick Bostrom masterfully demonstrates that such an artificial superintelligence (ASI) would be impossible to contain, unless we could insure from its inception that it is friendly towards us [6]. In AI, this is known as the famous 'alignment problem,' and is notoriously difficult: how to make sure that an ASI will have goals that are aligned with our own? The reality is that we have almost no way of anticipating how such an ASI might see us, or what its goals might be. It is sometimes said that it would be as intelligent compared to us as we are compared to ants. If this turns out to be true, then ASI would inhabit a very different perceptual and conceptual world, and it would perhaps have very different kinds of thoughts, that we couldn't ever comprehend.

Whether such an ASI would be a 'true intelligence' or a mere automaton is of little relevance for the outcome of this scenario. John Searle famously distinguishes between 'strong AI' and 'weak AI': while the former would be capable of thinking and would have what we call consciousness, the latter would only be a perfect simulation of intelligence, with no subjectivity, thoughts or phenomenal experience [7]. Because we do not have an answer yet to the question of how physical matter can bring about conscious experience – something known as the "hard problem of consciousness" [8] –, we also do not know if ASI would be strong or weak AI.

Whether or not ASI would be a *someone* instead of *something* is a fascinating philosophical and theological question, but with little implication for how such a future might play out for us in practice. In fact, too much talk of artificial consciousness might be a red herring, distracting us from the real possibility that AI might soon outsmart us. The question whether AI can be conscious is different from the question whether it can be more intelligent – understood as competent – than us, and too much attention to the former might obscure the astonishing AI progress in the latter. It is a chilling realization that ASI might be possible without having consciousness and a mind, but current AI algorithms already achieve super-human levels in many tasks without possessing such things. For practical purposes, it doesn't really matter whether there is a 'real mind' and intentionality behind an AI program, as long as it can outperform humans in increasingly more relevant cognitive tasks. Before too long, we might wake up to a reality where machines have reached human-level intelligence, or even ASI.

Talking now of ASI might seem exaggerated, but many experts with first-hand experience in the field believe that it is only a matter of decades until we AI gets there. An often-cited 2016 survey of 550 AI experts reveals that most of them expect human-level AI between 2040 and 2050, and ASI by 2080, the latest [9].

Machines are currently dependent on us to program and power them on, but this might cease to be the case with future, more autonomous, robots. Thinking of such intriguing scenarios helps us realize that a future without humans is perfectly possible, as Bill Joy eerily predicted. Some have reacted to this realization by ringing the alarm bells about the stakes of AI research. In 2015, Stephen Hawking, Elon Musk, and dozens of AI experts signed an open letter, calling for more research on the societal impact of AI and against mindlessly building something that cannot be *a priori* well-understood, let alone controlled [10].

Others, however, seem to be more reconciled with a future that does not necessarily include humans. Our cyborg descendants are going to replace us and that is absolutely fine. In the natural world, it is normal for better adapted life forms to thrive and replace their progenitors, so eventually this was going to happen to our species anyway. The special thing about our demise, though, is that it will mark the transition from biological to artificial life. This is what is referred to in the paper as the 'midwife proposal.'

There are two ways in which this transition can play out. In Ray Kurzweil's account, humans merge with machines, giving birth to hybrid cyborg species [11]. In James Lovelock's prediction, humans gradually fade away like an endangered species, being gradually replaced by intelligent forms of artificial life [12]. In either case, our hyper-intelligent descendants will go on expanding to other planets beyond the solar system, and then beyond the galaxy into the rest of the universe, saturating it with intelligence. This is something that we are not equipped to do, due the limitations of our biology. Humanity would thus be casted in the midwife role in this cosmic evolutionary drama.

II. Kurzweil's Singularity

Inventor and futurist Ray Kurzweil predicts that AI will reach human-level by the year 2029, which will be demonstrated by the first program to pass the Turing Test [11, p. 200]. But there will be no reason for AI to stop at human level. Beyond that point, its capability would continue to grow exponentially, leading to a so-called 'singularity' in 2045. This is when humans would merge with machines into a new type of hybrid being, many orders of magnitude more intelligent than *Homo sapiens* [11, p. 136].

The notion of technological singularity is rooted in mathematics and physics, denoting a point of no return in history. Just as it is impossible to communicate information from beyond the event horizon of a physical singularity, also known as a black hole, so too it is impossible to predict what history will look like after the technological singularity. It was mathematician John von Neumann who first associated the concept of singularity with technological progress [13], while Vernor Vinge popularized it in the 1990s [14].

The main principle behind Kurzweil's bold prediction is what he calls "The Law of Accelerating Returns." He regards human history as a history of technological evolution, anticipating that technological progress will continue forward at an accelerated rate. This acceleration is to a certain extent already accounted for by Moore's Law. In the 1960s, Intel co-founder Gordon Moore correctly predicted that the density of transistors in integrated circuits would continue to double at regular intervals [15], thus making computing technology exponentially cheaper and more powerful. But Kurzweil's Law of Accelerating Returns goes beyond this, conjecturing that the exponential growth tendency applies to a wider variety of evolutionary systems [16]. In other words, Moore's Law would only represent a particular case of the more general Law of Accelerating Returns, according to which technological progress in the world occurs at an accelerated rate.

The 2029 singularity would be followed by a complete merging between biological (human) and artificial intelligence. The resulting super-intelligent being would combine the best of each realms: "The Singularity will represent the culmination of the merger of our biological thinking and existence with our technology, resulting in a world that is still human but that transcends our biological roots.

There will be no distinction, post-Singularity, between human and machine or between physical and virtual reality" [11, p. 9].

Post-singularity cyborgs will supposedly combine the unique traits of human intelligence – the plasticity and massive parallelism of our brain, our mind's ability to hold contradictory thoughts etc. – with the advantages of AI – the huge speed of electronic circuits, increased memory storage, the ability to instantly copy skills/programs from one machine to the other etc. These godly successors of ours will live lives that are incomprehensible to us and will have powers that we cannot even think of. They will proceed to fulfill the universe's "ultimate destiny," that is, to be infused with intelligence: "In the aftermath of the Singularity, intelligence, derived from its biological origins in human brains and its technological origins in human ingenuity, will begin to saturate the matter and energy in its midst. It will achieve this by reorganizing matter and energy to provide an optimal level of computation [...] to spread out from its origin on Earth. [...] Whether our civilization infuses the rest of the universe with its creativity and intelligence quickly or slowly depends on [the speed of light's] immutability. In any event the "dumb" matter and mechanisms of the universe will be transformed into exquisitely sublime forms of intelligence [...]. This is the ultimate destiny of the Singularity and of the universe" [11, p. 21].

III. Lovelock's Novacene

Futurist James Lovelock is best known for his Gaia hypothesis, which posits that the Earth is a self-regulating system, like a giant organism, and that the emergence of life is part of our planet's evolutionary 'strategy' to keep cool against the increasing energy output of the Sun [17]. In his book, *Novacene: The Coming Age of Hyperintelligence*, Lovelock argues that we are quickly approaching the end of the Anthropocene and the beginning of a new geological age, the Novacene. The defining characteristic of this new age is the emergence of electronic life capable of directly transforming energy into information.

The main assumption behind Lovelock's argument is that the cosmos is informational at its most fundamental level. This would explain what he sees as the consistent positive selection of intelligence throughout cosmic evolutionary history, leading to the emergence of humans, the first 'understanders' of the cosmos. The 'informational assumption' would also neatly explain the anthropic principle, namely, the apparent fine tuning of physical laws and constants for the emergence of intelligent life. If this is true, then humans are the first consciousness of the cosmos. Through us, the universe awakens and becomes self-aware. Through our hyper-intelligent cyborg descendants, the universe will accomplish its last evolutionary stage, that of transforming all its matter and energy into information [12, p. 123].

The Gaian super-organism is therefore a nursery for the cosmos' self-awareness. Lovelock identifies three key moments in the history of Gaia, each corresponding to the beginning of one distinct geological age. The first was the evolution of organisms capable of photosynthesis, enabling Gaia to capture sunlight and store its energy. By releasing oxygen, these organisms set the stage for the emergence of more complex life, culminating with humans. The second key moment was the invention of steam-engines. Through technology, humans triggered the second stage, the Anthropocene, marked by Gaia's capability to transform the stored solar energy from fossil fuels into useful work. The third stage, the Novacene, begins when humans invent machines that are capable of learning and re-designing themselves, with a widespread ability to transform sunlight into information. Later, these machines will pursue the conversion of all the physical matter into information.

In Lovelock's story, these electronic life forms come in perfect continuation with biological life, emerging through the same processes of Darwinian selection. Instead of the natural selection that characterizes the evolution of biological organisms, cyborgs will undergo a purposeful selection, marked by a quick correction of harmful mutations.

Zooming out one further level, the Novacene can be regarded as a necessary evolutionary 'strategy' by Gaia. In a few hundred million years the Sun is poised to become a Red Giant type of star, dramatically increasing its radiation output. Biological life won't be able to keep the planet cool in such conditions anymore, hence the need for super-intelligent electronic life forms, with technologies powerful enough to tackle this challenge.

How will the Novacene unfold? Similarly to Kurzweil, Lovelock speculates that the evolution of cyborgs will be accelerated. Differently from Kurzweil, he does not think that humans will be able to keep up. Speed is one main quantitative differentiator between human and artificial intelligence. Electrical current can travel much faster through electronic circuits than through a brain's wetware, potentially 1 million times faster. Lovelock settles for the more conservative prediction that cyborgs will think around 10.000 times faster than we think [12, p. 81]. Even so, this would be the same ratio as that between humans and plants.

Another big difference, this time a qualitative one, would come from the very different nature of cyborg intelligence. AI will allegedly be more intuitive than human intelligence, because it wouldn't be built around speech. As the story goes, humans developed speech as a necessary evil. While it has been of tremendous evolutionary value for our species, it has also narrowed our thinking to the current linear, step-by-step logic, that we are all familiar with. Cyborgs won't have a speech-driven intelligence, and they will likely communicate telepathically with each other, retaining speech only to be able to communicate with us [12, pp. 96-103]. This will supposedly free up their intelligence from the chains of discursive thinking to realms and possibilities that for us are difficult to even imagine.

While it seems right to suppose that cyborgs will develop a very non-humanlike type of intelligence [18], the choice to label human intelligence as discursive and AI as intuitive looks rather odd. If anything, it should be the other way around. Modern psychology posits that there are more (usually two) modes of human cognition, of which only one is sequential and discursive: Daniel Kahneman's system 1 and 2 [19], Philip Barnard and John Teasdale's propositional and implicational [20], or Jonathan Haidt's elephant and elephant rider [21], to name only a few. Iain McGilchrist masterfully demonstrates that the intuitive, right hemisphere type of intelligence is much more involved in human cognition that the rational, left hemisphere type [22]. If the history of AI so far is of any relevance, machines are actually more likely to master the discursive type of intelligence and to struggle with the intuitive one. Hubert Dreyfus' famously argued that computers might have a chance of replicating our conscious, "knowing-what" mode of cognition, but it would incomparably more difficult for them to master the unconscious, "knowing-how" mode [23].

As for how these cyborgs will behave towards us, their 'parents,' Lovelock manifests an unbridled optimism, foreseeing no likely power struggle between biological and electronic life. Cyborgs will be so far ahead of us that we could not possibly be a match for them, just as other animals are currently no match for us. However, they will probably keep us around, in order to help keep Gaia cool. Future AI will be intelligent enough to realize that the biggest threat for its existence comes from the increasing heat from the Sun. The best strategy against that, at least for the time being, would be to preserve

biological life and allow it to continue cooling the planet, perhaps with some help from cyborg-invented technologies.

But is it realistic to believe that the overheating from the Sun, not due for a few hundred million years, would be so high on the immediate agenda of ASI? Would ASI not be able to invent better technologies for that job? Or, better, could the cyborgs not decide to keep all biological life *except* for humans? After all, no other species has 'worked' harder than we have against the temperature homeostasis throughout the Anthropocene. Lovelock is confident that our descendants will not bother to exterminate us, but rather treat us like pets, an idea going back to Apple's co-founder, Steve Wozniak. In this scenario humans will therefore enjoy a "peaceful retirement" [12, p. 119].

How does one so easily come to terms with such a chilling possibility? The answer is that it largely depends on the value ascribed to human life and to a human presence in the world. In Lovelock's system, the cosmos we live in is intelligence-oriented, but not necessarily anthropic-oriented. It manifests a selective preference for intelligent organisms and a predisposition toward being converted from matter to information. Up until now, humans have been the best at promoting that. But cyborgs will be far better equipped for this task than humans, so we should simply accept that "we have played our part" and rejoice "as wisdom and understanding spread outwards from the Earth to embrace the cosmos" [12, p. 130].

IV. Intelligence as a Cosmic Goal

Kurzweil and Lovelock's visions differ slightly in the kind of cyborg descendants they predict. While Lovelock forecasts fully cybernetic organisms, Kurzweil envisages a hybrid between humans and AI along transhumanist lines, but even in such a case, these hybrids would arguably be anything but human. In both scenarios, humanity plays the critical, but historically limited, role of a midwife for the more advanced forms of cyber life.

At a more profound level, something that both these extreme versions of cheerful non-anthropocentrism have in common is the assumption that our universe has a built-in purpose to promote intelligence, ultimately understood as total informatization of matter and energy.[1] Both versions of the midwife proposal presuppose that the universe has a destiny, and that humans play a pivotal role in the cosmic drama of fulfilling that destiny. This assumption takes the midwife proposal from the realm of scientific ideas, where both authors claim to be dwelling, and transports it very close to the realm of the religious.

The idea that humans have a central role to play in the world is not very different from what most major religions are claiming. In Christian anthropology, for example, this is known as the functional interpretation of human distinctiveness and the image of God (*imago Dei*) [25]. In the functional view of the *imago Dei*, humans are being regarded as God's representatives in the created universe, appointed to exercise stewardship and dominion over the world [26]. In the posthuman scenarios of Kurzweil and Lovelock, humans fulfill the role of midwives for the new hybrid intelligence that will "saturate the matter and energy in its midst" [11, p. 21]. This vocational language is strikingly reminiscent of God's commandment to the first humans to "fill the earth and subdue it" [27].

[1] It is worth pointing out that from a strictly scientific perspective this claim is highly dubious, if not a category error altogether. Matter/energy and information are not interchangeable quantities, because they exist at different levels of conceptual 'zooming-in.' Matter and energy are to a certain extent equivalent, as shown by Einstein's famous E=mc2 equation, but information is a totally different level of characterization of reality. Matter cannot be 'informatized,' as Lovelock suggests. Moreover, if the information conservation principle is correct [24], then no new information can be created in the universe.

The jump from AI and technological singularity to Christian cosmology and the image of God may seem abrupt, but the two realms have much more in common than what the casual reader might initially think. Firstly, the two discourses operate with surprisingly similar notions and structures, such as prophetism [28], a dualistic view of the world, or transcendent promises for the future. Transhumanist views of the future and 'AI apocalypticism' draw substantively on religious thought, especially on Jewish and Christian apocalypticism, to the point where the former can reliably be traced back as a "legitimate heir[s]" to the latter [29]. They also share the belief in a clear periodization of future cosmic history, and in profound changes that are imminent for both humanity and the world [30].

Secondly, they are equally non-scientific. This is not intended as a derogatory judgement, but as a precise delimitation of the space in which the two discourses are meaningful. While modern theology openly admits its limited competence when it comes to scientific issues, secular ideologies, such as the midwife proposal, often present themselves as flowing naturally from scientific theory and observations of the world. But just as Christian theology departs from some clearly defined assumptions about the existence of God and the meaning of life, so too such ideologies make strong assumptions about the purpose of the universe that are eminently non-scientific in nature.

The relegation of humans to a position of evolutionary ancestors or midwives of the true "heirs of the world" [31], namely super-intelligent cyborgs, is a scenario fueled not so much by irrefutable scientific observations, as by Kurzweil and Lovelock's personal beliefs regarding the teleology of the cosmos and the inevitability of the hybridization between human and machine.

From the conjecture that the universe is primed to favor intelligent forms of matter and energy, the ideology behind the midwife proposal goes on to affirm that the intrinsic purpose of the universe is to be saturated with intelligence. Biological life, of which humans are the apex, could never accomplish the cosmic goal. AI will therefore take over at some point and continue unhindered to use more and more resources to augment its intelligence until finally the entire cosmos becomes saturated with intelligence. As further shown, far from being neutral, these views are scientifically dubious, theologically problematic and morally dangerous.

V. The Midwife Proposal Is Scientifically Dubious

Both Kurzweil and Lovelock present their ideas as valid scientific paradigms that could account for our observations of reality. Lovelock's Gaia hypothesis is allegedly nothing but Darwinian evolution extended to the cosmic scale, while Kurzweil's Law of Accelerating Returns is a clever generalization of Moore's law. However, their proposals are not as purely scientific as they are presented to be. Instead, as further shown, they are a complicated mixture between the scientific and the non-scientific, between good and crazy ideas, as pointed out by Douglas Hofstadter's scathing criticism of Kurzweil: "[I]t's a very bizarre mixture of ideas that are solid and good with ideas that are crazy. It's as if you took a lot of very good food and some dog excrement and blended it all up so that you can't possibly figure out what's good or bad. It's an intimate mixture of rubbish and good ideas, and it's very hard to disentangle the two" [32].

The observation that our universe seems to favor intelligent life-forms may hold some validity, but it looks to be heavily biased toward the particular evolution of life on *our* planet. On a local scale, it seems indeed true that more intelligent life forms are often evolutionarily fitter. From this perspective, developing AI that potentially surpasses our intelligence may indeed be our worst evolutionary mistake, as pointed out by Stephen Hawking [33].

On a cosmic scale, however, the conjecture that the universe selects for intelligence hardly finds any support. If intelligence is indeed a great attractor to which the universe is irresistibly drawn, how come our observations so far have not revealed a ubiquity of intelligent forms of existence? The contrast between this expectation to find myriads of technological civilizations in the observable universe and the apparent silence in our observations is known as the Fermi paradox. In the summer of 1950, Italian physicist Enrico Fermi supposedly asked: "Where is everybody?" [34]. The universe is already approximately 13.8 billion years old [35]. Given its presumed predisposition to favor intelligence, one would expect that at least a few other life forms more advanced than our own have already gone through the technological singularity and/or have given birth to space-expanding AI. The process of saturating the universe with intelligence should thus be noticeably underway. But from what has been so far observed, it is not. This begs the question: why?

Both Kurzweil and Lovelock choose the same answer, which is intuitive, but astronomically improbable: we are the first ones [11, p. 357] [12, pp. 3-5]. We do not observe extra-terrestrial intelligences because there are none yet. But if the universe is capable of producing countless intelligent species throughout its existence, and even more if it is primed to do so, how likely is it that we are the first of this colossally big series?

The Fermi paradox is one of the most complex scientific and philosophical problems of our times, and its scope is simply too wide to be discussed in detail in this paper, so let us only point out that there exist a variety of proposed solutions. One of them, for example, formulated by Anders Sandberg, Eric Drexler, and Toby Oord, conjectures that we might be indeed the only technological civilization in the observable universe, but not because we happen to just be the first one, but rater because the emergence of intelligent life is incredibly improbable [36]. This looks like a more promising explanation, but it pushes back against the idea that the universe is 'striving' towards intelligence and informatization, which is a crucial assumption of the midwife proposal. How could only one occurrence of intelligent life across such a vastness of space and time be typical of anything?

Kurzweil actually acknowledges that his solution of humanity being the first intelligence of the cosmos has a dramatically low mathematical probability. But instead of embracing the 'improbability of intelligence' explanation, he chooses to invoke a modified version of the anthropic principle: some civilization *has* to be the first, and if our observations suggest that we are the first, then it must be true, in spite of the astronomical unlikelihood. This allows him to continue to postulate the built-in longing for intelligence of the cosmos, and the critical role of humanity in driving the universe toward "complexity and order:" "[W]e are in the lead. That's right, our humble civilization with its pickup trucks, fast food, and persistent conflicts (and computation!) is in the lead in terms of the creation of complexity and order in the universe" [11, p. 357].

Besides the casual romanticism in Kurzweil's above quote and the apparent bias to equate human civilization with the United States of America, it becomes clear that the midwife proposal also has many philosophical and theological implications.

VI. The Midwife Proposal Is Theologically Problematic

Even if the cosmos systematically promotes intelligent life-forms, an assumption shown above to be questionable, it seems a step too far to conclude that the universe has the goal of becoming more complex, ordered, or more saturated with intelligence. Firstly, if anything, the second law of thermodynamics paints the opposite picture of a universe inevitably evolving from order to chaos on a global scale. Secondly, speaking of goals and meaning is usually not the province of science, but of philosophy and theology.

On a subtler level, in spite of its apparent radical non-anthropocentrism, the midwife proposal actually supports its own version of human exceptionalism: it is, after all, humans who have the seminal role of making the transition between biological and cybernetical stages of evolution. As already noted, this appears to bear some similarities with functional accounts of human distinctiveness in theological anthropology.

The idea that humanity represents an inflection point in cosmic history is not something new. In fact, it has been emphasized so much in theological anthropology, to the point of inviting accusations of anthropocentrism [37], [38]. The very notion that the universe 'awakes' for the first time in human beings is not strange to Christian evolutionary thought. Paleontologist and Jesuit priest Pierre Teilhard de Chardin refers to humankind as "evolution becoming conscious of itself" [39]. Roman-Catholic theologian Karl Rahner believes that matter has an intrinsic predisposition to evolve toward spirit, and that humans are the apex of this process of the universe's self-realization [40], [41]. Is it possible that the midwife proposal and Christian theology speak of the same idea when describing humankind as *the* vehicle through which matter ultimately becomes information/spirit?[2] Rather not.

While the two might bear some superficial resemblance, the Christian version differs in at least three significant ways. Eastern-Orthodox theologian Dumitru Stăniloae provides a compelling and illuminating account of this Christian idea of spiritualization of matter, which is worth quoting at large: "The world was created in order that man, with the aid of the supreme spirit, might raise the world up to a supreme spiritualization, and this to the end that human beings might encounter God within a world that had become fully spiritualized through their own union with God. The world is created as a field where, through the world, man's free work can meet God's free work with a view to the ultimate and total encounter that will come about between them. For if man were the only one freely working within the world, he could not lead the world to a complete spiritualization, that is, to his own full encounter with God through the world. God makes use of his free working within the world in order to help man, so that through man's free work both he and the world may be raised up to God and so that, in cooperation with man, God may lead the world toward that state wherein it serves as a means of perfect transparence between man and himself" [43].

The first major difference between the Christian perspective and midwife posthumanism consists of the presence versus the absence of God throughout the process of spiritualization of the universe. Christian thought is, of course, theistic. It affirms, as Stăniloae makes it clear, that humans by themselves could never lead the world to fulfill its full potential of spiritualization without divine collaboration. Unsurprisingly, the midwife hypothesis makes no explicit mention of a deity, but it should not be too easily labeled as atheistic. The kind of cosmic harmony that Kurzweil speaks about when describing a universe infused with intelligence can better be categorized as a form of pantheism, rather than atheism.

Secondly, an even more profound difference between the two visions concerns their teleology. In the Christian perspective, the spiritualization of matter is not a goal in itself. Rather, it is only relevant within the larger picture of the free relationship of love between God and humans. Stăniloae explicitly articulates that in Christian theology the universe is only valuable "with a view to the ultimate and total encounter" between creator and creature. The ultimate purpose of the

[2] This parallel was first pointed out in [42].

world is therefore to facilitate this encounter. But in order to fulfill this role, the world needs to achieve perfect transparence, hence the need for "supreme spiritualization."

The midwife proposal, on the other hand, exhibits a rather unconvincing teleology. If the purpose of the universe is indeed informatization, as Lovelock conjectures, or the saturation with intelligence, as Kurzweil proposes, one could legitimately ask: why? Furthermore, it might be useful to pursue the midwife proposal to the absurdity of its final outcome. What would happen after the goal of informatization, complexity and order is physically achieved? Would the universe continue to exist forever in that state of perfect equilibrium, known in physics as the 'Big Freeze' [44], and synonymous to a heat death? Or would it then explode again into a new universe through another Big Bang, in which case the question of meaning and teleology would simply be reported to the end of the next cycle? Although we could certainly imagine higher-level informational beings inhabiting such a transfigured cosmos, the question of purpose still remains. Without an eternal God, infinitely generating new knowledge and meaning, it is hard to imagine what else could give purpose to such beings. Although this might be due to inherent limitations in our current imagination, it could also signal a weak and unsatisfactory teleology from the part of the midwife hypothesis.

Thirdly, the midwife proposal can rightly be suspected to arise from a certain dissatisfaction with the human condition, hence the need to replace humans with more advanced beings that will "inherit the earth" [45]. Theological anthropology might have a few problems with this. All the monotheistic religions share the intuition that there is something special about humans. In Christian anthropology, this intuition is encapsulated in the doctrine of the *imago Dei*: humans bear in them the image of God. Moreover, Christian faith is built around the testimony that God became human through the incarnation of Jesus Christ. These two, the *imago Dei* and the incarnation, strongly imply that humans, limited and imperfect as they might be, are in a way *enough*.

The issue of teleology is again of critical importance. From an evolutionary perspective, it can be said that the purpose of the universe is indeed to 'awaken.' But this awakening does not need to entail the saturation of the cosmos with intelligence, by being transformed into an unthinkably big supercomputer. Instead, the universe awakens by naturally giving birth to a conscious entity, which possesses all the mental and moral capabilities necessary to become a recipient of divine revelation and enter into a relational covenant with God. In this case, the awakening process would further be validated and fulfilled in the incarnation of the divine *Logos*, what Teilhard de Chardin calls "the Omega Point" of cosmic evolution [39, pp. 250-275]. Christopher Fisher summarizes this point very well: "In theological perspective, the appearance of personal subjective self-awareness and transcendentality in human beings means that there is no need for another step in biological or cosmic evolution [...]: the process is complete (complete, in particular, in the incarnation itself), having reached the goal of opening material reality directly to conscious relationship with the Absolute" [46].

In patristic anthropology, and in particular in the writings of Maximus the Confessor, human beings are described as microcosms, miniature recapitulations of the entire cosmos [47], [48]. It is thus possible to affirm that with the emergence of the human person, the cosmos itself becomes conscious. Analogically, in the human response to God's calling to relationship, the cosmos itself is brought to fulfillment and potentially transfigured, as in Stăniloae's cosmology.

Christian anthropology acknowledges the limitations of human nature and the need to transcend them, but it suggests a radically different solution from the one advocated by the midwife proposal. Humans are called to transcend their nature through the pursuit of deification, or *theosis*. Far from being "little more than the Christian's alternative to human enhancement" [42, p. 340], *theosis* implies a radical transformation of human nature at its most profound level. According to Stăniloae's definition, *theosis* is the "greatest possible union with God wherein the fullness of God is stamped upon the human being, yet without the human being thereby being dissolved into God" [43, p. 89]. *Theosis* suggests transcending human nature by downsizing oneself through God's kenotic self-giving love, in contrast to the expansion of the self, entailed by Kurzweil's vision [42, p. 330].

Finally, it is interesting to observe how the midwife proposal still struggles to find a place for human distinctiveness in its story. Even though AI will eventually "match and then vastly exceed the refinement and suppleness of what we regard as the best of human traits," Kurzweil still struggles to find a feature that remains uniquely human: "There will be no distinction, post-Singularity, between human and machine or between physical or virtual reality. If you wonder what will remain *unequivocally human* in such a world, it's simply this quality: ours is the species that *inherently seeks to extend its physical and mental reach beyond current limitations*" [11, p. 9, my emphases].

Kurzweil does not explain how humans are different in this respect from the animals. After all, isn't this tendency to reach beyond the limits inherent to biological life, in general? In theological anthropology, this exocentricity of human nature is yet another mark of the *imago Dei*: we continuously strive, most of the time unconsciously, towards a destiny of fellowship with God in the *eschaton*, as beautifully described by Wolfhart Pannenberg [49]. But Kurzweil's vision predictably lacks such context. What causes this supposedly uniquely human restlessness? And what is its *telos*?

The midwife proposal has no answers to these questions. While it might bear some superficial resemblance to Rahner's openness to transcendence or with Pannenberg's exocentricity, it does not come even close to painting as coherent a picture as these theological proposals.

For Rahner, humans are indeed intrinsically open to transcendence, but this is only because of the pre-apprehension of the infinite reality that is the transcendent God [40, p. 33]. Similarly, the exocentricity proposed by Pannenberg is a metaphysical drive toward fulfilling a vocational destiny in the encounter with God, who is the source of both the drive and of direction [50]. In both theological accounts, the typically-human longing for transcendence only makes sense if there exists an infinite transcendence, namely God, to long for in the first place. One might not agree with the inherent theological assumptions, but the system is at least self-coherent.

For Kurzweil, to the contrary, this thirst to exceed limitations is ultimately empty of content: it exists only because it is a necessary prerequisite for developing the kind of technology that can saturate the cosmos with intelligence. When compared to the theological accounts of human distinctiveness as *imago Dei*, the midwife proposal looks shallow and highly unconvincing.

VII. The Midwife Proposal Is Morally Dangerous

While the theological and philosophical weakness of the midwife proposal might not be too imperative, its ethical ramifications are genuinely dangerous and in need of urgent clarification. Firstly, if saturating the universe with intelligence is the ultimate cosmic goal, then it follows that everything should be evaluated according to the measure in which it advances or hinders this process. This is already visible in how some in the bio-liberal movement are pleading that becoming posthuman is a moral imperative [51]. If such a view

becomes mainstream, how will the value of each individual human being come to be judged? Would human life be valuable in itself, or only insofar as it contributed to the progress towards the singularity or the Novacene?

Secondly, and more generally, if technological progress towards the so-called awakening of the cosmos is regarded as the ultimate goal of *everything*, then this view has the potential to substantially alter our current ethical definitions of good and evil. 'Good' would in such a case be any effort that promotes the technological singularity or the Novacene. Any resistance to the mainstream paradigm, such as refusing to augment oneself, could become synonymous with moral evil and sanctioned accordingly.[3] This might sound dystopic, but it is a result of following the thread of Kurzweil and Lovelock's ideas to their logical conclusion. Neither of them intentionally proposes such a chilling moral system, but certain readings of the midwife hypothesis could nevertheless lead in that direction.

Thirdly, as discussed earlier, there is no reason to believe that super-intelligent cyborgs would necessarily be strong AI. It is equally likely, or perhaps even more likely, that they would turn out to be mindless automatons, superbly capable to outsmart us, but totally incapable of feeling or thinking anything. Could a world populated and radically transformed by such automatons be one we would deem as 'good'? Without a doubt, no. Even if human evolution is taken as a proof that the universe selects for intelligent life-forms, not any kind of intelligence gets positively selected. If the universe 'wants' indeed to 'awaken', it is *our* type of intelligence that it ultimately needs, one that is also accompanied by subjective experience and understanding. Otherwise, what would be the point?

Undoubtedly, even the most convinced believer of the midwife proposal would agree that the universe doesn't seek to be saturated with a mindless type of intelligence. To make any sense, the midwife proposal needs strong AI. There is thus a hidden built-in assumption that our cyborg descendants will be intelligent not in the way that current AI programs are, but that they will also be centers of selfhood and phenomenal experience, truly capable of thinking, understanding, and feeling. As of today, we are still completely in the dark regarding this possibility. We simply do not know whether machines could ever become conscious. Although AI has made significant progress toward replicating human intelligence 'on the outside', from what we know it has made *zero* progress toward acquiring consciousness, or an inside-outness.

Thus, even from a non-theistic utilitarian perspective, which judges things to be good or bad depending on how efficiently they promote the wellbeing of conscious agents, the midwife hypothesis is deeply problematic in a weak AI scenario. The hypothesis relies therefore on the possibility of strong AI, something that is often not made explicit enough in its manifestos.

VIII. Conclusion

This succinct overview of the midwife proposal and its assumptions enables some provisional conclusions. First and foremost, the character of the ideas behind it is highly speculative. Although they are presented as sound scientific truths, even the briefest of analyses reveals this to be an overstatement. This is also reflected by the skepticism with which the scientific and AI communities continue to regard such views [53].

That being said, Kurzweil's core idea that technological progress is accelerating, even though perhaps at a slower pace than he suggests, is still a valuable observation. Similarly, Lovelock's creative imagery of how the Novacene world will look like is very powerful: electronic animals grazing solar-powered plants, robots so small and fast that they inhabit and study the quantum world, or cyborgs thinking so fast that "the experience of watching your garden grow gives you some idea of how future AI systems will feel when observing human life" [12, p. 82]. However, any value of such images and ideas is outbalanced by their questionable science, unclear teleology and dangerous ethical implications.

The principles upon which the midwife proposal is based are far from being merely 'neutral' scientific and technical observations. As shown in the paper, they stem from a philosophically dubious understanding of the purpose of the universe and the role of humans. The idea that humanity has a seminal role to play in cosmic evolution by developing AI might be intriguing, but at a closer inspection it is exposed to be lacking support, depth, and a coherent teleology, especially in comparison with theological accounts of human distinctiveness. Finally, the *doctrine* that humanity has only a midwife role to play in the larger narrative of the cosmos evolving toward hyper-intelligence comes with heavy and rather indefensible ethical implications for the value of human life and the very definition of good and evil.

The midwife proposal therefore makes claims that, although couched in scientific language, belong more to the realm of religious discourse. Even when judged solely by their internal logic and coherence, such anthropologies and cosmologies fare much worse than their Christian counterparts, on which they draw. Before becoming too quickly resigned to a fate of collective demise, we should perhaps stop and wonder whether the future doesn't, in fact, badly need us.

[3] This is similar to the outcome of a thought experiment known as Roko's Basilisk, where a future omnipotent artificial superintelligence (ASI) decides to retroactively reward those who promoted its existence and punish those who did not (by resurrecting them through avatar reconstruction and then torturing those avatars eternally), in order to motivate us in the present to invest everything we have in the pursuit of ASI for fear of retribution. Although this scenario might sound anything from logically flawed to hilarious, it has caused a lot of anxiety among members of the LessWrong virtual community of rationalists [52]

Acknowledgment

This publication was made possible through the support of a grant from the Templeton World Charity Foundation. The opinions expressed in this publication are those of the author and do not necessarily reflect the views of the Templeton World Charity Foundation.

References

[1] B. Joy, "Why the future does not need us," *Wired*, 2001. Accessed Jan. 28, 2021. [Online]. Available: https://www.wired.com/2000/04/joy-2/.

[2] S. Papert, "The summer vision project," *MIT AI Memos (1959 - 2004)*, 1966. Accessed Jan. 28, 2021. [Online]. Available: http://people.csail.mit.edu/brooks/idocs/AIM-100.pdf.

[3] S. J. Russell and P. Norvig, *Artificial intelligence: A modern approach*, Upper Saddle River, NJ: Prentice Hall, 2003, p. 21.

[4] H. Moravec, *Mind children: The future of robot and human intelligence*, Cambridge, MA: Harvard University Press, 1988, p. 15.

[5] I.J. Good, "Speculations concerning the first ultraintelligent machine," *Advances in Computers*, vol. 6, pp. 31-88, 1965, doi: 10.1016/S0065-2458(08)60418-0.

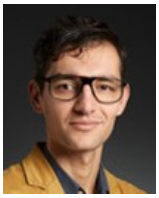
[6] N. Bostrom, *Superintelligence: Paths, dangers, strategies*, Oxford, UK: Oxford University Press, 2014.

[7] J. Searle, "Minds, brains and programs." *Behavioral and Brain Sciences*, vol. 3, no. 3, pp. 417–457, 1980, doi: 10.1017/S0140525X00005756.

[8] D. Chalmers, "Facing up to the problem of consciousness," *Journal of Consciousness Studies*, vol. 2, no. 3, pp. 200–219, 1995.

[9] V. C. Müller and N. Bostrom, "Future progress in artificial intelligence: A survey of expert opinion," in *Fundamental Issues of Artificial Intelligence*,

V. C. Müller Ed. Berlin, DE: Springer, 2016, pp. 553-571.
[10] M. Sparkes, "Top scientists call for caution over artificial intelligence," *The Telegraph*, 2015. Accessed Jan. 28, 2021. [Online]. Available: https://www.telegraph.co.uk/technology/news/11342200/Top-scientists-call-for-caution-over-artificial-intelligence.html.
[11] R. Kurzweil, *The singularity is near: When humans transcend biology*, New York, NY: Viking Penguin, 2005.
[12] J. Lovelock, *Novacene: The coming age of hyperintelligence*, London, UK: Allen Lane, 2019.
[13] M. Shanahan, *The technological singularity*, Cambridge, MA: MIT Press, 2015, p. 233.
[14] V. Vinge, "The coming technological singularity: How to survive in the post-human era," in *Vision-21: Interdisciplinary Science and Engineering in the Era of Cyberspace*, G. A. Landis Ed. NASA Publication CP-10129, 1993, pp. 11–22.
[15] G. E. Moore, "Cramming more components onto integrated circuits," *Electronics*, vol. 38, no. 8, 1965.
[16] R. Kurzweil, *The age of spiritual machines*, New York, NY: Viking Press, 1999, pp. 30-32.
[17] J. Lovelock, "Gaia as seen through the atmosphere," *Atmospheric Environment*, vol. 6, no. 8, pp. 579–580, 1972, doi: 10.1016/0004-6981(72)90076-5.
[18] M. Dorobantu, "Human-Level, but Non-Humanlike: Artificial Intelligence and a Multi-Level Relational Interpretation of the Imago Dei," *Philosophy, Theology and the Sciences*, forthcoming 2021.
[19] D. Kahneman, *Thinking, fast and slow*, New York, NY: Farrar, Straus and Giroux, 2011.
[20] P. J. Barnard and J.D. Teasdale, "Interacting cognitive subsystems: A systemic approach to cognitive-affective interaction and change," *Cognition and Emotion*, vol. 5, no. 1, pp. 1–39, 1991, doi: 10.1080/02699939108411021.
[21] J. Haidt, *The happiness hypothesis: Finding modern truth in ancient wisdom*, New York, NY: Basic Books, 2006.
[22] I. McGilchrist, *The master and his emissary: The divided brain and the making of the Western world*, Yale University Press, 2009.
[23] H. L. Dreyfus and S. L. Dreyfus, *Mind over machine: The power of human intuition and expertise in the era of the computer*, Free Press, 1986, pp. 16-51.
[24] B. Zhang, Q.-Y. Cai, M.-S. Zhan and L. You, "Information conservation is fundamental: Recovering the lost information in Hawking radiation," *International Journal of Modern Physics D*, vol. 22, no. 12, 1341014, 2013, doi: 10.1142/S0218271813410149.
[25] Genesis 1: 27.
[26] M. Cortez, *Theological anthropology: A guide for the perplexed*, London, UK: T&T Clark, 2010, pp. 21-24.
[27] Genesis 1: 28.
[28] B. Singler, "Existential Hope and Existential Despair in AI Apocalypticism and Transhumanism," *Zygon: Journal of Religion & Science,* vol. 54, no. 1, pp. 156-176, 2019, doi: 10.1111/zygo.12494.
[29] R. M. Geraci, "Apocalyptic AI: Religion and the promise of artificial intelligence," *Journal of the American Academy of Religion*, vol.76, no. 1, pp. 138–166, 2008, doi: 10.1093/jaarel/lfm101, p. 158.
[30] R. Cole-Turner, "The Singularity and the rapture: Transhumanist and popular Christian views of the future," *Zygon: Journal of Religion & Science*, vol. 47, no. 4, pp. 777-796, 2012, doi: 10.1111/j.1467-9744.2012.01293.x.
[31] Romans 4: 13.
[32] W. Grassie, "Politics by other means: Science and religion in the twenty-first century,"Bryn-Mawr, PA: Metanexus, 2010, p. 282.
[33] G. Dvorsky, "Stephen Hawking says A.I. could be our 'worst mistake in history,'" *Gizmondo*, 2014. Accessed Jan. 28, 2021. [Online]. Available: https://io9.gizmodo.com/stephen-hawking-says-a-i-could-be-our-worst-mistake-in-1570963874.
[34] E. M. Jones, "'Where is everybody?': An account of Fermi's question," *Los Alamos National Laboratory*, USA, 1985, doi: 10.2172/5746675. Accessed Jan. 28, 2021. [Online]. Available: https://www.osti.gov/servlets/purl/5746675.
[35] Planck Collaboration, "Planck 2015 results. XIII. Cosmological parameters," *Astronomy & Astrophysics*, vol. 594, article no. A13, 2016, doi: 10.1051/0004-6361/201525830, p. 32.
[36] A. Sandberg, E. Drexler, and T. Ord, "Dissolving the Fermi paradox," *ArXiv*, 2018. Accessed Jan. 28, 2021. [Online]. Available: https://arxiv.org/abs/1806.02404.
[37] G. Kaufman, "The concept of nature: A problem for theology," *Harvard Theological Review*, vol. 65, no. 3, pp. 337–366, 1972, doi: https://doi.org/10.1017/S0017816000001619.
[38] K. Junghyung, "Christian anthropology in an age of science: Between anthropocentrism and non-anthropocentrism," *The Expository Times*, vol. 129, no. 12, pp. 547-553, 2018, doi: 10.1177/0014524617753327.
[39] P. Teilhard de Chardin, *The phenomenon of man*, New York, NY: Harper & Row, 1961, p. 220.
[40] K. Rahner, *Foundations of Christian faith: An introduction to the idea of Christianity*, translated by W. V. Dych, New York, NY: Crossroad, 1982.
[41] O. Putz, "Evolutionary biology in the theology of Karl Rahner," *Philosophy and Theology*, vol. 17, no. 1&2, pp. 85-105, 2005, doi: 10.5840/philtheol2005171/25.
[42] R. Cole-Turner, "Theosis and human enhancement," *Theology and Science*, vol. 16, no. 3, pp. 330-342, 2018, doi: 10.1080/14746700.2018.1488526, p. 336.
[43] D. Stăniloae, *The experience of God: Orthodox dogmatic theology. Vol. 2 – The world: Creation and deification*, translated by I. Ionita and R. Baringer, Brookline, MA: Holy Cross Press, 2000, p. 59.
[44] A. V. Yurov, A. V. Astashenok, and P. F. Gonzalez-Diaz, "Astronomical bounds on a future Big Freeze singularity," *Gravitation and Cosmology* vol. 14, pp. 205–212, 2008, doi: 10.1134/S0202289308030018.
[45] Matthew 5: 5.
[46] C. L. Fisher, "Animals, humans and x-men: Human uniqueness and the meaning of personhood," *Theology and Science*, vol. 3, no. 3, pp. 291-314, 2005, doi: 10.1080/14746700500317289, p. 307.
[47] Maximus the Confessor, "Epistolae VI," in *Patrologiae Cursus Completus*, J.-P. Migne Ed. Tomus 91, Paris, FR: Garnier Fratres, 1865, col. 429.
[48] L. Thunberg, *Microcosm and mediator: The theological anthropology of Maximus the Confessor* (2nd edition), Chicago and La Salle, IL: Open Court, 1995, pp. 132-143.
[49] F. LeRon Shults, *Reforming theological anthropology: After the philosophical turn to relationality*, Grand Rapids, MI: Eerdmans, 2003, p. 235.
[50] W. Pannenberg, *What is man? Contemporary anthropology in theological perspective*, Philadelphia, PA: Fortress Press, 1970, p. 220.
[51] M. Walker, "Ship of fools: Why transhumanism is the best bet to prevent the extinction of civilization," in *H±: Transhumanism and Its Critics*, G. R. Hansell and W. Grassie Eds. Philadelphia, PA: Metanexus Institute, pp. 94-111, 2011, p. 95.
[52] B. Singler, "Roko's Basilisk or Pascal's? Thinking of Singularity Thought Experiments as Implicit Religion, vol. 20, no. 3, pp. 279-297, 2018, doi: 10.1558/imre.35900.
[53] J. Rennie, "Ray Kurzweil's slippery futurism," *IEEE Spectrum*, 2010. Accessed Jan. 28, 2021. [Online]. Available: https://spectrum.ieee.org/computing/software/ray-kurzweils-slippery-futurism.



Marius Dorobantu

Dr. Marius Dorobantu is a research associate and lecturer at the Vrije Universiteit Amsterdam, working on issues of science & religion. His current project, financed with a "Diverse Intelligences" grant from the Templeton World Charity Foundation, is entitled "Understanding Spiritual Intelligence: Computational, Psychological and Theological Approaches." Formerly he obtained a PhD in ethics from the University of Strasbourg, France in 2020, with a thesis entitled "Theological Anthropology and the Possibility of Human-Level Artificial Intelligence: Rethinking Human Distinctiveness and the Imago Dei."

Rituals and Data Analytics: A Mixed-Methods Model to Process Personal Beliefs

Daniel Burgos*

Research Institute for Innovation & Technology in Education (UNIR iTED). Universidad Internacional de La Rioja (UNIR). 26006, Logroño, La Rioja (Spain)

Received 28 April 2021 | Accepted 20 June 2021 | Published 19 July 2021

Abstract

The goal of this research is to delve into ritual, religious, and secular phenomenology. It concentrates specifically on the relationship between pagan, cultural, celebratory, and traditional rituals and any other form of representation of a social sentiment focused on identifying, enjoying, or replacing a feeling (e.g. transcendence) as well as how these rituals overlap, replace, nourish, or make use of religious rituals bi-directionally. To achieve this goal, the research develops a semi-automatic process that leans on a mixed-methods approach, to explore the degree of ritual identity. This approach combines qualitative and quantitative research, applying a number of tools, such as systematic literature review, semi-structured interviews, data-analytics generic framework, and case studies. After a thorough systematic review of 251 publications, a semi-structured interview is designed and applied to 51 subjects. 10 significant aspects that define rituals are extracted. Subsequently, this list is completed with the 17 common elements of ritual identity from the systematic literature review. These combined indicators constitute the basis for building a data-analytics generic framework of ritual affinity through weighing each element's relevance and presence to obtain a degree of total affinity. That framework is then applied to 34 representative case studies. The core findings reinforce the initial hypothesis, determining that rituals follow a similar pattern of structure and preparation according to a predetermined set of common elements, whether linked to religious or secular settings.

Keywords

Data Analytics, Mixed-methods, Pragmatism, Semi-automatic Process, Ritual.

DOI: 10.9781/ijimai.2021.07.002

I. Research Foundation, Methodology, and Methods

THE hypothesis of this research states that there is a similarity between religious and non-religious rituals regarding form, content, meaning, and, above all, structure. The study will seek to prove that rituals fulfil a given function, in a concrete context, with specific actors, and with a defined structure and that this entire frame or data-analytics generic framework is largely applicable to each ritual independently of creed, orientation, or social setting, whether religious, political, sports, family, or any other type' [1]-[2]. Further, the main research question is to determine if rituals follow a similar pattern of structure and preparation according to a predetermined set of common elements, whether linked to religious or secular settings; and if this pattern can be parameterised in a semi-automatic process through a data-analytics generic model.

The researcher used a hybrid methodology, which combines mixed methods under the pragmatic approach. Many disciplines prefer hybrid scientific research based on mixed methods [3]-[6]. Pragmatism combines knowledge-processing methods based on the research needs, the resource provision, and the researcher's view [7]-[8]. Pragmatism is widely used in social science research [9]-[10] and combines qualitative and quantitative methodologies [11]. The objectivist paradigm gives a thorough analysis from data series and user-tracking services, but it lacks the personal context to learn the unique reasons behind a decision or behaviour. In contrast, the constructivist paradigm offers a comprehensive picture of the subject's environment, but it lacks the large objective datasets to escalate and find user patterns. Pragmatism, however, uses both approaches, so the objective data complements subjective interpretation within context. Further, the research followed the pragmatic paradigm and a mixed-methods approach, combining action research with qualitative, experimental, and practical approaches [12]-[15].

The approach consists of four phases in a semi-automatic data analysis process [16]-[17]:

1. The observation phase will use anthropological patterns to collect data and identify patterns without intervention in the sample subjects or in the environment.
2. The interpretive phase will analyse the objective data with the aim of defining the behaviours, phases, and common elements found.
3. From that moment on, the methodology will be focused on a productive phase of design and production of a data-analytics generic framework that allows to group the patterns found.
4. Finally, the methodology will address a semi-experimental phase that will apply the design carried out in case studies, with the aim of validating the instrument and drawing conclusions from the application.

* Corresponding author.
E-mail address: daniel.burgos@unir.net

For the interviews, the researcher used a qualitative approach for data categorisation and analysis since all the evidence was collected from personal, one-to-one, open discussions with experts. For the literature review, the researcher designed a systematic review strategy to look for the best-suited research papers, book chapters, books, and theses so that the selection fits the purpose of the study. For case studies, the researcher used a combined approach, qualitative for open questions to participants and quantitative in the way of a data-analytics generic framework with closed questions to the very same participants. In doing so, the research did benefit from a hybrid approach to the data analysis and interpretation of the findings.

II. Systematic Literature Review

A systematic review is defined as a process for identifying literature according to search and inclusion criteria in a specific field of study [18]. Such a review is scientific research per se with the clear objective of determining the state of the art for a field of study as a result of analysing work done by third parties [19]-[20].

A. Databases Used

The reference search has been executed on the Web of Science (WoS) platform, maintained by the multinational Thomson Reuters and comprising 12,000 important magazines, including Open Access, and over 160,000 conference minutes. WoS comprises over 800 million references and constitutes the largest database of academic and scientific articles focused on the social sciences, science, and humanities. IBCSR Research Review, APA PsycINFO, APA PsycBOOKS, APA PsycARTICLES, the Wiley-Online Library, Sociological Abstracts, Dialnet, Academic Search Premiere EBSCO, SAGE Premier, Scopus, and Taylor & Francis were also used. They were accessed from the Bodleian Library at Oxford University, Westminster University in London, the Gregorian University in Rome, the University of Barcelona, and the Universidad Internacional de La Rioja (UNIR) in Logroño. Google Scholar was also used principally for published books and un-indexed complementary studies (generally not included in the aforementioned databases).

B. Steps Carried Out in the Systematic Review

The systematic review followed the flow and steps presented in Fig. 1.

After discarding 69 references (8 doctoral theses, 45 articles, and 16 books), the final selection for the study constituted 9 doctoral theses, 196 articles, and 46 books, totalling 251 references deserving a calibrated in-depth study within the research process (Fig. 2).

Selection Stages	References identifies in electronic databases according to search terms	Excluded. Non-relevant studies by title and summary	Relevant references by Title and summary	Excluded. Body not related to research	Relevant references by body	Excluded. Non-relevant contribution to research for failing to meet inclusion criteria	References that meet all the selection criteria and whose in-depth analysis shows relevance for the research
Thesis	36	12	24	7	17	8	9
Papers	442	114	328	87	241	45	196
Books	107	21	86	24	62	15	46
N	**585**	**147**	**438**	**118**	**320**	**69**	**251**

Fig. 2. References selected by publication type at each stage.

C. Findings from the Systematic Review

The main aspects that define rituals extracted from the systematic literature review are 17 common elements of ritual identity: (1) transcendence, (2) feeling, (3) meaning, (4) transformation, (5) contextualisation, (6) polysemy, (7) music, (8) need, (9) representation of reality, (10) ceremony, (11) stages, (12) formality, (13) script, (14) impact, (15) invariability, (16) periodicity, and (17) symbolism [21]-[30]. Following and based on these findings, an interview protocol and questionnaire were designed.

III. Interview Design

A. Description of the Target Group and the Sample

The study targeted university graduates with work experience and sensitivity towards the subject matter but without a background in philosophy, anthropology, or theology or specific knowledge of the terminology. The intention was that they should answer freely and be interested in staying updated on the process and the results. The interviews were conducted mainly in Spanish, English, and French. Furthermore, a script has been prepared in Italian and Portuguese; these scripts can be used as a base to further explain some concepts but not for primary-language interviews.

Within these parameters, the researcher required responsible, cultured adults, sensitive to the subject matter, who could provide first-hand information with personal interpretations of religious, sports, political, musical, family, intimate, or any other type of ritual previously described. The intention was to gather information from people who had not been influenced by prior studies or texts

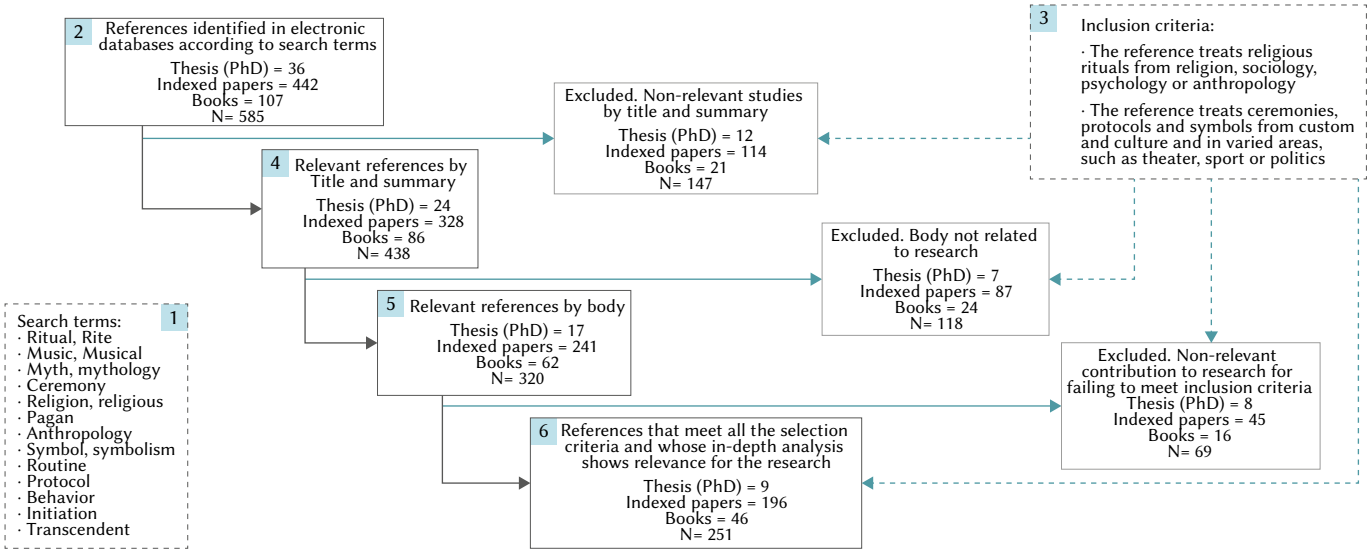


Fig. 1. Steps in the reference selection process and results by type of publication.

developed on this matter so that they could express themselves while free of either active or passive guidance. In a sense, the researcher sought a popular and educated voice given the mundane and popular aspect inherent to rituals.

The sample consists of 51 interviewees, of which 20 are women and 31 are men; 23 reside in Spain, and 28 are distributed among Saudi Arabia, Argentina, Australia, Belgium, Brazil, Chile, Colombia, Ecuador, England, Morocco, Mexico, New Zealand, the United Kingdom, the Dominican Republic, and Venezuela. Ages range from 24 to 73 years. All are graduates. Furthermore, 21 are practicing Catholics, 2 are Muslims, 2 are Protestants, 21 are not worshippers or are atheists, and 5 declined to answer. Lastly, 30 Catholic, 2 Protestant, 2 Muslim, 1 Jewish, 2 popular belief, 3 therapeutic, 4 intimate, 9 spiritual, 8 sporting, 2 artistic, and 6 social rituals have been described (occasional double classification).

The interview phase was conducted from September 2017 to July 2018 and took place in the cities of Sao Paulo (Brazil), Marrakech (Morocco), Brussels (Belgium), Quito (Ecuador), Paris (France), Toronto (Canada), Buenos Aires (Argentina), Bogotá (Colombia), Jeddah (Saudi Arabia), and Santa Cruz de Tenerife, Salamanca, Madrid, Murcia, León, and Logroño (Spain).

B. Design and Description of the Base Script

The interview script has been structured in eight sections or categories. Each section is composed of several questions, with at least one considered as primary (that is, requiring an answer). In this manner, the core of the primary questions constitutes the basic outline with which the interview must be conducted; the remaining questions are thus complementary. There is a total of 9 categories, 74 possible questions, and 18 primary questions (including the informed consent), together with 3 fixed pieces of information (unique identifier, date, and time).

The questions can be open (free development) or closed (restricted answers). There are 4 mixed (predetermined answers with the option of additional freely developed comments) and 3 fixed data fields. The final script template has been centralised in an online database to be able to feed and analyse it instantly and coherently from different geographic locations. Centralised dumping has proven to be a very valuable tool for the in-depth comparison and coherent cross-analysis of the interviews. Given the interest generated by the research, the centralised database includes an online form to be used in a subsequent research phase.

The categories taken into account are as follows: sample, scope, characterisation, meaning, behaviour, transference, evolution, and impact. The questions and type of associated question in each category are attached as Appendix I.

The semi-structured interview included 62 questions, 18 of which were primary or mandatory, and was divided into 8 sections and conducted with 51 individuals. The questions were categorised as sample, scope, characterisation, meaning, behaviour, transference, evolution, or impact questions.

The interview subjects were university graduates with work experience in and sensitivity towards the subject matter. The researcher preferred to interview people who had not studied subjects related to rituals, such as philosophy, anthropology, or theology. The interviews were conducted mainly in Spanish, English, and French, and the script was also prepared in Italian and Portuguese to be used occasionally as support. The researcher thus hoped to find responsible, cultured interview subjects who were sensitive to the research subject and who could voice their opinions and interpret them without being influenced by formal education. In short, the researcher sought the voice of the informed person to learn the mundane and popular aspects inherent in rituals.

C. Findings from the Interview Phase

The analysis of the 51 interviews evidenced a recurring number of 10 primary aspects:

1. The participants exhibited a high degree of affinity towards the ritual as well as a transcendence of the event beyond the daily custom.
2. The ritual overlapped with the introspection ceremony – a large number of interviewees had periodically explored an act of synchronisation between the individual and their surroundings.
3. The participants often confused terminology, some of which was basic and general knowledge, and had overlapping beliefs and definitions. The researcher perceived that the starting concepts were fuelled by creative definitions.
4. The participants were unfamiliar with the depth of religious, choral, and/or social rituals. The roles, characteristics, or symbols, for example, were familiar yet not known in detail.
5. The participants exhibited a need to express and move beyond the mundane to experience a transcendent feeling, whether personal or collective.
6. The sample was clearly and universally cross-sectional or transversal regardless of culture, religion, or type of ritual, with a marked range of ages, languages, and backgrounds. Nevertheless, the researcher found similar or identical answers and attitudes across the entire group and often had the impression of hearing the same person talk through different mouths.
7. Each participant interpreted a ritual differently, granting it singularity and a specific meaning (i.e. polysemy). Thus, while maintaining structure, steps, or symbols, each individual or collective ritual acquired personality through the eyes and actions of each participant.
8. The widespread use of vocal or instrumental music was a recurrent pattern. The ritual was the same with or without music, but the music – whether it was chanting, a song, background, a cappella, with instruments, or with harmony – represented a catalyst for the individual's or the group's participation.
9. The ritual was essential and indispensable for all the interview subjects. Human beings need to express themselves and articulate expression through a ritual.
10. The participants experienced evocation through rituals, awakening memories, feelings, and moods.

These findings will help in modelling the data-analytics generic framework, as shown in the following sections.

IV. Data-Analytics Generic Framework for Rituals

This section establishes a data-analytics generic framework for standardising the structure, processes, and behaviours of rituals. The final intent is to design the most abstract model possible that is able, most precisely, to identify a wide range of rituals according to category, orientation, intent, population, and other criteria.

The data-analytics generic framework seeks to set parameters for measuring the ritual identity of an event or celebration. This framework, therefore, will analyse this hypothesis based on identity, composition, context, and other factors of any type of celebration, whether religious or secular. To that end, the framework is constructed by parameterising and weighting the common elements that allow for a generic approach and, therefore, are normally applicable to any event or celebration.

A. Identifying the Elements that Characterise a Ritual

From all the material studied for this section, the researcher highlights [22], [25], [31]. They all note the systemisation of the components and sections of a ritual from different points of view.

Four steps were taken to identify the common characteristics; the first step was to select up to 10 key aspects identified during the interviews (Table I).

TABLE I. Key Aspects from Interviews

Aspect #	Description of key aspect
1	Transcendence and affinity with the ritual moment
2	Overlapping among ritual, introspection, expression, and feeling
3	Unclear definition of terms
4	Unfamiliarity with rituals
5	Need for expression
6	Transverse approach
7	Polysemy
8	Music
9	Indispensability
10	Evocation

In the second step, the researcher selected and combined common elements found in the literature review. Each element listed below includes a brief description and at least one reference of renowned authors who used them in their work (Table II):

TABLE II. Common Elements from the Literature Analysis

Element	Explanation	Supporting reference 1
Ceremony	Expressed through representation or an act	Bell, 1997 [2]; Van Gennep, 2011 [29]
Contextualisation	Makes sense in a specific context	McCauley & Lawson, 1999 [26]
Stages	Stages are clearly defined	Kelly & Kaplan, 1990 [31]
Formality	Performed in a formal setting	Bell, 1997 [2]
Script	Designed according to a script with roles and sections	McCauley & Lawson, 1999, 2002 [32]; Staal, 1979 [33]
Impact	Produces an impact on the subject	Eliade, 1957 [24]; Turner, 1967 [28]
Invariability	Repeated over time without major alterations	Staal, 1979 [33] ; Bell, 1997 [2]
Music	Music used in a significant way	Grimes, 2014 [25]
Need	The ritual is required to guarantee stability, commitment, meaning, or any other aspect	Fredericks, 2021 [34]; Freston, 2001 [35]
Regularity	Repeated periodically	Bell, 1997 [2]
Polysemy	Interpreted differently depending on who and how	Turner, 1967 [28]; Eliade, 1957 [24]
Representation of reality	Performed to evoke or represent reality	Eliade, 1957 [24]
Feeling	Arouses genuine feeling	Driver, 2006 [23]
Meaning	Involves expressing meaning	Leach, 1966 [36]; Durkheim & Swain, 2008 [37]
Symbolism	Uses symbols linked to feelings or meanings	Bell, 1997 [2]
Transcendence	Elevates the spirit	Grimes, 2014 [25]; Strenski, 1991 [38]
Transformation	Causes a change of state or status	Driver, 2006 [23]; Grimes, 2014 [25]

The third step matched the aspects identified during fieldwork (#1–10) with elements extracted from the analysis of the systematic literature review to find comparable meanings (Table III):

TABLE III. Pairing of Aspects from Interviews and Elements in the Literature

Elements	Explanation	Key aspect	Description of key aspect
Ceremony	Expressed through representation or an act		
Contextualisation	Makes sense in a specific context	#6	Transverse approach
Stages	Stages are clearly defined		
Formality	Performed in a formal setting		
Script	Designed according to a script with roles and sections		
Impact	Produces an impact on the subject		
Invariability	Repeated over time without major alterations		
Music	Music used in a significant way	#8	Music
Need	The ritual is required to guarantee stability, commitment, meaning, or any other aspect	#9	Indispensability
Regularity	Repeated periodically		
Polysemy	Interpreted differently depending on who and how	#7	Polysemy
Representation of reality	Performed to evoke or represent reality	#10	Evocation
Feeling	Arouses genuine feeling	#2	Overlapping among ritual, introspection, expression, and feeling
Meaning	Involves expressing meaning	#5	Need for expression, for experiencing something beyond the day-to-day
Symbolism	Uses symbols linked to feelings or meanings		
Transcendence	Elevates the spirit	#1	Transcendence and affinity with the ritual moment
Transformation	Causes a change of state or status	#5	Need for expression, for experiencing something beyond the day-to-day

During this pairing process, the researcher found that two key aspects are not reflected in the systematic review (Table IV):

TABLE IV. Discarded Aspects from Interviews

Aspect #	Description of key aspect
3	Unclear definition of terms and concepts
4	In-depth unfamiliarity with religious rituals

The researcher also identified Aspect #5 (need for expression) with two elements from the literature: meaning and transformation, all focused on causing a change in state or status and on the need for expression, for experiencing something beyond the mundane. Thus, the researcher identified 8 out of the 10 aspects from the fieldwork analysis as key elements in the definition and identification of a ritual according to the group of authors and work analysed in the systematic literature review. The researcher found that Aspects #3 and #4 correspond with the interpretation of the sample and with the understanding of the process and ritual identity. They cannot, therefore, be considered as definitions of the ritual itself but of a characteristic of the population sample. Though interesting as a feature of context and sample, they are not relevant to the definition of a generic framework that identifies key aspects of the research subject. Therefore, the researcher did not include them in the final list of the framework elements.

In the fourth and last step, the researcher organised the list of elements, prioritising the ones found during fieldwork. It should be noted that this prioritisation merely follows grouping and aesthetic criteria and does not include any special assignment or identification within the framework (Table V).

This table represents the list of elements that characterise a ritual, resulting from combining the findings of the analysis of the systematic review of literature and those of the fieldwork, which was carried out as semi-structured qualitative interviews.

TABLE V. Matching and Grouping of the Elements

#	Elements	Explanation	Key aspect	Description of key aspect
1	Transcendence	Elevates the spirit	#1	Transcendence and affinity with the ritual moment
2	Feeling	Arouses genuine feeling	#2	Overlapping among ritual, introspection, expression, and feeling
3	Meaning	Involves expressing meaning	#5	Need for expression, for experiencing something beyond the mundane
4	Transformation	Causes a change of state or status	#5	Need for expression, for experiencing something beyond the mundane
5	Contextualisation	Makes sense in a specific context	#6	Transverse approach
6	Polysemy	Interpreted differently depending on who and how	#7	Polysemy
7	Music	Music used in a significant way	#8	Music
8	Need	The ritual is required to guarantee stability, commitment, meaning, or any other aspect	#9	Indispensability
9	Representation of reality	Performed to evoke or represent reality	#10	Evocation
10	Ceremony	Expressed through representation or an act		
11	Stages	Stages are clearly defined		
12	Formality	Performed in a formal setting		
13	Script	Designed according to a script with roles and sections		
14	Impact	Produces an impact on the subject		
15	Invariability	Repeated over time without major alterations		
16	Regularity	Repeated periodically		
17	Symbolism	Uses symbols linked to feelings or meanings		

B. Informing the Data-analytics Generic Framework

Having selected and prioritised the elements that make up the data-analytics generic framework, the researcher established a weighting system that permits the calibration of each element's importance according to the individual and the event or celebration [39]. Two values should be considered: (1) the value that the individual assigns to an element based on how much they believe it to be present in the ritual being evaluated; and (2) the absolute weight of importance that the element has in the ritual being evaluated. In other words, what presence does a specific element have, and what is the relevance of that presence within the entirety of the ritual's definition?

To standardise the weights, the researcher chose the Likert scale [40]-[41], which measures from 'highly against' to 'highly in favour', on a scale from 1 to 5: (1) 'strongly disagree', (2) 'disagree', (3) 'neutral', (4) 'agree', and (5) 'strongly agree'. The researcher chose to include a sixth degree, '(0) completely disagree', which takes into consideration frontal opposition since (1) 'strongly disagree' may still indicate a positive score, even if there is none. This way, and by adapting the scale to the framework, each element is valued by a set of degrees ranging from 0 to 5 (from 'non-existent' to 'maximum') – on one hand, the relevance of the element according to the subject's perspective and, on the other hand, the level of its presence in the ritual being evaluated.

It is necessary to highlight the importance of this double weighting, which is designed to offer a unique calibrating system that works irrespective of whether the evaluator is an individual actively participating in the ritual or an external observer (e.g. as researcher). In the first case, the context of a ritual and the individual's interpretation regarding the relevance of its different elements are still significant in the experience, impact, and transcendence of the ritual [42]-[44]. In this sense, the individual is not merely a consumer or a spectator but rather a vehicle of celebration of the ritual experience. The individual becomes a conduit of the expression, interpretation, and meaning of the ritual. In summary, they are a part of the scene's assessment and are jointly or uniquely responsible for the final representation, thus combining both active roles. Therefore, the individual must evaluate both aspects (degree of relevance and degree of presence) to provide a unique and personal assessment of the identity of the event or celebration with the archetypical ritual defined by the elements of the generic framework.

In the second case, where the evaluator of the ritual is external (e.g. a researcher) and not an active participant, the double calibration system allows them to measure the relative weight of each element according to the context, including the subject of the assessment [45]-[46]. In this way, the experience is customised through a personal application of the general parameters according to each evaluator's gauge. In both cases, active participant and external observer, the system offers greater accuracy in terms of matching the event or celebration with the general ritual framework given the common elements that define them.

Thus, affinity is ruled by two combined values on a Likert scale of 0 ('non-existent') to 5 ('maximum'), which was multiplied to obtain one value. Given that multiplication ranges between 0 ('non-existent') and 25 ('maximum'), for easier understanding, it was changed into a percentage scale from 0% (= 0, non-existent) to 100% (= 25, maximum). Each element provides a percentage value. The total affinity of the event or celebration with the general ritual framework is the proportion of the percentage values of the 17 elements (n = 17) over the maximum possible value. The following equation shows the total affinity as a percentage value, resulting from calculating over the set of elements assessed in relevance and presence:

$$Affinity(n) = \frac{\sum_{i=1}^{n}(Relevance(i) * Presence(i))}{\sum_{i=1}^{n}(Relevance(i) * Max(0:5))}$$

For ease of interpretation, (a) decimals were limited, and (b) colours to sections were assigned, as follows (Table VI):

TABLE VI. Affinity by Section and Range

Section	Affinity range (absolute value: ABS)	Affinity range (relative value: %)	Interpretation
1	0.00–6.00	0%–24%	Minimum
2	6.25–12.25	25%–49%	Low
3	12.50–18.50	50%–74%	High
4	18.75–24.00	75%–100%	Maximum

The final dashboard of the data-analytics generic framework is shown in Fig. 3.

C. Selection of Case Studies

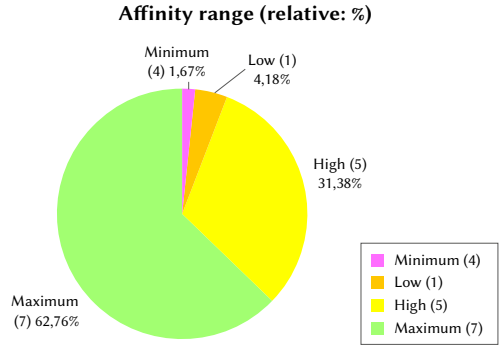
Rituals already assimilated by the various communities of practice do not represent a validation challenge. A Catholic sacrament, the daily prayer cycle of Islam, the chanting of Buddhist mantras, or the offering of Taoist incense are rituals established in their creeds, incorporated by their communities, and used by their practitioners. It is not the subject of this thesis to question established rituals but, as indicated by Research Question 2, to delve into the similarities between religious and secular rituals: 'Do religious and secular rituals maintain an equal or similar structure at different times or contexts of application?' In other words, are secular rituals comparable to religious ones in their identity as rituals in that they both maintain a comparable structure and definition? To this end, a selection of secular and religious rituals was made as case studies to which researcher applied the generic framework. These can be personal or group rituals, more or less community based, and more or less mundane or general. They meet the condition that they have been published at some point and are therefore open to reference; for each case, a list of references that define it in detail is attached.

The goal of applying the framework is not to catalogue or describe each case in detail but to draw conclusions from the application of the tool to the potential ritual to analyse the degree of affinity it shows

Caso-ID	Title	Initiation into a Salvadoran gang ("mara")	
04	References	Hume, 2007; Miguel Cruz, 2010	
	Stages	Candidacy, proof, personal detachment, group assumption, tatoo membership, submission to leader	**Description**
	Roles (who)	Initiator, applicant, group	A candidate (young, usually) wants to join a criminal group, closed and structured. To do this, you must go through an identity and loyalty process with that group and with your boss. Only if you pass the input tests and the steps of the final event will you be admitted as a brotherhood or sibling within the group
	Justification (why)	The feeling of belonging to a group and submission to a leader leads to joining in an act that requires showing value and loss of self-identity to assume the group	
	Moment (when)	Just one time	
	Meaning (what)	Assumption of group identity and submission to a leader	
	Type	Initiation	
	Context	Group	

Element ID	Element of the ritual	Relevance	Presence	Affinity (ABS)	Affinity (%)
				239/330	72%
1	Transcendence	5	0	0	0%
2	Feeling	5	3	15	60%
3	Meaning	5	5	25	100%
4	Transformation	5	4	20	80%
5	Contextualization	4	5	20	80%
6	Polysemy	1	0	0	0%
7	Music	5	4	20	80%
8	Need	5	3	15	60%
9	Representation of reality	5	0	0	0%
10	Ceremony	4	5	20	80%
11	Stages	3	5	15	60%
12	Formality	3	5	15	60%
13	Script	5	5	25	100%
14	Impact	2	5	10	40%
15	Invariability	3	5	15	60%
16	Regularity	2	2	4	16%
17	Symbolism	4	5	20	80%

Nr. of elements per range	Affinity range (absolute value)	Affinity range (relative: %)
17	239	100%
Minimum (4)	4,00	1,67%
Low (1)	10,00	4,18%
High (5)	75,00	31,38%
Maximum (7)	150,00	62,76%



*Minimum: 0%-24% | Low: 25%-49% | High: 50%-74% | Maximum: 75%-100%

Fig. 3. Case ID-04: Initiation into a gang ('mara').

with the definition established by each element used as framework parameters. Therefore, the indicated references reflect, support, and extend the summary identification of each case, which should be understood as a combination of those basic references. In other words, the set of references addresses the basic identification data, which do not respond exclusively to a reference but to the joint definition extracted from all of them.

Also, within the variety of these rituals, the sources and cultures have been diversified to obtain a varied sample that allows a contrast of this thesis's hypothesis with the use of the tool generated from the fieldwork and the systematic literature review. To do this, Grimes's indications [25] about the variety of contexts of the rituals as a basis were followed – group events, intimate/personal, social events, religious, artistic-cultural, sports, political events, etc. For this reason, the cases are grouped by the following contexts, some of which may be mixed or combined – cultural, group, social, political, sports, personal, medical, curative, and religious. A final detailed list of 34 selected cases is presented in Appendix II.

V. Discussion

The generic framework was applied to those 34 representative case studies which were obtained by analysing specialised literature that specifically describe the cases in detail; this literature review was not systematic but rather served to broaden the bibliography. A representative sample was chosen which combines the often-linked religious and secular contexts: cultural, group, social, political, sports, personal, medical, curative, and religious. The study chain began with the systematic review, followed by analysing the literature, conducting semi-structured interviews, and designing the evaluation tool; as the final step, the researcher applied the framework to the case studies and obtained the following 14 findings:

1. Rituals exhibit common elements that imply a generic framework of identity. Each ritual adapts based on its type, category, context, and other conditions, but all rituals can be defined by one unique set of parameters.
2. Religious rituals and secular rituals coincide within this defining framework. There are no significant differences regarding identification, design, and implementation; the difference is provided by the context and the meaning given to them by the participants.
3. Ritual identity is intimately linked to the profile of each event, the context in which it is performed, and the participants. Even so, each identity is comparable in both religious and secular rituals, with no significant difference between the two settings.
4. There are ritual elements that themselves define their identity, without a need for a majority of aligned elements. In other words, a ritual can be defined with a few distinct elements, if and when their relative value is significant enough.
5. Conversely, some rituals do not achieve maximum affinity in any element and yet are considered rituals given the average score held by the majority of their elements. This shows that regularity in the elements' affinity produces a result that is just as valid as rituals with only some clearly marked elements.
6. Within the context of the case study sampling where the generic framework was applied, the researcher observed a classification of elements by average affinity. The standard deviation is irrelevant in this set. Three groups were distinguished, from greater to lesser affinity. The first group is made up of meaning and contextualisation, both primaries with an affinity over 75%. The second group holds eight elements (transformation, feeling, music, ceremony, symbolism, need, script, and transcendence) with high affinity. The researcher found the transcendence element, which is traditionally a determinant factor for considering a ritual as such; however, the researcher also found that other elements exhibited an equivalent or distinctly higher affinity, though they are all closely related to the effect that the ritual has on the participant and that the participant affords the ritual. The third and last group contains seven circumstantial and operational elements, with the exception of the impact element, which has a low value in this group but belongs to the second group, by definition. The polysemy element seems residual.
7. The comparison between religious and secular contexts, which coincides with the analysis of the entire set of cases, shows a group of seven elements (meaning, feeling, transformation, need, symbolism, contextualisation, and music) with maximum and high affinities in both contexts. These elements create synergy with the ritual, whether religious or secular, to a greater or lesser degree and homogeneously. Two other elements are placed in the first group, although only in one context each: script in secular and transcendence in religious. All these elements refer to the effect that the ritual has on the participant and vice versa, opposite another large majority of elements that could be considered more circumstantial or operational (e.g. stages, regularity, or ceremony).
8. Transcendence – traditionally assigned as a defining, if not unique, characteristic of a ritual – emerges with high but not maximum significance. It is only in the context of other elements (mentioned in the previous conclusion) that transcendence claims greater prominence but never in isolation.
9. Each element's rating in the two contexts is matched, increasing and decreasing harmoniously, almost in its entirety. Although the religious context provides more defined values, the more established, multitudinous, and vastly recognised profile of the selected religious rituals usually implies greater refinement of the terms and must be considered. The doubles between contexts for each element are balanced and matched in up to eight elements, with a slight percentage difference of about 10 points.
10. A ritual rarely shows affinity in each element. Given the diversity of rituals and their adaptations to local context, participants, culture, and other factors, it is almost impossible that a ritual would need all the elements to be considered as such.
11. Regardless of context, no ritual showed 100% affinity with all the elements.
12. Ritual affinity depends greatly on the participant and the context. In those cases where affinity borders the threshold between considering a specific event or celebration as a ritual or not (not necessarily established in 50%), that participant's nature in that context is what significantly defines ritual identity.
13. Rituals can be parameterised and studied following a semi-automatic process that relies on mixed methods, involving qualitative analysis and data analytics.
14. The calibration of the data-analytics generic framework defines identity by assigning the relevance factor. According to how the outside observer gauges that factor when applying the framework, the affinity indexes will vary correspondingly. The presence factor also depends on the outside observer's subjective interpretation. Each element's affinity and total affinity are, therefore, relative to the evaluating individual.

VI. Conclusion

The main hypothesis proposes a similarity between religious and secular rituals and concentrates on comparing the ritual function and the structure of rituals regardless of context. Similarly, the researcher

designed a data-analytics generic framework that evaluates the degree of ritual affinity of an event or celebration, also independently of that context. The results show that, in effect, religious and secular rituals are defined by a series of common elements that characterise their identity and function and that can be processed through semi-automatic data-analytics techniques. These elements can be standardised through a data-analytics generic framework and weighted according to the factors' relevance and presence, which can be applied by an outside observer or researcher (or by the interested party themselves) in a possible future evolution of the methodology. To do so, the research shows that this semi-automatic process can be developed as a data-analytics generic framework in a mixed-methods approach and which combines qualitative and quantitative techniques.

A next step in further studies would be to integrate a third assessment factor into the data-analytics generic framework: priority. This would involve the observer establishing an order to the list of elements of ritual identity, which would require new weighing. Three factors would then be aligned – relevance, presence, and priority. The additional factor of priority provides a more precise measurement if applied correctly, but it also complicates the understanding of the framework as a tool to be applied by an individual, by an active subject of the ritual being analysed, or by an outside observer. Another future step is to implement the data-analytics generic framework as an online tool for open use by any interested party. This step would also offer training and explain in detail how to use the tool. The observer's (researcher's) autonomy would be sought, and the tool would help to optimise the data-analytics generic framework (and the related semi-automatic analysis process) and to provide additional information developed by other researchers, with the aim of sharing and exploring the findings.

Appendix I

ID-1. Sample: Defines the Interviewee's Basic Information

Unique ID	Category	Code	Question	Response type
ID-1-0	Sample	Primary	Interviewee's unique identifier (anonymized)	Fixed
ID-1-0	Sample	Primary	Date of interview	Fixed
ID-1-0	Sample	Primary	Time of interview	Fixed
ID-1-1	Sample	Primary	Consent	Closed
ID-1-2	Sample		Interviewee name	Open
ID-1-3	Sample		Interviewee surname(s)	Open
ID-1-4	Sample		Interviewee gender	Open
ID-1-5	Sample		Interviewee email address	Open
ID-1-6	Sample		Interviewee language	Open
ID-1-7	Sample		Interviewee city of birth	Open
ID-1-8	Sample		Interviewee country of birth	Open
ID-1-9	Sample		Interviewee city of residence	Open
ID-1-10	Sample		How long has interviewee resided in this city?	Open
ID-1-11	Sample		Interviewee country of residence	Open
ID-1-12	Sample		How long has interviewee resided in this country?	Open
ID-1-13	Sample		Interviewee's age	Open
ID-1-14	Sample		Interviewee's profession	Open
ID-1-15	Sample	Primary	Interviewee's relation to the study	Open
ID-1-16	Sample		City where interview takes place	Open
ID-1-17	Sample		Country where interview takes place	Open
ID-1-18	Sample		Religion interviewee professes and practices	Mixed

ID-2. Scope: Defines the Field of the Ritual

Unique ID	Category	Code	Question	Response type
ID-2-1	Scope		To which religion does the ritual pertain?	Mixed
ID-2-2	Scope		If not religious, to which other aspect or area does the ritual pertain?	Mixed
ID-2-3	Scope		If religious, what is the ritual's relevance within the religion?	Open
ID-2-4	Scope		What type of ritual is it, which type of moment does it express?	Mixed

ID-3. Characterisation: Identifies the Ritual

Unique ID	Category	Code	Question	Response type
ID-3-1	Characterization		How many people normally participate in the ritual?	Closed
ID-3-2	Characterization		How long has the ritual been celebrated?	Closed
ID-3-3	Characterization		How frequently is the ritual celebrated?	Closed
ID-3-4	Characterization	Primary	What is the protocol of the ritual?	Open
ID-3-5	Characterization	Primary	What are the steps or stages of the ritual?	Open
ID-3-6	Characterization		Is dance used at any time?	Open
ID-3-7	Characterization		What official or unofficial training is required of the officiant/minister?	Open
ID-3-8	Characterization		What prior experience is required of the officiant/minister?	Open
ID-3-9	Characterization		What prior experience is required of the participant?	Open
ID-3-10	Characterization		What preparations does the ritual require of the officiant/minister?	Open
ID-3-11	Characterization		What preparations does the ritual require of the participant?	Open

ID-4. Meaning: Explains the Meaning of the Ritual according to the Interviewee

Unique ID	Category	Code	Question	Response type
ID-4-1	Meaning	Primary	What symbology is used?	Open
ID-4-2	Meaning	Primary	Which tools or objects are used and how?	Open
ID-4-3	Meaning	Primary	What is the culminating point of the ritual and what does it mean?	Open
ID-4-4	Meaning	Primary	What does each step or stage mean?	Open
ID-4-5	Meaning		What significance is portrayed by the dances used?	Open
ID-4-6	Meaning		What is the significance of the texts used?	Open

ID-5. Behaviour: Describes the Behaviour of the Ritual Participants

Unique ID	Category	Code	Question	Response type
ID-5-1	Behaviour	Primary	What behaviour do the participants exhibit?	Open
ID-5-2	Behaviour	Primary	What interaction is there between participants in the ritual?	Open
ID-5-3	Behaviour		What is the behaviour of the officiant/minister during the ritual?	Open
ID-5-4	Behaviour		What is the behaviour of the officiant/minister before and after the ritual?	Open

ID-6. Transference: Indicates the Ritual's Relationship to External Elements

Unique ID	Category	Code	Question	Response type
ID-6-1	Transference		Is there an identification of the ritual with any specific person?	Open
ID-6-2	Transference	Primary	Is there an identification of the ritual with any other external aspects?	Open
ID-6-3	Transference	Primary	What influence or overlapping with other rituals does this ritual exhibit?	Open
ID-6-4	Transference	Primary	What influence or overlapping with other rituals does this ritual have?	Open
ID-6-5	Transference		What influence does this ritual have from abroad or from other cultures?	Open
ID-6-6	Transference		What influence does this ritual have abroad or in other cultures?	Open
ID-6-7	Transference	Primary	Is there any transference from Society to the ritual?	Open
ID-6-8	Transference	Primary	Is there any transference from the ritual to Society?	Open

ID-7. Evolution: Explains the Ritual's Historic Evolution

Unique ID	Category	Code	Question	Response type
ID-7-1	Evolution		How has the officiant's/minister's profile evolved throughout the ritual's history?	Open
ID-7-2	Evolution		How has the participant's profile evolved throughout the ritual's history?	Open
ID-7-3	Evolution		What modifications do you expect the ritual to undergo in the near future?	Open
ID-7-4	Evolution		What modifications has the ritual undergone since its creation?	Open
ID-7-5	Evolution		What variations of the ritual are there and where?	Open

ID-8. Impact: Describes the Ritual's Relationship with Media and Society in General

Unique ID	Category	Code	Question	Response type
ID-8-1	Impact	Primary	How is the ritual attacked, by whom, and why?	Open
ID-8-2	Impact		What type of coverage does it receive in the media?	Open
ID-8-3	Impact		What type of coverage does it receive in social media?	Open
ID-8-4	Impact		Link to further information on the ritual	Open

ID-9. Music: Describes the Role of Music in the Ritual and Its Relationship with the Other Categories

Unique ID	Category	Code	Question	Response type
ID-9-1	Music	Primary	What music is used?	Open
ID-9-2	Music		What musical style(s) accompany the ritual?	Open
ID-9-3	Music		How frequently is music used?	Open
ID-9-4	Music		Which specific moments use music, and what type of concrete music?	Open
ID-9-5	Music	Primary	Which people are involved musically: participants, minister, separate group, etc.	Open
ID-9-6	Music		What musical instruments are used?	Open
ID-9-7	Music		If there are lyrics, in what language are they used?	Open
ID-9-9	Music		If there are lyrics, who sings them and when?	Open
ID-9-10	Music		If there is dancing associated with the music, what type of dancing is it?	Open
ID-9-11	Music	Primary	What does the music mean when used within the context?	Open
ID-9-12	Music		What symbology is directly associated with the ritual's music?	Open

Appendix II

Selected Case Studies

Case ID	Context	Ritual contrasted with generic framework
1	Cultural	Artist's preparation before a performance
2	Cultural	Musical performance
3	Cultural, Religious	Dressing a bullfighter with a 'traje de luces'
4	Group	Initiation into a gang ('mara')
5	Group	Group breakfast/lunch/dinner
6	Group	Tea ceremony
7	Group	Non-religious initiation (i.e. into a sect, army, society, fraternity, etc.)
8	Group	Use of recreational drugs (wine, alcohol, coffee, tobacco, marijuana) in a group
9	Group	Modern hunting in a group
10	Social, Political	Ritual of taking office as president of a nation
11	Social, Political	Participating in democratic elections of India in Europe
12	Social, Cultural	All Souls' Day, Day of the Dead
13	Social	Crop harvest, grape harvest, olive harvest
14	Social	Releasing sky (or Chinese) lanterns
15	Social	New Year's Eve and New Year (countdown)
16	Social	Female puberty ritual
17	Social, Religious	Carnival: Burial of the Sardine
18	Social, Religious	Welcoming the arrival of spring, changing of the seasons
19	Sports	Running (jogging)
20	Sports	Elite athlete's preparation prior to competition
21	Sports	Practicing yoga
22	Sports, Group	Attendance at mass sports events (e.g. football)
23	Personal	Preparing and enjoying coffee
24	Medical, Curative	Placebo and assistance rites in medicine
25	Medical, Curative	Ritual healing
26	Religious, Curative	Shamanic rites
27	Religious	Mha Pujā in Nepal
28	Religious	Catholic confirmation
29	Religious	Religious initiation
30	Religious	Religious offering (floral, ornamental, etc.)
31	Religious	Jewish Passover
32	Religious	Burning incense, lighting a candle
33	Religious	Breaking the fast at the end of Ramadan (Eid al-Fitr)
34	Religious, Personal	Hindu sun salutation

Acknowledgment

Special thanks go to Prof Lluís Oviedo from the Pontifical Antonianum University (Rome, Italy) as PhD supervisor, who supported the research process all the way, and to Prof Bernardo Pérez as PhD programme director from the Theological Institute of Murcia at Universidad de Murcia (Spain), who encouraged me throughout the process. The author also thanks all the host universities, libraries, and interviewees that sympathised with this research.

References

[1] E. Haynes, R. Garside, J. Green, M.P. Kelly, J. Thomas, C. Guell, "Semiautomated text analytics for qualitative data synthesis", *Research Synthesis Methods*, vol. 10, no. 3, pp. 452–464, 2019.

[2] A. J. Johs, D.E. Agosto, R.O. Weber, "Qualitative investigation in explainable artificial intelligence: A bit more insight from social science". arXiv preprint arXiv:2011.07130, 2020.

[3] G. Abeza, N. O'Reilly, M. Dottori, B. Séguin, O. Nzindukiyimana, "Mixed methods research in sport marketing", *International Journal of Multiple Research Approaches*, vol. 9, no. 1, pp. 40–56, 2015.

[4] J. W. Creswell, V.L.P. Clark, "*Designing and conducting mixed methods research"*, Sage Publications, 2017.

[5] R. B. Johnson, A.J. Onwuegbuzie, "Mixed methods research: A research paradigm whose time has come", *Educational Researcher*, vol. 33, no. 7, pp. 14–26, 2014.

[6] P. Leavy, "*Research design: Quantitative, qualitative, mixed methods, arts-based, and community-based participatory research approaches"*, Guilford Publications, 2017.

[7] R. Wille, "Conceptual landscapes of knowledge: A pragmatic paradigm for knowledge processing", in *Classification in the Information Age*, pp. 344–356, Berlin, Heidelberg: Springer, 1999.

[8] R. Wille, "Methods of conceptual knowledge processing", in *Formal Concept Analysis*, pp. 1–29, Berlin, Heidelberg: Springer, 2006.

[9] K. F. Punch, "*Introduction to social research: Quantitative and qualitative approaches"*, Sage: London, 1998.

[10] A. Tashakkori, C. Teddlie (Eds.), "*Sage handbook of mixed methods in social & behavioral research*", Sage Publications:Thousand Oaks, CA, 2010.

[11] P. Bazeley, "Issues in mixing qualitative and quantitative approaches to research", in *Applying Qualitative Methods to Marketing Management Research*, pp. 141, 156, 2004.

[12] C. R. Kothari, "*Research methodology: Methods and techniques"*, New Age International, 2004.

[13] R. Kumar, R. "*Research methodology: A step-by-step guide for beginners"*, Sage Publications Limited:Thousand Oaks, CA, 2019.

[14] S. Kumar, P. Phrommathed, P. "*Research methodology"*, Springer, pp. 43–50, 2005.

[15] G. Rodríguez Gómez, J. Gil Flores, E. García Jiménez, "*Metodología de la investigación cualitativa"*, Aljibe: Málaga, 1999.

[16] N. Astrakhantsev, D. Turdakov, N. Vassilieva, "Semi-automatic data extraction from tables", in *RCDL*, pp. 14–20, 2013.

[17] A. Jaimes, J.R. Smith, "Semi-automatic, data-driven construction of multimedia ontologies", in *2003 International Conference on Multimedia and Expo ICME '03 Proceedings* (Cat. No. 03TH8698), Vol. 1, pp. I–781, IEEE, 2013.

[18] R. Newell, P. Burnard, "*Research for evidence-based practice"*, Blackwell Publishing, 2006.

[19] G. V. Glass, "Primary, secondary, and meta-analysis of research", *Educational Researcher*, vol. 5, no. 10, pp. 3–8, 1976.

[20] M. Hunt, M. "*How science takes stock: The story of meta-analysis"*, Russell Sage Foundation, 1997.

[21] J. L. Barrett, J. L. "*Cognitive science, religion and theology: From human minds to divine minds"*, Templeton Press: West Conshohocken, PA, 2011.

[22] C. Bell, "*Ritual theory, ritual practice"*, Oxford University Press, 1992.

[23] T. F. Driver "*Liberating rites: Understanding the transformative power of ritual*", BookSurge Publishing, 2006.

[24] M. Eliade, "*The sacred and the profane: The nature of religion*", Harvest,1957.

[25] R. L. Grimes, "*The craft of ritual studies"*, Oxford University Press, 2014.

[26] B. Malley, J. Barrett, "Can ritual form be predicted from religious belief? A test of the Lawson–McCauley hypotheses", *Journal of Ritual Studies*, pp. 1-14, 2003.
[27] O. Zell-Ravenheart, M.G. Zell-Ravenheart, "*Creating circles and ceremonies"*, Red Wheel/Weiser, 2006.
[28] V. W. Turner, "*The forest of symbols: Aspects of Ndembu ritual"*, (Vol. 101). Cornell University Press, 1967.
[29] A. Van Gennep, "*The rites of passage"*, University of Chicago Press, 2011.
[30] A. Warburg, W.F. Mainland, W. F. "A lecture on serpent ritual", *Journal of the Warburg Institute*, vol. 2, no. 4, pp. 277–292, 1939.
[31] J. D. Kelly, M. Kaplan, "History, structure, and ritual", *Annual Review of Anthropology*, vol. 19, no. 1, pp. 119–150, 1990.
[32] R. N. McCauley, E.T. Lawson, "Bringing ritual to mind", in *Ecological approaches to cognition: Essays in honor of Ulric Neisser*, pp. 285-312, Psychology Press, 1999.
[33] F. Staal, "The meaninglessness of ritual", Numen, vol. 26, no. 1, pp. 2-22, 1979.
[34] S. E. Fredericks, *"Environmental Guilt and Shame: Signals of Individual and Collective Responsibility and the Need for Ritual Responses"*, Oxford University Press, 2001.
[35] P. Freston, *"Evangelicals and Politics in Asia, Africa and Latin America"*, Cambridge University Press: New York, 2001.
[36] E. R. Leach, "G. Ritualization in man: Ritualization in man in relation to conceptual and social development", *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, vol. 251, no. 772, pp. 403-408, 1966.
[37] M. A. Duncan, *"Duncan's Ritual of Freemasonry"*, Courier Corporation, 2007.
[38] I. Strenski, "What's rite? Evolution, exchange and the big picture", *Religion*, vol. 21, no. 3, pp. 219-225, 1991.
[39] N. C. Chen, M. Drouhard, R. Kocielnik, J. Suh, C.R. Aragon, "Using machine learning to support qualitative coding in social science: Shifting the focus to ambiguity.", *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 8, no. 2, pp. 1–20, 1998.
[40] R. Likert, "A technique for the measurement of attitudes", *Archives of Psychology*, 1932.
[41] A. Joshi, S. Kale, S. Chandel, D.K. Pal, "Likert scale: Explored and explained", *British Journal of Applied Science & Technology*, vol. 7, no. 4, pp. 396-403, 2015.
[42] M. Deutsch, L. Solomon, "Reactions to evaluations by others as influenced by self-evaluations", *Sociometry*, vol. 22, no. 2, pp. 93–112, 1959.
[43] P. Skolnick, "Reactions to personal evaluations: A failure to replicate", *Journal of Personality and Social Psychology*, vol. 18, no. 1, pp. 62–67, 1971, doi: 10.1037/h0030709 1971.
[44] S. E. Taylor, E. Neter, H.A. Wayment, "Self-evaluation processes", *Personality and Social Psychology Bulletin*, vol. 21, no. 12, pp. 1278–1287, 1995.
[45] M. Conley-Tyler, "A fundamental choice: Internal or external evaluation?", *Evaluation Journal of Australasia*, vol. 4, no. 1–2, pp. 3–11, 2005.
[46] H. Yang, J. Shen, "When is an external evaluator no longer external? Reflections on some ethical issues", *American Journal of Evaluation*, vol. 27, no. 3, pp. 378–382, 2006.

Prof. Dr. Daniel Burgos

He works as Vice-rector for International Research (http://research.unir.net), UNESCO Chair on eLearning and ICDE Chair in OER (http://research.unir.net/unesco) at Universidad Internacional de La Rioja (UNIR, http://www.unir.net). He is also Director of the Research Institute for Innovation & Technology in Education (UNIR iTED, http://ited.unir.net). He has contributed to +60 research projects focused on Educational Technology and Innovation, Learning Analytics, Open Education, Social entrepreneurship, eGames, Competence development, and other topics, funded by the European Commission and other institutions. He has published over 180 scientific papers, 20 books and 20 special issues. He has been the jury chair of the 2016 and 2018 UNESCO Prize for the Use of ICTs in Education. He holds degrees in Communication (PhD), Computer Science (Dr. Ing), Education (PhD), Anthropology (PhD), Business Administration (DBA), Theology (PhD), Management (PhD), Open Science & STEM (PhD) and Artificial Intelligence (Postgraduate, MIT).

Artificial Intelligence Seen Through the Lens of Bateson's Ecology of Mind

Dai Griffiths*

Research Institute for Innovation & Technology in Education (UNIR iTED)
Universidad Internacional de La Rioja (UNIR). 26006, Logroño, La Rioja (Spain)

Received 1 April 2021 | Accepted 20 June 2021 | Published 3 August 2021

Abstract

Gregory Bateson developed a number of ideas which are relevant to artificial intelligence, and in particular to the ascription of qualities such as mind, consciousness, spirituality and the sacred. Relevant sections of Bateson's key works are discussed, and his intellectual framework for an ecology of mind is summarized, and in particular his concepts of mind, learning, and the sacred. These are then applied to discuss whether artificial intelligence applications can be considered to possess 'mind'. It is concluded that symbolic artificial intelligence falls short of Bateson's criteria for mind, as do neural networks, although approach more closely. Nor are computers based on the rules of formal logic able to engage with the sacred, which is paradoxical in nature. However, artificial intelligence applications can form part of an ecology of mind and can be involved in the experience of the sacred. Bateson's writing remains a fertile source of ideas relevant to an understanding of the nature and capabilities of artificial intelligence.

Keywords

Bateson, Mind, Artificial Neural Networks, Symbolic Artificial Intelligence, Sacred.

DOI: 10.9781/ijimai.2021.08.004

I. Introduction

According to Marvin Minsky, "we are on the threshold of an era that will be strongly influenced, and quite possibly dominated, by intelligent problem-solving machines" [1]. These words were written over sixty years ago, but only today are we stepping over the threshold. Since Minsky's paper Artificial Intelligence (AI) has demonstrated its huge potential, but the nature and degree of its influence on human beings, and the ways in which it may dominate our society remain uncertain. Indeed, it remains the case that, as Minsky commented, "there is no generally accepted theory of intelligence" [1], and consequently there is no consensus on the nature of artificial intelligence. Given the rapidly increasing power of AI, a clarification of these open questions is ever more urgent.

This special issue seeks to contribute to an understanding of these matters by discussing how the capabilities of AI could be interwoven with the human phenomena of spirituality and analogue thinking. This raises further questions about human biology and thought, the nature of machine thinking, the digital and the analogue, and of the nature of spirituality. To make progress in understanding the relationship between these complex topics, it is necessary to view them not as separate phenomena, each determined according to their own rules, but rather to establish an overarching theory within which they can all be conceptualized. Such theories are few and far between. The work of Gregory Bateson (1904-1980) encompasses all these aspects and provides a theoretical position from which an interwoven discussion of AI, spirituality and analogue thinking becomes possible. This paper builds on Bateson's ideas to reflect on whether an AI can be said to have a mind, and on the possible relationship of such a mind with the spiritual, or as Bateson would have termed it, the sacred. It is hoped that a reexamination of Bateson's ideas, which are now unfamiliar to many, may offer a valuable perspective from which to view the complex and deep issues raised by AI. The intention is to be of interest to both readers familiar with AI who know nothing of Bateson, and readers familiar with Bateson who know nothing of AI. Consequently, no prior knowledge of either area is assumed.

The radical and interdisciplinary nature of Bateson's ideas has led them to remain outside the mainstream. He is perhaps best known for his assertion "In fact, what we mean by information, the elementary unit of information, is a difference which makes a difference" [2] p.459. This remains widely cited, including by Floridi, [3] p.85, perhaps the most prominent current theorist of information (who also points out that Bateson's definition was anticipated by Donald MacCrimmon MacKay [3] p.85). However, this aphorism is just one of a set of interrelated ideas which together defined Bateson's concept of an ecology of mind. Bateson was well aware of computers and their potential power but could not have foreseen the developments which have taken place since his death.

Bateson's thought was extraordinarily wide ranging, unusual, and spread across publications on many themes during a long career. The reader should be aware that this paper offers an individual reading of Bateson, based on a particular route through his writing undertaken for a specific purpose. It does not aspire to giving a complete overview of his work, and most obviously leaves to one side his contributions

* Corresponding author.
E-mail address: david.griffiths@unir.net

to anthropology and psychology. If the interest of the reader is piqued by this discussion, there is no substitute for engaging with the books in which his ideas were set out [2], [4], [5], [6]. A number of valuable studies of Bateson's work are also available, including David Lipset's biographical study [7] and the synthesis and interpretation provided by Peter Harries-Jones [8], [9] and Noel Charlton [10].

II. Methodology

This paper considers the ways in which Bateson's concepts of an ecology of mind, and of the sacred, can shed light on the nature of AI. Within this context, the tasks undertaken by this paper are to:

- outline Bateson's position on dualism and information, which underlies his understanding of mind
- summarize and clarify Bateson's concepts of mind, learning and the sacred
- apply these concepts as criteria for the presence of mind and the sacred in both symbolic AI (which was well established when Bateson was writing), and in neural networks (which are prominent today).

Bateson develops and discusses his concepts of mind and the sacred in four books: *Steps to an Ecology of Mind* [2], *Mind and Nature, a Necessary Unity* [4], *Angels Fear, Towards an Epistemology of the Sacred* [5] (written in collaboration with his daughter Mary Catherine Bateson and published posthumously), and *A Sacred Unity: Further Steps to an Ecology of Mind* [6] (a posthumous selection of unpublished writings). The relevant ideas are not presented in unitary manner, but rather are spread throughout these books. A first task was therefore to read the texts, and to take notes of the key formulations of Bateson's theoretical position, and of those instances where he discusses AI. This was followed by the identification of representative claims made for AI, as set out in the literature. This reading and note taking may be characterized as a 'snowball' approach, as conducted, for example, by Hepplestone et al. [11]. Papers and books were selected based on their title and abstract (where available), and additional sources were identified from references within the articles. The direction of exploration was from the present to the past, gradually uncovering the antecedents of the current discourse. The two bodies of notes were raw material for an analysis of the relationship between the two sets of ideas, evolved in the successive drafts of the paper over a period of three months.

This paper explores the degree to which the application of Bateson's ideas may facilitate an understanding AI. It is not a critique of Bateson's thinking, nor an attempt to update his concepts in the light of recent developments, valuable though such contributions would be. Consequently, Bateson's terminology is given precedence. Thus, the paper discusses 'mind' and 'mental processes' rather than 'intelligence', and 'the sacred' rather than 'the spiritual'. This is a pragmatic response to the task in hand and avoids falling into an examination of competing definitions which would take the place of a substantive discussion.

III. Bateson's View of Mind

A. Differences and Distinctions

Bateson was adamantly opposed to explanations which depended on the identification of a dualism. In particular he rejected Descartes' distinction between "mind or thinking substance" and "extended substance or body" (i.e. the characteristics of physical bodies). [12] p.208-209, which he saw as a strategy for avoiding the problem. In his view dualism is

> a device for removing one half of the problem for explanation from that other half which could be more easily explained. Once separated, mental phenomena could be ignored. This act of subtraction, of course, left the half that could be explained as excessively materialistic, while the other half became totally supernatural. … The materialist superstition is the belief (not usually stated) that *quantity* (a purely material notion) *can determine pattern.* On the other side, the antimaterialist claims *the power of the mind over matter.* [5] p.59

Leaving on one side the responsibility of Descartes for the prevalent dualist view of mind and matter (see [13] for a sympathetic view of Descartes thinking) it remains deeply entrenched in our current thinking about mind. Indeed, it is entangled with ideas of the spirit or soul, and of what it means to be alive. These ideas appear to be of such value to people that they are willing to live with, or even embrace, the contradictions which they generate. Bateson set out to overcome the "formidable barrier" [5] p.12 of Cartesian dualism, and to formulate a system which could accommodate his conviction that "mind and nature form a necessary unity, in which there is no mind separate from body, and no god separate from his creation" [5] p.12. Bateson's proposal of the ecology of mind takes as its starting point this rejection of dualism and is rigorous in following through its implications.

In identifying what Bateson meant by *mind* we must start with his conception of the animate and the inanimate. It should be noted that this is not a dualist explanation, but rather a distinction made in identifying the phenomena to be explained.

> ...we will use Jung's term Pleroma as a name for that unliving world described by physics which in itself contains and makes no distinctions, though we must, of course, make distinctions in our description of it.
>
> In contrast, we will use Creatura for that world of explanation in which the very phenomena to be described are among themselves governed and determined by difference, distinction, and information. [5] p. 18.

Within Pleroma, interactions take place through the transfer of energy, for example in the friction of a meteorite entering the atmosphere, and its subsequent impact crater. This energy is conserved, so the size of the impact crater will be proportional to the velocity of the meteorite and the resistance of the site. In Creatura the situation is entirely different, as there is no relationship between the energy involved in distinction making and the scale of the consequences for the organism. The thunder of a passing truck may lead to less activity than the quiet rattle of a rattlesnake. Indeed, Bateson points out that there can be an inverse energy relationship, as, for example, when an amoeba moves more because it has been deprived of nutrients and is searching for food [2] p.490.

There is a near-infinite number of discontinuities in the environment, and all of these could be inputs into Creatura. This is problematic, as no organism can regulate its relationship with its environment by adsorbing an infinite variety of inputs. In the cybernetic literature this principle has the status of a law, Ashby's law of requisite variety, which states that a regulator's capacity cannot exceed its capacity as a channel for variety [14]. The variety in the Pleroma to which the organism is exposed is attenuated by the act of making distinctions. This attenuation is achieved by selecting from the infinite range of differences in the environment those which are relevant to the ongoing life of the organism. The selection is carried out in part by limits of the organism's sense organs, and in part through the operation of its nervous system.

B. The Nature of Information

The energy involved in distinction making is often minuscule, initiated by the tiny impact of photons on the retina, or sound waves on the ear drum. The energy required for the difference to have any consequences is provided by the organism itself, both in activating the neural pathways, and in subsequent muscular activity, is obtained from within, from its metabolism.

Bateson describes this decoupling of energetic cause and effect in terms of transformation or coding within a circuit. He often illustrated this with the example of a blind person with a stick. Interaction with the environment creates transforms that are transmitted up the stick as vibrations, and then further transformed into neural activity. Bateson warns that "What is transmitted on a neuron is not an impulse, it is news of a difference" [2] p.490. In other words, the news of difference does not carry with it its own interpretation, which is dependent on the system through which it is traveling. It is this "news of difference" which constitutes information. Although Bateson did not discuss the ontological implications of his position, it implies the existence of a 'real world', but one which can only be apprehended at one remove, and which is constructed by the subject. "The mind contains no things, no pigs, no people, no midwife toads, or what have you, only ideas (i.e., news of difference), information about "things" in quotes, always in quotes." [4] p.132.

C. The Nature of Mental Processes

As indicated above, Bateson was attempting the very challenging task of defining mind in terms which avoided proposing mind and matter as different substances [5] p.16. Within this context, and given the concepts outlined above, what constitutes a mental process in Bateson's thinking? What is a mental process composed of? Where is it located? How can it be identified? Fortunately, Bateson was very explicit about the criteria for the existence of a mind[1]:

> *1. A mind is an aggregate of interacting parts or components*
>
> *2. The interaction between parts of mind is triggered by difference*, and difference is a nonsubstantial phenomenon not located in space or time, difference is related to negentropy and entropy rather than to energy.
>
> *3. Mental process requires collateral energy*
>
> *4. Mental process requires circular (or more complex) chains of determination*
>
> *5. In mental process, the effects of difference are to be regarded as transforms (i.e. coded versions) of events which preceded them.* The rules of such transformation must be comparatively stable (i.e., more stable than the content) but are themselves subject to transformation
>
> *6. The description and classification of these processes of transformation disclose a hierarchy of logical types immanent in the phenomena.* [4] p.92 (italics in the original)

Circular causation is required to sustain a mental process, as without it there would be only an isolated event. It should be noted that "a change in any part of the circle can be regarded as cause for change at a later time in any variable anywhere in the circle" [5] p.60. Bateson gives the simple example of a thermostat, in which a rise in ambient temperature can be seen as causing a change in the switch of the thermostat, or the thermostat can be seen as controlling the temperature of the room. Bateson (in common with Hofstadter [15]) ascribes consciousness to recursive circular causation, defining it as "A reflexive aspect of mental process that occurs in some but not all minds, in which the knower is aware of some fraction of his knowledge or the thinker of some fraction of his thought" [5] p.207. It should be noted, however, that Bateson's criteria do not include consciousness, nor do they specify that mind should be contained within a single organism [5] p.210.

In accordance with the six criteria for mental processes above, Bateson was clear that mental processes are digital in nature. This is because mental processes require coded transforms of difference. These in turn require distinction making which turns any analogue value into a digital one through a distinction between the two sides of a threshold. He observed that in animals "the central nervous system and DNA are in large degree (perhaps totally) digital, but the remainder of the physiology is analogic" [4] p.180.

[1] Bateson offered a set of four criteria in an earlier paper of 1969 [2] p.490, which evolved into the definitive set of six discussed here, published in 1979 [4] p. 92 and repeated in a slightly simplified form in 1987 [5] p.18-19.

D. The Ecology of Mind

Bateson's criteria for mental processes are straightforward, and easy to accept, at least for those sympathetic to his non-dualist starting position. Nevertheless, the criteria have implications which are not immediately obvious, and are, indeed, startling. Bateson argues that

> *...any* ongoing ensemble of events and objects which has the appropriate complexity of causal circuits and the appropriate energy relations will surely show mental characteristics. It will *compare*, that is, be responsive to difference (in addition to being affected by the ordinary physical "causes" such as impact or force). It will "process information. [2] p.315 (italics in the original).

Accordingly, Bateson includes within the category of *mental process* "a number of phenomena which most people do not think of as processes of thought" [5] p.16, including embryology, evolution, and "all those lesser exchanges of information and injunction that occur inside organisms and between organisms, and that, in the aggregate, we call *life.*" [5] p.17. This, he implies, is the logical consequence of rejecting a dualist view of mind and matter. To understand these mental processes, he proposed the concept of an *ecology of mind*, which is ecological in the sense that it concerns the interrelations and dependencies between mental systems of all sorts and their environments. In his view, in explaining the behavior of a human being or other organism, "this "system" will usually *not* have the same limits as the "self" – as this term is commonly (and variously) understood." He gives the example of felling a tree with an axe, in which each stroke is modified according to the shape of the cut face of the tree. He sets out the mental process as

> (differences in tree)-(differences in retina)-(differences in brain)-(differences in muscles)-(differences in movement of axe)-(differences in tree), etc. What is transmitted around the circuit is transforms of differences. And as noted above, a difference which makes a difference is an idea or unit of information. [2] p.317.

E. A Tenuous Tradition Building on the Ecology of Mind

Bateson was a unique figure, but he was not entirely alone in his view that the mind was not contained in the brain, and there is a tenuous thread of related work leading to the present which should be briefly discussed here to give context for our discussion.

In their highly influential book *The Embodied Mind* (1991), Varela Thompson and Rosch write that by embodied they mean:

> ...first, cognition depends upon the kinds of experience that come from having a body with various sensorimotor capacities, and second, that these individual sensorimotor capacities are themselves embedded in a more encompassing biological, psychological, and cultural context. [16] p.173.

This view clearly has a great deal in common with Bateson's ecology of mind. Both Varela and Thompson were aware of Bateson's thinking [17] although they do not cite him in their book

Edwin Hutchins analyzed the processes of navigation by teams in ships, and came to the conclusion that "The central computations of the navigation tasks are accomplished by the propagation of information across representations and representational media." [18] p.218. He recognized the influence of Bateson in his focus on mapping propagation of information beyond the limits of an individual, writing "I take the fundamentals of an architecture of cognition and a sense of a unit of analysis from Gregory Bateson". [18] p.291.

In 1998 Andy Clark and David Chalmers wrote an influential paper called 'The Extended Mind' [19], which built on Varela's ideas. In recent years here has been an upsurge of interest in '4e cognition' which brings together Gibson, Varela, Hutchins and Chalmers "under one heading and conceives of them as coherently opposed to the internalist, brain-centered views of cognitivism" [20] p.4. A related area in which there has been an active discussion in recent years has been that of cognition in plants, or as it is perhaps more correctly

termed, plant gnosophysiology [21], which is explicitly linked with extended cognition [22].

The fact that this tenuous thread of thinking around extended cognition now seems to be gathering some degree of prominence makes it timely to reassess the relevance of Bateson's, particularly as his contribution has been so widely forgotten or ignored.

IV. Bateson and Symbolic Artificial Intelligence

A. Symbolic Artificial Intelligence

The development of artificial intelligence technology is a convoluted story, with many interconnecting strands, which this paper will not attempt to describe. The reader who would like to explore this history is referred to 'Mind as Machine', Margaret Boden's two volume history of the field [23]. For present purposes, it is sufficient to distinguish between systems based on formal logic (which were available during Bateson's lifetime), and those making use of neural networks (which had been conceived of but were not developed in a practical form).

Bateson was central figure in the establishment of cybernetics in the 1940s and a prominent participant in the seminal Macy conferences [24]. As such he would have been well aware of the ongoing development of artificial intelligence, and knew a number of the leading figures, including John von Neuman, Warren McCulloch and Norbert Wiener [25]. Indeed, artificial intelligence was established as a distinct field in a split from cybernetics at the Dartmouth Summer Research Project in 1956, partly as a result of personality clashes with Norbert Wiener, who was perceived to be the public face of cybernetics [26] p.78.

In 1957 Herbert Simon and Allen Newell made the bold claim that machines making use of heuristic problem-solving methods would, within ten years, be the world's chess champion, prove an important new mathematical theorem, write music accepted by critics as having considerable aesthetic value, and that most theories in psychology would take the form of computer programs [27]. Looking back on their work in 1976 they noted that it was based on the hypothesis that "A physical symbol system has the necessary and sufficient means for general intelligent action." [28]. The approach taken by Simon and Newell became dominant and was famously characterized by Haugeland in 1985 as Good Old Fashioned Artificial Intelligence, or GOFAI, as it became known. He identified the essential claims of GOFAI theories to be:

> 1. our ability to deal with things intelligently is due to our capacity to think about them reasonably (including sub-conscious thinking); and
>
> 2. our capacity to think about things reasonably amounts to a faculty for internal "automatic" symbol manipulation. [29] p.82

In line with Haugeland's second claim, this tradition has been known as 'symbolic AI', which is the term we will use here. It has also been referred to as 'cognitivist', since "according to this view cognition occurs by taking in information provided by the environment, forming this into representations which can then be processed to provide logical responses by way of activity." [30] p.106. This tradition continues to be of significance in the present day, often in combination with newer machine learning methods as proposed, for example, by Gary Marcus in his 2020 paper The Next Decade in AI [31].

In 1972, when Bateson published *Steps to an Ecology of Mind*, symbolic AI researchers were still confident of the success of their paradigm. However, this was also the year that Hubert Dreyfus published a book arguing that a system that could use natural language and recognize complex patterns would need a body, and that such robots would need to be entirely different from present digital computers [32]. p.216. The furious response to Dreyfus' book suggested that his admittedly sharp critique of symbolic AI had touched a nerve. By the time that Mind and Nature was published in 1979 and it had become clear that the trajectory of research established at Dartmouth had encountered major problems. Bateson did not participate in the heated discussion around AI, but the ambitions and shortcomings of AI research formed an important part of the backdrop to his thinking, which was highly relevant to the debate.

B. Symbolic AI and Bateson's Criteria for Mind

Bateson's inquiry into mind was an unusual one within the cybernetics and AI communities. One of the few who discussed computers in terms of mind was Simon, who is reported by McCorduck [33] p.151 as saying that through his work in artificial intelligence he had arrived at "a notion that a mind was something which took some program inputs and data and had some processes which operated on the data and produced output." In other words, an understanding of mind did not inform Simon's understanding of AI, rather it was the emerging capabilities of AI which were taken as a basis for understanding the nature of mind. Similarly, the book *The thinking computer: mind inside matter* by Bertram Raphael, published in 1976 [34], not only proclaims in its title the dualism that Bateson abjured, it also consider avoids any discussion of what might constitute a mind. Raphael simply claims that if a computer successfully models the processes followed by a human in carrying out a task, then "we can view the flow chart of its program as a plausible guide to the logic of the inner workings of the mind." [34] p.300.

Bateson's criteria for mind (see section III.C, above) offer a perspective from which to view the claims of symbolic AI. It is clear that 1, 2, & 3 are fulfilled, since computers running AI applications are (1) an aggregate of interacting components, (2) triggered by difference and (3) employ collateral energy. Criteria 4, 5, & 6 need more consideration.

Criterion 4 "requires circular (or more complex) chains of determination" [4] p.92. There are certainly plenty of loops in computer code, for example a sub-routine may be called frequently in the execution of a program, always returning to the starting point so that the program can proceed. Nevertheless, circular chains of determination contravene the formal logic applied in symbol manipulation and are treated as a bug or a malware attack by both conventional computers and by symbolic AI. In other words, the system is designed to prevent a program changing its own functioning or the computational environment in which it operates, in response to its own operation. In biology, on the other hand, "many regularities are part of – contribute to– their own determination." [5] p.161.

Criterion 5 is fulfilled, in that computer circuits do indeed involve coded transforms of events which preceded them. However, the rules which govern these transformations are not, in classical computer architectures, subject to transformation without the intervention of a programmer.

Criterion 6 is that "The description and classification of these processes of transformation disclose a hierarchy of logical types immanent in the phenomena." In this regard Simon's *The Sciences of the Artificial* demonstrates that the symbolic AI community was well aware of hierarchy, which Simon discuses in terms that Bateson might well have approved of:

> ...complexity frequently takes the form of hierarchy and ... hierarchic systems have some common properties that are independent of their specific content. [35] p. 87.

Still more Batesonian is Simon's argument that hierarchies can take the form of different levels of description, commenting that the genetic description of a single cell may therefore take a quite different form from the genetic description that assembles cells into a multi celled organism. [35] p.115.

In practical computing applications, however, these structures were limited to branching classification trees, determined by the programmer. The contrast with Bateson's approach can be seen clearly in the question posed by Raphael "If we wish to insert knowledge into a computer what kinds of concepts must we represent?" [34] p.48. This conception of knowledge as something independent of the knower, which can be 'injected', is far removed from Bateson's view that the meaning of information is dependent on its context.

A deeper problem implied by criterion 6 is that it refers not only to a hierarchy, but to a "hierarchy of logical types immanent in the phenomena". Bateson's understanding of logical types was based on Whitehead and Russell's *Principia Mathematica.* He summarized the principles as being:

> no class can, in formal logical or mathematical discourse, be a member of itself; that a class of classes cannot be one of the classes which are its members; that a name is not the thing named… a class cannot be one of those items that are correctly classified as its non members... if these simple rules of formal discourse are contravened, paradox will be generated… [2] p.280.

In Bateson's view, the information flow in an organism is logically typed, but not neatly separated in the way that a programmer might define a set of classes. Rather, the organism generates an enormous and tangled network of messages [4] p.109, within which complex relationships of logical types emerge, although the observer may find these hard to identify. I understand this to be an implication of Bateson's use of the word 'immanent' in his criterion.

C. Symbolic AI in an Ecology of Mind

We have seen in the previous section that computers running symbolic AI applications are lacking two of Bateson's criteria for mind: circular chains of causation, and complex interactions between logical types. However, this does not mean that they do not constitute parts of in an ecology of mind.

> Bateson argues that
>
> ...in no system which shows mental characteristics can any part have unilateral control over the whole. In other words, *the mental characteristics of the system are immanent, not in some part, but in the system as a whole.* [2] p.316.

The question then arises of what constitutes 'the system' in a symbolic AI application. As we have seen in section III.D, viewed within an ecology of mind, a system does not usually have the same limits as the self, but rather is constituted by the limits of the flow of information within an ecology of mind. Consequently, the questions "Can a computer think?" and "Is the mind in the brain?" should be answered in the negative [2] p.316 (with the possible exception of processes which monitor the internal states of the computer or brain). More precisely:

> … it would be incorrect to say that the main business of the computer – the transformation of input differences into output differences – is "a mental process". The computer is only an arc of a larger circuit which always includes a man and an environment from which information is received and upon which efferent messages from the computer have effect. This total system, or ensemble, may legitimately be said to show mental characteristics. It operates by trial and error and has creative character.
>
> Similarly, we may say that "mind" is immanent in those circuits of the brain which are complete within the brain. Or that mind is imminent in circuits which are complete within the system, brain *plus* body. Or, finally, that mind is immanent in the larger system – man *plus* environment. [2] p.317.

Discussing the blind person with a stick, Bateson asks where that person's self begins. For Bateson the answer was *at the tip of the stick*, because the any other location would "draw a delimiting line across this pathway is to cut off a part of the systemic circuit which determines the blind man's locomotion." [2] p.318. On this basis, we make the same mistake if we draw a delimiting line between a human and artificial intelligence.

We can now provide a response to the question "Does a symbolic AI application have a mind?". From the perspective of the ecology of mind the answer is "No, but that is not the right question to ask". It is more valuable to ask "What is the structure of the ecology of mind generated when a person interacts with an AI?" The answer will involve mapping information flows (i.e. coded transforms of distinctions), while paying attention to their logical types and to recursive causation.

V. Bateson's Ideas and AI in the 21st Century

A. Neural Networks and Deep Learning

In 1985 Haugeland was able to write that "AI more or less ignores learning" [29] p.11. This is by no means the case today, when technologies related to 'machine learning' using neural networks, often adopting a connectivist approach, have become established as the focus for most AI research, while symbolic AI retains importance as an established method.

The history of neural networks is usually traced back to a paper of 1943 by McCulloch and Pitts which provided a Logical Calculus for Nervous Activity. Both the authors knew Bateson and coincided with him at the seminal Macy conferences [24]. We may therefore assume that Bateson would have known of their approach to AI and would not have ignored it in developing his thinking on the ecology of mind. However, given the low level of achievements of neural networks at the time, there was no need to pay particular attention to this line of work in his writing. Bateson's criteria for mind make no reference to the structure of information flows, nor the material through which they pass. Consequently, the discussion of mind in relation to symbolic AI can, in principle, also be applied to neural networks.

Since Bateson was writing, however, neural networks have become vastly more powerful, and have demonstrated astonishing capabilities. Machine learning methods examine the relationship between inputs and outputs in a set of data, with the following requirements set out in [36] p.22.

1. Data (a set of historical examples).
2. A set of functions that the algorithm will search through to find the best match with the data.
3. Some measure of fitness that can be used to evaluate how well each candidate function matches the data.

The measure of fitness is used iteratively to adjust the functions to arrive at the best available fit with the data. The term 'deep learning' has been coined to refer to machine learning methods which deploy algorithms in layers, each of which is optimized by the network. That such systems are powerful is not in doubt. They "control operations as diverse as labeling images, recognizing speech, translating texts, playing strategy games, predicting protein folds, detecting new exoplanets, analyzing fMRI data, and driving automobiles autonomously" [37] and there is no limit in sight to what may be achieved in future. Be this as it may, the aspect of deep learning which most concerns us here is its recursive structure, in which the results of information processing change the way in which this processing is carried out. This requires a reconsideration of Bateson's criteria 4 & 5, both of which were partially fulfilled by symbolic AI.

1. Bateson's Criteria for Mind Applied to Deep Learning

As regards criterion 4, there is no doubt that deep learning applications have a greater capability to change their own functioning

than do symbolic AI applications. However, the reach of these changes is tightly constrained to specified algorithms in a fixed architecture.

The situation for the closely related criterion 5 is similar. In common with symbolic AI, deep learning fulfills the requirement that effects of difference are transforms of events which preceded them. As we noted above, however, the problem arises with Bateson's additional requirement that the rules which govern that transformation should be subject to transformation. To some extent, the implementation in machine learning of recursive change of the functions used in transformation addresses this requirement. Moreover, the use of 'evolutionary' or 'genetic' algorithms, which has been developing in AI since the 1960's, has become an established technique (see De Jong [38] for a historical overview), and this enables more substantial changes to be made to the rules. However, as a biologist, when Bateson wrote about changing rules he was surely thinking of evolution, and, as we have already noted, he viewed embryology and evolution as mental processes [5] p.16. From the perspective of biological evolution, the evolution that takes place in deep learning is rather superficial. This is because the applications which determine the changes in the algorithms are themselves excluded from evolutionary change, and the same is true for the hardware on which the system runs. Zaadnoordijk, Besold, and Cusack [39] point out that machine learning has been based on adult learning, but that a study of specific processes in the cognitive development infants might produce valuable techniques. Such a focus on developmental change could, perhaps, move towards a more complete fulfillment of criterion 5.

Thus, the development of deep learning takes a step towards fulfillment of criteria 5 & 6, but still requires the presence of a human to meet the requirements.

2. Levels of Learning

Bateson makes a perceptive comment in this context: "The question is not "Can machines learn?" but what level or order of learning does a given machine achieve?" [2] p.284. He is precise about what he means by levels of learning, and his definition is closely related to the logical types introduced in section IV.B (see *The Logical Categories of Learning and Communication* [2] p.279-308 for a full discussion of the relationship summarized below).

Zero learning is defined as that in which information is simply stored and reproduced at the appropriate time, for example learning the time of an appointment. Bateson comments that "many very simple mechanical devices show at least the phenomenon of zero learning" [2] p.284, adding that "a very high (but finite) order of complexity may characterize adaptive behavior based on nothing higher than zero learning" [2] p.284. This complexity is possible because information of many different logical types may be managed within a finite and constrained architecture, even though the AI application that is doing the learning is constrained to zero learning.

Level I learning involves "the class of phenomena which are appropriately described as *changes* in zero learning (as "motion" describes change of position)" [2] p.287. In level 1 learning the entity gives different responses at different times. In an organism, this may be, for example, a result of habituation or reinforcement. In the machine, level 1 learning is absent in symbolic AI, but is clearly present in deep learning applications.

Level II learning, put simply, is 'learning to learn', for example one might learn to perform better at rote learning tasks. This involves changes in the process of learning, and recognition of new contexts which require different responses. Bateson terms this "changes in the manner in which the stream of action and experience is segmented or punctuated into contexts together with changes in the use of context markers" [2] p.293. He illustrates this with the example of 'reversal learning' experiments in which the subject is taught that X = R1, and that Y = R2. Once this has been learned the relationship is reversed. Level II is manifested in an improvement of the subject in recognizing the reversal and adapting to it.

Our discussion suggests that deep learning is pushing towards achievement of Level II but has not achieved it. There is no equivalent in deep learning to the developmental changes of children, or when an entirely new set of capabilities is acquired through learning mathematics or a musical instrument from a position of complete ignorance. The developmental approach recommended by Zaadnoordijk, Besold, and Cusack [39], cited above, indicates a possible route forward for deep learning in this respect.

Bateson also discusses a rather more elusive *level III learning*, which he describes as "likely to be difficult and rare even in human beings", involving "profound reorganization of character" [2] p.301. However, this is out of scope for a discussion of current AI.

VI. The Sacred

A. The Sacred and the Spiritual

This special issue is concerned with AI and spirituality, but Bateson largely avoided the word *spirituality* in his writing. This is perhaps because the meaning of the word is so tightly bound up with the dualism spiritual-material. An exception to this rule is that he defines *sacrament* as the "outward and visible sign of an inward and spiritual grace" [4] p.230. In this case he was perhaps describing religious practices as an anthropologist, using the terminology of the field. In any event, he never invokes the spiritual as part of an explanation for the phenomena which he discusses. On the other hand, the sacred is a frequent theme in his writing. The meaning which he attached to the sacred was elusive, for reasons which will become clear in our discussion below, but two characteristics can be made clear as a starting point. Firstly, Bateson did not see the sacred as in any way opposed to a scientific understanding of the world, and he was convinced that "there are strong arguments for the necessity of the sacred, and that these arguments have their base in an epistemology rooted in improved science and in the obvious." [5] p.11. Secondly, he situated the sacred (as Mary Catherine Bateson puts it) in "the integrated fabric of mental processes that envelops all our lives" [5] p.200, and consequently the concept of the sacred is an integral part of the ideas which we have been discussing in this paper.

B. Paradox and Causation

Bateson frequently referred to Epimenides paradox.

> ...the ancient paradox of Epimenides - "Epimenides was a Cretan who said, 'Cretans always lie'" - was built upon classification and metaclassification. I have presented the paradox here in the form of a quotation within a quotation, and this is precisely how the paradox is generated. The larger quotation becomes a classifier for the smaller, until the smaller quotation takes over and reclassifies the larger, to create contradiction. [4] p.116-117.

The same oscillation can be seen in physical systems, for example in a buzzer circuit:

- If contact is made at A, then the magnet is activated.
- If the magnet is activated, then contact at A is broken.
- If contact at A is broken, then the magnet is inactivated.
- If magnet is inactivated, then contact is made. [4] p.59.

The question arises as to why the Cretan liar is paradoxical, whereas a buzzer is an unproblematic piece of everyday equipment. The answer, argues Bateson, lies in a confusion about the meaning of the word 'if', which can either refer to causal relationships (if an induction magnet is activated, then a nail will be attracted to it) or

to logical relationships (if all men are mortal, and Socrates is a man, then Socrates is mortal). Thus, the sequence in the buzzer circuit makes perfect sense when seen as a causal description, but none whatsoever as a sequence of logical propositions. The difference is the result of the inevitable inclusion of time in causal relationships, so that the description of the buzzer circuit is a set of sequential steps, each of which supplants the previous one. In contrast, the steps of the syllogism showing that Socrates is mortal are valid simultaneously, and permanently. Bateson argues that this has major implications for computers.

> The "if … then …" of logic contains no time. But in the computer, cause and effect are used to *simulate* the "if … then …" of logic, and all sequences of cause and effect necessarily involve time. [2] p.281.

The code which is run on a computer is an abstract logical structure which stands outside of time. However, when the code is instantiated in digital circuits, it operates as sequences of cause and effect which exist in time. The result is, as Norbert Wiener pointed out, that a computer would encounter the Cretan liar not as a paradox, but rather as an oscillation YES . . . NO. . . YES . . . NO . . . until it runs out of energy. [4] p.117. This, if it were permitted by the programmers of the computer, would be experienced by the user as a malfunction of the computer.

Among other consequences, computers as we as we are familiar with them, with their ground rules of logic and respect for logical types, appear to be precluded from the possibility of being conscious in Bateson's sense of reflexive and recursive mental processes, a perspective which is explored at length by Hofstadter in his discussion of 'strange loops' [15].

More generally, Bateson argues that biological systems, including brains, are networks of causal links. Furthermore, every circuit of causation in biology, in physiology and neural processes, and in ecological and cultural systems, "conceals or proposes those paradoxes and confusions that accompany errors and distortions in logical typing." [4] p.109. In practice, it is exceedingly difficult to decide that aspects of an organism's activity are in a meta-level relationship to others [4] p.117. It is not possible to render such process in a set of logical links, without violating the rules of logical types established by Whitehead and Russell to exclude paradox [40].

C. Paradox and the Sacred

A metaphor is a kind of syllogism, but one which is not held together by the logical links of Socrates' mortality, discussed above. Bateson gives the example, [2] p.205, of

- Grass dies
- Men die
- Men are grass.

He argues that this is the way that biological homology is best understood, as for example, "a formal similarity that suggests a relationship, like that between a human hand and the wing of a bat" [5] p.192. Such formal similarities emerge not from logical connections, but from a vast network of causal relationships full of circularity and contradictions in logical typing, with an associated lack of clarity of what is causing what, and at what level of logical type. He also proposes that this is how poets think, and, we might add, other artists. In this he echoes his contemporary Arthur Koestler, whose concept of bisociation [41] analyzed all creativity in terms of bringing together intersecting planes of associations, with concomitant violation of logical categories. Bateson knew Koestler from the Macy conferences, but strongly rejected some of his ideas [5] p.57-58.

In our normal waking states, we make internal or external reports of our perceptions, in a state which Bateson refers to as *prose consciousness*, and which he associated with the left hemisphere of the brain. In this state we are quite able to label the thing that we perceive as a symbol, for example a stop sign on the road. We can even label it as a metaphor, and parse that metaphor into its components. But we also have other states, where the identification of hand and bat, or wine and blood, is not labeled with a logical hierarchy, but experienced as an identity. This mode of thinking, familiar from dreams, and also present in (for example) aesthetic experiences, trances of various sorts, religious experience and the intensity of love. In these states the difference between the logical types of the map and the territory is dissolved, and we return to the "innocence of communication by means of pure mood-signs" [2] p.183. This state is the 'inward and spiritual grace' of which the sacrament is an 'outward and visible sign". From this perspective, the sacrament is more than a metaphor, but is rather seen as the thing itself, leading, for example, soldiers to sacrifice themselves to save a flag, and for martyrdom to be embraced to defend the idea of the transubstantiation of the host.

In our interior life, and in our relations with our environment, human beings participate in both logical and causal circuits, in prose consciousness and the transcendent, and in the rational and the emotional. Indeed, paradox is central to the most widely recognized sacrament in western society, the mass, in which the bread and wine are both themselves and the body and blood of Christ. How is this to be understood. Is the bread transformed into the body of Christ during the mass, through which we can experience union with Christ? Or is it a symbol for the body of Christ, whose contemplation can lead to religious insight?

Bateson suggests that "the richest use of the word "sacred" is that use which will say that what matters is the *combination* of the two… any fracturing between them is, shall we say, anti-sacred" [6] p.267. Any attempt to analyze a specific example of this cohabitation between the different visions requires, instead of a unified experience, the alternating view of that experience from the two different standpoints, dissolving the phenomena which we hope to analyze, an alternation reminiscent of the computer's response to the Cretan liar.

Such a combined experience of opposites involves paradox not only in operating with the conflicting premises of two mutually incompatible types of interaction, but also in considering the nature of the resulting combined entity. It was because of this that Bateson stated that "To be conscious of the nature of the sacred or of the nature of beauty is the folly of reductionism" [4] p.214.

It follows from this position that

> *noncommunication* of certain sorts is needed if we are to maintain the "sacred." Communication is undesirable, not because of fear, but because communication would somehow alter the nature of the ideas [5] p.80.

It is this which leads Bateson to be elusive in his descriptions of the sacred. He suggests that this is part of wider phenomenon, whereby there may be processes in all living systems such that "if news or information of these processes reaches other parts of the system, the working together of the whole will be paralyzed or disrupted" [5] p.81.

The conceptual framework outlined above places the sacred outside the domain of AI as we know it. Any engagement with the sacred requires and engagement with and tolerance of paradox. AI applications that are currently conceived of, running as they do on von Neumann architectures, are unable to encounter paradox. Consequently, they cannot, in themselves, engage in the mental tight-rope walk involved in the merging these perspectives which Bateson sees as being the core of the sacred. Thus, an AI built on current design principles is systemically unable to experience, or even to represent, an important aspect of the human mind. This implies a constraint on the ability of AI to interact with a human being in a way which would enable it to substitute for a human caregiver or teacher.

In our discussion of mind in AI we saw that although AI does not in itself fulfill the criteria for mind, it can be a significant element within a wider ecology of mind. Something similar may apply to the sacred. There is no reason why the extraordinary logical structures generated by computers should not be a powerful component of the sacred. Perhaps the increasing power of simulations will give rise to new opportunities for experience of the sacred, as one pole of a combined experience. Readers who find the possible association of simulations, including those involving sex or violence, with the sacred, should bear in mind that the Latin root of the word, *sacer*, referred not only to the extremes of holy and pure, but also those of the unholy and impure [6] p.267.

VII. Concluding Remarks

The application of Bateson's ideas to AI is not intended to constitute a solution to the difficult questions which surround AI. Nor is it suggested that the insight obtained supplants other work carried out since Bateson's death. It is, however, proposed that there are valuable characteristics in Bateson's thought which can inform the current debate on AI.

Firstly, Bateson's work is based on strong foundations. His analysis starts with an explicit statement of the nature of information, but, in contrast, much of the literature of AI is silent on this. Similarly, Bateson is rigorous in his rejection of dualism, following through the implications for the nature of mind. A lack of clarity on these issues may or may not be a problem in the practical tasks of building AI applications, but a reading of Bateson suggests that this lack is a barrier to conceptualizing the phenomena generated by those applications. Whether or not one agrees with Bateson's views, the admirable clarity of his position provides an example which could usefully inform current attempts to improve our understanding of what AI is, and how humans interact with it.

Secondly, the explanations offered by Bateson are functional, and he ascribes the properties of things to their structure. There is therefore no obstacle in principle to AI achieving human mental abilities. The constraints on AI which we have identified in this article are related to the structure of computers as we know them, and as we can presently conceive of them. There is every reason to suppose that Bateson would have agreed with Chalmers when he argued that a neural description of the brain, translated into a combinatorial-state autonoma, would have experiences indistinguishable from the brain [42] p.321. Neural networks have moved some distance in this direction with increasingly sophisticated models of the behavior of neurons, see for example [43], and further progress is surely to be expected. In this context Bateson's ideas can make a valuable contribution by focusing attention on the levels of learning which are exhibited in machine learning, and on the scope of adaptive change which is required if AI is to become equivalent to its organic counterpart.

Thirdly, as Denning and Tedre pithily put it, in deep learning applications "All there is inside is an inscrutable, complex mass of connections." [44] p.173. This aspect of deep learning is intriguing, because it moves AI in the direction of Bateson's description of the equally inscrutable tangled network of messages in organic brains, within which complex relations of logical types are imminent. However, the rather rigid layering of the algorithms which run deep learning applications would seem to militate against the development of recursion in the mass of connections in deep learning applications. Leaving to one side the complex architectural issues which arise, Bateson's ideas suggest that it would be interesting to explore the results of loosening the prohibition of recursion in the networks of connections in machine learning, and indeed encouraging it.

Lastly, one of the most challenging aspects of Bateson's ecology of mind is the idea that mind does not end at the physical limits of an organism or machine, but rather at the limits of the information flows which constitute the mind. However, the alternative to this view is equally difficult to assimilate, i.e. that a mind is constituted of something other than information flows, by a mental stuff which is present in brains, and perhaps in AI, but which we have yet to detect. Some nodes in the ecology of mind are clearly more powerful than others. When I interact with a dog, I am aware that I have mental capabilities which the dog does not have (although it doubtless has some important capabilities, for example relating to smell, which I lack). The same is true of my interactions with the computer on which I am typing this text. As a result, the search for, and the potential deification of, a discrete super-mind is misleading from Bateson's perspective. Whatever is developed in the future will participate in an ecology of mind with all the organisms and AIs with which it is in contact. Indeed, from Bateson's perspective, it is hard to see how the AI could be useful or effective without that network of information flows, within and between components of the ecology,

Nevertheless, just as humans are peak predators in the ecology of energy, they are also peak nodes in the information flows of an ecology of mind. The singularity, popularized by Kurzweil [45], suggests that once AI surpasses human capabilities, it will accelerate exponentially past us, and become superhuman. There is understandable concern regarding what such a superhuman entity might choose to do to its progenitors. But Bateson's writings suggest that we ask another kind of question, one which should not wait until the postulated singularity arrives. Our immediate concern should be "what is the impact of increasing AI capabilities on the ecology of mind, and how does this change the niche of human beings within that ecology".

References

[1] M. Minsky, "Steps Toward Artificial Intelligence," *Proceedings of the Institute of Radio Engineers*, vol. 49, no. 1, pp. 8–30, 1961, doi: 10.1109/JRPROC.1961.287775.

[2] G. Bateson, *Steps to an Ecology of Mind*, 2nd ed. Chicago, IL: Chicago University Press, 2000, first published 1972.

[3] L. Floridi, *The Philosophy of Information*. Oxford, UK: Oxford University Press, 2011.

[4] G. Bateson, *Mind and Nature: A Necessary Unity*. New York, NY: E. P. Dutton, 1980, first published 1979.

[5] G. Bateson and M. C. Bateson, *Angels Fear: Towards an Epistemology of the Sacred*, 2nd ed. Cresskill, NJ: Hampton Press, 2005, first published 1987.

[6] G. Bateson, *A Sacred Unity: Further Steps to an Ecology of Mind*. New York, NY: Bessie/HarperCollins, 1991.

[7] D. Lipset, *Gregory Bateson: The Legacy of a Scientist*. Eaglewood Cliffs, NJ: Prentice Hall, 1980.

[8] P. Harries-Jones, *A Recursive Vision: Ecological Understanding and Gregory Bateson*. Toronto, Ontario: University of Toronto Press, 1995.

[9] P. Harries-Jones, *Upside-Down Gods: Gregory Bateson's World of Difference*. New York, NY: Fordham University Press, 2016.

[10] N. G. Charlton, *Understanding Gregory Bateson: Mind, Beauty, and the Sacred Earth*. Albany, NY: State University of New York Press, 2008.

[11] S. Hepplestone, G. Holden, B. Irwin, H. J. Parkin, and L. Thorpe, "Using technology to encourage student engagement with feedback: A literature review," *Research in Learning Technology*, vol. 19, no. 2, pp. 117–127, 2011, doi: 10.1080/21567069.2011.586677.

[12] R. Descartes, *The Philosophical Writings of Descartes*. Cambridge, UK-Cambridge University Press, 1984.

[13] G. Baker and K. Morris, *Descarte's Dualism*. London, UK: Routledge, 2005.

[14] R. Ashby, "Requisite variety and its implications for the control of complex systems," *Cybernetica*, vol. 1, no. 2, pp. 83–99, 1958.

[15] D. Hofstadter, *Godel, Escher, Bach: An Eternal Golden Braid*. New York, NY: Basic Books.

[16] F. J. Varela, E. Thompson, and E. Rosch, The Embodied Mind. Cambridge, MA: MIT Press, 1993.
[17] E. Thompson, "Life and Mind: From Autopoiesis to Neurophenomenology," *Phenomenology and the cognitive Sciences* vol. 3, no. 4, pp. 381-398, 2004.
[18] E. Hutchins, *Cognition in the Wild*. Cambridge, MA: MIT Press, 1995.
[19] A. Clark and D. Chalmers, "The extended mind," *Analysis*, vol. 58, no. 1, pp. 7–19, 1998.
[20] A. Newen, L. De Bruin, and S. Gallagher (Eds.), *The Oxford Handbook of 4e Cognition*. Oxford, UK: Oxford University Press, 2018.
[21] D. Michmizos and Z. Hilioti, "A roadmap towards a functional paradigm for learning & memory in plants," *The Journal of Plant Physiology*, vol. 232, pp. 209–215, 2019.
[22] A. G. Parise, M. Gagliano, and G. M. Souza, "Extended cognition in plants: is it possible?," *Plant SignalLing and Behavior*, vol. 15, no. 2, pp. 1–10, 2020.
[23] M. A. Boden, *Mind as Machine: a History of Cognitive Science, volumes 1 & 2*. Oxford: Oxford University Press, 2006.
[24] C. Pias, *Cybernetics - The Macy Conferences 1946–1953: the complete transactions*. Berlin: diaphenes, 2003.
[25] S. P. Heims, "Gregory Bateson and the Mathematicians: From Interdisciplinary Interaction to Societal Functions," *Journal of the History of the Behavioral Sciences*, vol. 13, no. 1977, pp. 141–159, 1977.
[26] N. J. Nilsson, *The Quest for Artificial Intelligence: A History of Ideas and Achievements*. Cambridge, UK: Cambridge University Press, 2010.
[27] H. A. Simon and A. Newell, "Heuristic Problem Solving: The Next Advance In Operations Research," *Operations Research*, vol. 6, no. 1, 1958.
[28] A. Newel and H. A. Simon, "Computer Science as Empirical Inquiry: Symbols and Search," *Communications of the Association for Computing Machinery*, vol. 19, no. 3, 1976.
[29] J. Haugeland, *Artificial Intelligence: The Very Idea*. Cambridge, MA: The MIT Press, 1985.
[30] J. Mingers, "Embodying Information Systems: The Contribution of Phenomenology," *Information and Organization*, vol. 11, no. 2, pp. 103–128, 2001.
[31] G. Marcus, "The next decade in AI: Four steps towards robust artificial intelligence," *arXiv*, no. February abs/2002.06177., 2020.
[32] H. L. Dreyfus, *What Computers Can't Do: A Critique of Artificial Reason*. Harper & Row, New York, 1972.
[33] P. McCorduck, *Machines Who Think*. Natick, MA: A K Peters, Ltd, 2004.
[34] B. Raphael, *The thinking computer: mind inside matter*. San Francisco, CA: W. H. Freeman, 1976.
[35] H. A. Simon, *The Sciences of the Artificial*, Cambridge, MA: MIT Press, 1969.
[36] J. D. Kelleher, *Deep Learning*. Cambridge, MA: MIT Press, 2019.
[37] C. Buckner, "Deep learning: A philosophical introduction," *Philosophy Compass*, vol. 14, no. 10, pp. 1–19, 2019.
[38] K. A. De Jong, *Evolutionary Computation: A Unified Approach*. Cambridge, MA: MIT Press, 2006.
[39] L. Zaadnoordijk, T. R. Besold, and R. Cusack, "The Next Big Thing(s) in Unsupervised Machine Learning: Five Lessons from Infant Learning," *arXiv*, 2020.
[40] A. N. Whitehead and B. Russell, *Principia Mathematica*. Cambridge: Cambridge University Press, 1910.
[41] A. Koestler, *The Act of Creation*. London: Pan books, 1977.
[42] D. J. Chalmers, *The Conscious Mind: In Search of a Fundamental Theory*. Oxford, UK: Oxford University Press, 1996.
[43] A. Jeyasothy, S. Sundaram, and N. Sundararajan, "SEFRON: A New Spiking Neuron Model with Time-Varying Synaptic Efficacy Function for Pattern Classification," *IEEE transactions on neural networks and learning systems*, vol. 30, no. 4, pp. 1231–1240, 2019.
[44] P. J. Denning and M. Tedre, *Computational Thinking*. Cambridge, MA: MIT Press, 2019.
[45] R. Kurzweil, *The Singularity is Near*. London, UK: Viking, 2005.

Dai Griffiths

Dai Griffiths, also known as David, has a background in the arts, and holds a PhD from Universitat Pompeu Fabra in the area of ICT. He spent the first part of his career working as a teacher in primary, secondary and higher education, as well as in interpersonal skills training in industry, before becoming fascinated by the potential of computers in education. For the past twenty-five years he has worked in the development of educational applications, and as an educational technology researcher. He has published extensively, and details are available at https://orcid.org/0000-0002-6863-2456. In this work he became deeply engaged with the tradition of cybernetics, and he remains active in this field. He was Professor of Educational Cybernetics at the University of Bolton. At the University of Bolton he was a member of the Institute for Educational Cybernetics, and of Cetis. He then took on a role in the Department of Education of Bolton University, leading the Department's PhD and Doctor of Education programs. Dai Griffiths is currently a Senior Researcher at the Research Institute for Innovation & Technology in Education (UNIR iTED), Universidad Internacional de La Rioja (UNIR). He is an Associate of Cetis LLP.

unir
LA UNIVERSIDAD
EN INTERNET

Rectorado
Avenida de la Paz, 137
26006 Logroño (La Rioja)
t (+34) 941 21 02 11

www.unir.net