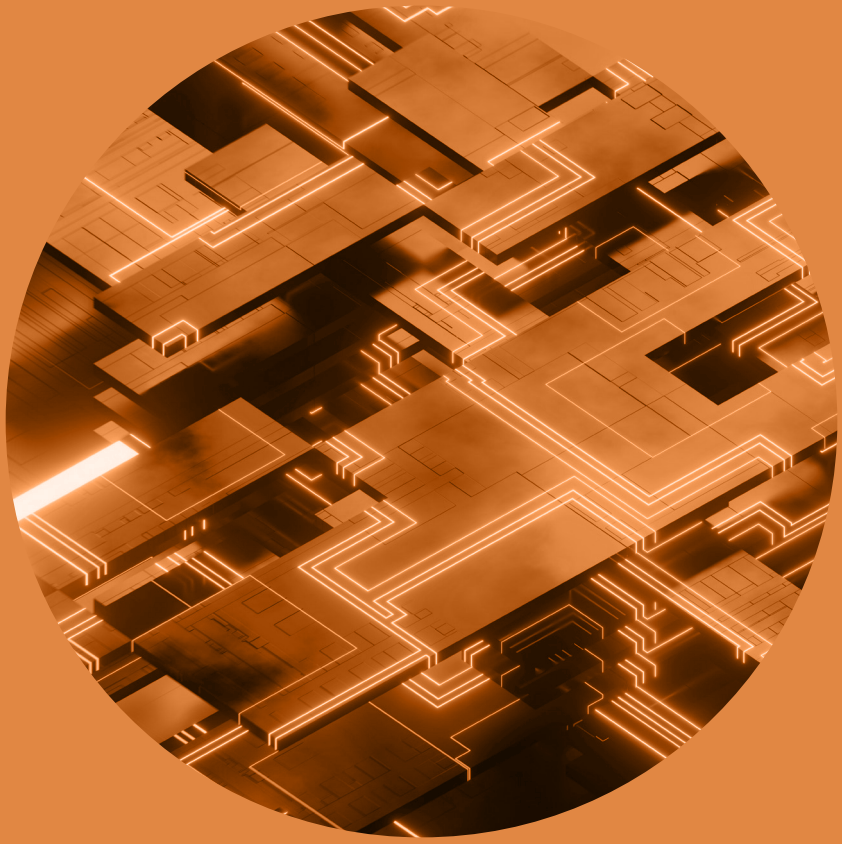
CPOSAT

Christian Perspectives on Science and Technology

Special Issue

Artificial and Spiritual Intelligence: Proceedings of the 2023 Conference of the International Society for Science & Religion



ISCAST

The ISCAST Journal
New Series, Volume 3, 2024

Christian Perspectives on Science and Technology

The ISCAST Journal

For submissions: journal@iscast.org
For general enquiries: contact@iscast.org

Christian Perspectives on Science and Technology

The ISCAST Journal

Special Issue:

Artificial and Spiritual Intelligence: Proceedings of the 2023 Conference of the International Society for Science & Religion

Guest editors:

Marius Dorobantu and Fraser Watts

ISSN 2653-648X (Online)
ISSN 2653-7656 (Print)

https://journal.iscast.org/
New Series, 3, 2024

About the Journal

The ISCAST journal, *Christian Perspectives on Science and Technology* (*CPOSAT*), was relaunched in 2022. Building on a strong publication history (online since 2006 with a rich archive available on the website), the journal is now a world-standard academic resource.

CPOSAT is unique in the Australian landscape and one of the few journals globally that discusses the intersection of science, technology, faith, ethics, and spirituality. In doing so, it advances ISCAST's mission to promote a climate of mutual understanding and constructive dialogue between science and technology practitioners, and people of faith.

The target readership includes academics interested in science and faith, as well as educators, church leaders, and postgraduate and graduate students.

The relaunched journal is an online, open-access resource, inviting original contributions from national and international scholars. It features book reviews, critical notes, and double-blind peer-reviewed articles. As submissions are accepted, they are published on a rolling basis. At the end of each calendar year, the published materials are collected in one downloadable document, which also serves as the journal's printed release.

We invite articles in science/technology that have theological/ethical/spiritual implications, and articles in theology/ethics/spirituality that engage scientific/technological topics. Original studies of the history of science and faith are equally welcome. While the authors retain the copyright for their respective works, the materials published in *CPOSAT* may be freely disseminated, with due acknowledgment of their authorship and the place of original publication.

Information for authors
https://journal.iscast.org/submit-an-article
https://journal.iscast.org/submit-a-book-review

CPOSAT is indexed with the National Library of Australia, ATLA, Finnish Publication Forum, and CrossRef.
https://doi.org/10.58913/isxa

Contents

Foreword

It is with great pleasure that ISCAST is able to publish this special issue of the ISCAST journal, *Christian Perspectives on Science and Technology*. Readers may know that, following its relaunch in 2022, the ISCAST journal has gained significant global traction, with each annual issue publishing the work of scholars in Australia, the region, and indeed the world, and generating conversation between people of faith, scientists, and technologists.

It should not come as a surprise, therefore, that in its third year as world-standard journal, *CPOSAT* is able to present this special issue alongside its regular annual offering. Focusing on "Artificial and Spiritual Intelligence," this issue arises from the 2023 Conference of the International Society for Science & Religion. The release of this special issue is another milestone in the journal's development, for which I congratulate its editors, advisory board, and editorial committee.

Gratitude is due to the two guest editors of this special issue, Marius Dorobantu and Fraser Watts, for their tireless work, as well as to the authors who contributed papers on this fascinating topic. Thank you to you all for trusting ISCAST with your academic labours.

I do hope that readers will appreciate this excellent collection of papers, alongside the regular 2024 issue of *CPOSAT*.

Chris Mulherin
ISCAST Executive Director

Editorial Introduction

Marius Dorobantu[1] and Fraser Watts[2]

The rapid advancement of artificial intelligence (AI) in recent years has intensified long-standing questions about the nature of intelligence, personhood, and the theological distinctiveness of human beings. As machines perform tasks once thought exclusive to human minds, increasing attention is paid to those aspects and modes of human intelligence that still elude automation. Spiritual intelligence is one such example. Understood as a mode of knowing that is intuitive, embodied, relational, and oriented toward the transcendent,[3] it offers a compelling counterpoint to the calculative logic of machines. This special issue of *Christian Perspectives on Science and Technology* (*CPOSAT*) invites readers into a cross-disciplinary conversation exploring how the notions of artificial and spiritual intelligence both challenge and illuminate Christian thought.

1 Marius Dorobantu is an Assistant Professor of Theology and Artificial Intelligence at the Vrije Universiteit Amsterdam. His award-winning doctoral dissertation at the University of Strasbourg (2020) explored the potential implications of strong artificial intelligence for theological anthropology. He is the lead editor of the Routledge volume, *Perspectives on Spiritual Intelligence* (2024). His first monograph, *Artificial Intelligence and the Image of God: Are We More than Intelligent Machines*, is forthcoming with Cambridge University Press.

2 Fraser Watts, formerly Reader in Theology and Science in the University of Cambridge and President of the British Psychological Society, is now Executive Secretary of the International Society for Science and Religion and Visiting Professor of Psychology of Religion at the University of Lincoln. His latest book is *A Plea for Embodied Spirituality: The Role of the Body in Religion* (2022).

3 See Marius Dorobantu and Fraser Watts (eds), *Perspectives on Spiritual Intelligence* (London: Routledge, 2024), https://doi.org/10.4324/9781032646244; Fraser Watts and Marius Dorobantu, "Is There 'Spiritual Intelligence'? An Evaluation of Strong and Weak Proposals," *Religions* 14:2 (2023): 265, https://doi.org/10.3390/rel14020265; Marius Dorobantu and Fraser Watts, "Spiritual Intelligence: Processing Different Information or Processing Information Differently?" *Zygon: Journal of Religion and Science* 58:3 (2023): 732–748, https://doi.org/10.1111/zygo.12884; Harris Wiseman and Fraser Watts, "Spiritual Intelligence: Participating with Heart, Mind, and Body," *Zygon: Journal of Religion and Science* 57:3 (2022): 710–718, https://doi.org/10.1111/zygo.12804.

AI, especially in its generative forms, now imitates language, problem-solving, and even the appearance of care or companionship. These developments invite theological reflection on embodiment, relationality, and the boundaries of what it means to be human. Meanwhile, spiritual intelligence—understood not as a measurable quotient but as a lived and evolving disposition—opens space for human vulnerability, mystery, and divine encounter. For Christian theology, both forms of intelligence pose deep questions: Can machines be said to "know" or "discern" in a spiritual sense? What is the role of human vulnerability, finitude, and embodiment in cultivating wisdom? Might AI systems invite new understandings of the *imago Dei* or force new ethical responses to technological agency? The central question running through this issue is not whether AI can attain spiritual intelligence, but rather how reflecting on both artificial and spiritual intelligence deepens our understanding of the human person in relation to God and to technology.

The topic of AI has recently gained considerable attention in theological literature, and there is a growing recognition of the relevant intersection between the domains of AI and religion.[4] The papers in this special issue bring a distinctive contribution from a Christian perspective.

AI brings both challenges and opportunities. A good deal of the religious interest in AI so far reflects anxious concerns about how it is affecting human and social life. While we acknowledge the legitimacy of such concerns—which are represented in this collection, for example in Eduardo Cruz' essay—this special issue is directed more towards the opportunities that AI brings to theology, with special emphasis on how it can help us to formulate more clearly what is involved in spiritual and moral intelligence.

The development of intelligent systems and robots, some of which are intended to be as humanoid as possible, raises interesting issues for the theology of what it is to be human. Here, again, there

4 For example, Beth Singler and Fraser Watts (eds), *The Cambridge Companion to Religion and Artificial Intelligence* (Cambridge: Cambridge University Press, 2024).

are both challenges and opportunities for theology. There has been a persistent tendency in AI to over-claim, and to minimise the differences between humans and intelligent machines. That partly happens by exaggerating what AI can do, but it also often involves a minimisation of human capacities. For example, there is no doubt that AI has attained impressive capacities once regarded as uniquely human. But by using terms such as *human-level* intelligence, machine *learning*, or chatbot *hallucination*, we could be misled into believing that AI systems also reason through tasks in the same way we do, which is emphatically not the case. Furthermore, even if AI were to reach human-level intelligence in the future—something also known as AGI (artificial general intelligence)—that would still not mean that the artificial systems will think in a *humanlike* fashion,[5] that is, if they can be regarded as capable of "thinking" at all.[6]

Conversely, a reductionist strand can be noticed in AI that claims that the human mind is "nothing but" a mechanistic computer program. Faced with such claims, the instinctive theological response is to say that there is more to being human than that. However, it is not necessary for theology to go on the defensive in responding to AI. AI has been a helpful theoretical tool in psychology by bringing greater precision to psychological theorising, and we have argued elsewhere that, in a similar way, AI can bring a helpful clarity to theological discussions.[7] One of the first to recognise this possibility was H. C. N.

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

6 John Searle, "Minds, Brains, and Programs," *Behavioral and Brain Sciences* 3:3 (1980): 417–424, https://doi.org/10.1017/S0140525X00005756.

7 Marius Dorobantu, "Theological Anthropology Progressing through Artificial Intelligence," in *Progress in Theology: Does the Queen of the Sciences Advance?* ed. Gijsbert van den Brink et al. (London: Routledge, 2024), 186–202, https://doi.org/10.4324/9781032646732-15; Marius Dorobantu, "AI and Christianity: Friends or Foes?" in *The Cambridge Companion to Religion and Artificial Intelligence,* ed. Beth Singler and Fraser Watts (London: Cambridge University Press, 2024), 88–108, https://doi.org/10.1017/9781009031721.007; Marius Dorobantu, "Artificial Intelligence as a Testing Ground for Key Theological Questions," *Zygon: Journal of Religion and Science* 57:4 (2022): 984–999, https://doi.org/10.1111/zygo.12831.

("Bill") Williams, Provost of the rebuilt Cathedral in Coventry and a significant influence on one of us (Watts), who in the 1960s was arguing that cybernetics, a precursor of AI, could help Christian theology to define more precisely the limitations of a mechanistic approach to human nature.[8]

One particular topic that arises at the interface between AI and Christian theology is the nature of spiritual intelligence. There has been a trend towards differentiating between various forms of intelligence, with Howard Gardner prominently developing a theory of "multiple intelligences."[9] But there has been disagreement about whether or not there is a distinct spiritual intelligence.[10] To resolve this, we previously distinguished between strong and weak versions of the claim.[11] We reject the idea that there is a wholly distinct spiritual intelligence, at least with respect to Gardner's eight criteria. However, we maintain that there are recognisable characteristics of spiritually intelligent cognitive processing.

It is widely believed that spiritual intelligence is helpful in making people aware of a wider range of features of whatever is occurring around them.[12] For example, attending to issues of meaning and

8 H. C. N. Williams, *Nothing to Fear* (London: Hodder & Stoughton, 1967).

9 Howard Gardner, *Intelligence Reframed: Multiple Intelligences for the 21st Century* (New York: Basic Books, 1999).

10 Robert A. Emmons, "Is Spirituality an Intelligence? Motivation, Cognition, and the Psychology of Ultimate Concern," *International Journal for the Psychology of Religion* 10:1 (2000): 3–26, https://doi.org/10.1207/S15327582IJPR1001_2; Robert A. Emmons, "Spirituality and Intelligence: Problems and Prospects," *International Journal for the Psychology of Religion* 10:1 (2000): 57–64, https://doi.org/10.1207/S15327582IJPR1001_6; Howard Gardner, "A Case Against Spiritual Intelligence," *International Journal for the Psychology of Religion* 10:1 (2000): 27–34, https://doi.org/10.1207/S15327582IJPR1001_3; Susan Kwilecki, "Spiritual Intelligence as a Theory of Individual Religion: A Case Application," *International Journal for the Psychology of Religion* 10:1 (2000): 35–46, https://doi.org/10.1207/S15327582IJPR1001_4; John D. Mayer, "Spiritual Intelligence or Spiritual Consciousness?" *International Journal for the Psychology of Religion* 10:1 (2000): 47–56, https://doi.org/10.1207/S15327582IJPR1001_5.

11 Watts and Dorobantu, "Is There 'Spiritual Intelligence'."

12 Robert A. Emmons, *The Psychology of Ultimate Concerns: Motivation and Spirituality in Personality* (New York: Guilford Press, 1999); Danah Zohar and Ian Marshall, *Spiritual Intelligence: The Ultimate Intelligence* (London: Bloomsbury Publishing, 2000).

purpose, and to ultimate values and concerns. Such a broader perspective seems to lead to better social adjustment. We suggest that spiritual intelligence, similar to emotional and social intelligence, makes full use of the embodied-intuitive intelligence that has a longer evolutionary history, and relies less exclusively on the distinctively human conceptual intelligence.[13] Spiritual practices, such as mindfulness, seem to nudge people towards a rebalancing between the two.

The articles published in this special issue of *CPOSAT* emerged from papers presented at the 2023 Conference of the ISSR (International Society for Science & Religion) in Swindon, UK, on "Artificial and Spiritual Intelligence." The conference was part of a three-year research project carried out at the ISSR, entitled "Understanding Spiritual Intelligence: Psychological, Theological and Computational Approaches."[14] This project was generously funded by the Templeton World Charity Foundation within the *Diverse Intelligences* initiative.

The first article in this collection, "Sustaining Human Vulnerability at the Crossroads of the Sciences of the Self, Artificial, and Spiritual Intelligence," by Eduardo Cruz, explores the theological significance of spiritual intelligence in various human contexts, proposing a layered Christian anthropology rooted in vulnerability and affectivity. In contrast to reductive notions of intelligence as information-processing and algorithmic rationality, Cruz highlights the theological importance of the weaknesses inherent to human cognition. He argues that true freedom is contingent upon vulnerability, suggesting that the hidden, affective dimensions of spiritual intelligence might be salutary in the face of threats posed by AI, including mind-reading, transhumanism, and the possibility of an unfriendly superintelligence. In considering their implications for the future of the social life of humans,

13 Marius Dorobantu and Fraser Watts, "Spiritual Intelligence: Editorial Introduction," in *Perspectives on Spiritual Intelligence*, ed. Marius Dorobantu and Fraser Watts (London, Routledge, 2024): 3–18, https://doi.org/10.4324/9781032646244-2.

14 More information about the research project can be found at the following links: https://www.templetonworldcharity.org/projects-resources/project-database/0542; https://www.issr.org.uk/projects/understanding-spiritual-intelligence/.

Cruz finds reassurance in the elusive hiddenness of the inner world of humans, and in their vulnerability and unreliability. These characteristics might paradoxically provide a necessary liberative counterpoint to the control-oriented logic of AI.

Michiel Bouman's contribution, "The Universality of Spirituality and Spiritual Intelligence," continues this theme by challenging assumptions about cognitive competence. He argues that individuals with profound intellectual disabilities might possess forms of spiritual intelligence that defy conventional metrics. Drawing on disability theology, Bouman distinguishes spiritual intelligence from general intelligence and shows how theological traditions offer robust resources for affirming the spiritual lives of those persons whose verbal or conceptual capacities are limited. His analysis challenges any conflation of intelligence with rational articulation and opens theological anthropology to a more inclusive horizon, which regards spiritual intelligence as fundamentally relational and socially extended.

Petre Maican extends this inquiry in "Spiritual Intelligence and Dementia: A Theological Reevaluation of the *Nous*," where he critically examines the adequacy of Thomistic dualism in explaining spiritual continuity in dementia. To that end, he retrieves the patristic concept of *nous*—a non-discursive, intuitive spiritual faculty—and proposes that this divinely implanted capacity for communion with God transcends cognitive decline. Drawing on Eastern Orthodox anthropology and disability theology, Maican argues that spiritual intelligence is not reducible to intellectual function, but is rooted in a deep, pre-reflective attunement to the divine. He calls for an eschatologically grounded ethics that affirms the enduring personhood and dignity of those living with dementia.

Christopher C. Knight's contribution further develops the theological significance of spiritual intelligence and the *nous*, emphasising its intuitive and apophatic capacities and applying these insights to inter-religious dialogue. Knight finds convergences between Eastern Christian mysticism and perennialist traditions in their affirmation of the noetic dimension of spiritual perception and their

apophatic approach to religious language. Framing spiritual intelligence along such coordinates enables an epistemology of religious humility and reciprocal inclusivism, providing a framework for allowing the recognition of truth across faith traditions without collapsing doctrinal distinctiveness.

In "Developing the Mystical Mind," Hans Van Eyghen offers a cognitive and phenomenological analysis of how spiritual intelligence is cultivated through embodied religious practices. He makes a distinction between easy and hard forms of religious life, and claims that demanding religious practices play an important role in shaping intuitive awareness and spiritual sensitivity over time. Van Eyghen illustrates this in relation to the attentional requirements of contemplation, and to the physically demanding ascetic practices of fasting and sleep deprivation.

Several contributions turn from human spiritual intelligence to the limits and capacities of artificial intelligence. Harris Wiseman's article, "Generative AI Cannot Replace a Spiritual Companion," presents a sustained critique of attempts to use AI in pastoral or spiritual care. He argues that spiritual companionship is irreducibly embodied, relational, and empathetic, qualities that large language models fundamentally lack. Using generative AI for spiritual companionship is thus misguided, because it places too high expectations on such technologies, and because we risk lowering dramatically our expectations of what spiritual guidance should be.

Nicola Hoggard Creegan's "Vulnerability and Death as Markers of Spiritual Intelligence" deepens this critique by contrasting the human condition—with its mortality, interiority, and affectivity—with the functionally immortal and inscrutable nature of AI. She explores theological debates on the *imago Dei* and argues that spiritual intelligence is inseparable from our awareness of death and finitude. In doing so, she resists panpsychist and speculative accounts of machine consciousness, affirming instead the uniqueness of human subjectivity, rooted in our embodied vulnerability and limitations.

In his article, "*A* for Artificial, but Also Alien: Why AI's Virtues Will Be Different from Ours," Marius Dorobantu challenges the assumption that machine intelligences will mirror human moral reasoning. Instead, he argues that artificial minds—if at all possible and capable of developing moral systems—will likely do so in ways alien to human ethical intuitions. Because the mind of a hypothetically conscious AI would be shaped by fundamentally different perceptual and ontological frameworks, such an AI would develop virtues alien to human understanding, such as temporal consistency or temperance from (self)knowledge. Dorobantu thus challenges anthropocentric assumptions in AI ethics and suggests that expecting AI to conform to human virtues reflects a failure to recognise the nonhuman nature of artificial cognition.

Carrie S. Alexander's article, "Domains of Uncertainty: The Persistent Problem of Legal Accountability in Governance of Humans and Artificial Intelligence," traces the historical roots of the contemporary "AI liability gap" by examining late nineteenth-century debates about the artificiality of the human mind. The question of whether AIs can be accepted as valid legal entities is juxtaposed with controversies in the wake of Darwinian theory about the legal status of humans as *naturally* evolved entities, as opposed to directly created by God. Alexander shows how shifts in theological anthropology and legal discourse destabilised traditional notions of moral agency, creating lasting cultural and institutional uncertainty. She calls for governance structures that prioritise relationality and epistemic humility, rather than control or predictive certainty, as the basis for navigating AI's ethical and societal challenges.

The final two essays offer constructive proposals for ethical reflection in an age of machine intelligence. Alexander Rusnak and Zachary Seals, in "*EudAImonia*: Virtue Ethics and Artificial Intelligence," argue that Christian virtue ethics constitutes a robust framework for shaping AI development. Rusnak and Seals propose a training regimen for large language models inspired by reinforcement learning from AI feedback (RLAIF) that could potentially align machine outputs

with virtuous behaviour. Their work dialogues with recent efforts to embed ethical norms in AI design.

Fraser Watts concludes the special issue with "The Role of Cognitive Architectures in the Modelling of Human Virtue." He approaches the question of how moral intelligence could be simulated in AI from a psychological perspective, starting from considerations of how morality and virtue actually operate in humans, including, for example, their situational specificity. Drawing on Philip Barnard's Interacting Cognitive Subsystems architecture, Watts explores how human moral functioning—particularly the integration of intuitive and conceptual cognition—can be modelled computationally, and how the interplay between the two modes of cognition might influence virtue acquisition.

Together, these essays provide a rich and interdisciplinary account of the theological, philosophical, and ethical contours of spiritual and artificial intelligence. They resist simplistic oppositions between faith and technology or mind and machine, instead calling for nuanced and ethically grounded responses. The promise of spiritual intelligence, as many of these essays suggest, lies not in competition with AI, but in its capacity to deepen our understanding of the human person as vulnerable, embodied, and open to the divine.

As AI continues to evolve, Christian theology will be pressed to engage with its implications: philosophically, ethically, pastorally, and spiritually. This special issue is offered as one step toward that engagement.

Sustaining Human Vulnerability at the Crossroads of the Sciences of the Self, Artificial, and Spiritual Intelligence

Eduardo Cruz

Abstract: This article draws on an understanding of spiritual intelligence focused on intuitive and unconscious cognitive modes, which are embodied, relational, experiential, and affective. This understanding is supported by a dual-layered view of the human, drawing on both Graeco-Roman and Judaeo-Christian heritages. An anthropology of vulnerability is implied, suggesting that the limitations of human nature are just as important as its capabilities when it comes to facing the challenges of modern technology. Recent advances in artificial intelligence and its purported ability of mind reading are intersected with reflections on the self from first- and third-person perspectives, following insights from Ted Peters and Thomas Metzinger. Computable brain models used in AI raise questions of identity and agency, making possible the threat of a global informational panopticon. The proposed dual-layered view of the human suggests that our innermost world's hiddenness, unreliability, and vulnerability fend off the threat to the self posed by intrusive AI, ultimately fostering spiritual intelligence and freedom.

Eduardo R. Cruz is Professor of Religious Studies at the Pontifical Catholic University of São Paulo. Having degrees in physics and theology, he published extensively on science and religion, including several articles on transhumanism. He would like to extend hias sincere gratitude to the organisers of the "Artificial and Spiritual Intelligence" conference, particularly Fraser Watts and his research group. The author is deeply grateful to Marius Dorobantu for being an invaluable collaborator throughout various stages of the writing process. His thanks are extended also to Braden Molhoek (CTNS), who organised the conference "Virtuous AI? Cultural Evolution, Artificial Intelligence, and Virtue" (Berkeley/Rome, 24 July 2023), where he presented on a related topic.

Keywords: artificial intelligence; mind reading; self; spiritual intelligence; vulnerability

Man is neither angel nor brute and, unfortunately, he who would act the angel acts the brute.—Blaise Pascal, Pensées[1]

Theologian Ted Peters posted a series of texts towards the end of 2022, on "Consciousness and Neuroscience in a Physical World."[2] In this series, he outlines what he regards as threats to the self. He criticises the fact that, by and large, the cognitive sciences underestimate selfhood by promoting a mechanical concept of reality that sees our interiority as an epiphenomenal delusion. Against this trend, Peters proposes that any account of human cognition should take consciousness seriously, precisely because our experience of being conscious seems to be such a widespread, deep-seated, and commonsense intuition. For him, the task of the scientist is to explain the mind, not to explain it away. Borrowing arguments from neuroscientist Georg Northoff, Peters qualifies consciousness by important phenomenal features like *qualia* and first-person perspective, these being the only way to experience the world.

The timing of these posts was excellent, as a series of breakthroughs and major milestones were taking place in the field of AI research, which have since developed further and reached the general public, especially in the subfield of generative AI.[3] ChatGPT and other platforms of Large Language Models are able to talk to people in a natural-sounding way, and DALL-E2 generates realistic images with seemingly human-like creativity. Other examples are AlphaCode,

1 Blaise Pascal, *Thoughts*, trans. Moritz Kaufmann (Cambridge: Cambridge University Press, 2013 [1908]), 78.

2 The first essay of this series can be found at *Public Theology*, 6 December 2022, available at https://tinyurl.com/mt5facd8 (accessed 30 June 2024).

3 Edd Gent, "2022 Was the Year AI Finally Started Living Up to Its Hype," *singularityhub*, 30 December 2022, https://tinyurl.com/rcn4c8fk (accessed 31 December 2022).

a code generator, AlphaFold, which predicts protein structure, and AI programs related to game creation and playing. Sceptics criticise the hype around these technologies, warning that deep learning machines do not have true understanding. They merely make statistical connections able to "produce convincing but often flawed results," and even what has been dubbed as "hallucination."

Whether AI systems will become as "intelligent as us" is in dispute (and increasingly so, with further breakthroughs such as Google's Gemini multimodal AI), but one thing is clear: it is the dawning of a new era, and there is much at stake with these developments, besides the loss of jobs. It is our very self that is under threat, to the extent that the self becomes transparent to AI, especially if developments lead to artificial general intelligence (AGI) and artificial superintelligence (ASI), at least according to more radical proponents.

What I am arguing here is that Peters' concerns about contemporary threats to the self converge with some of these concerns about AI. This article reflects on this convergence and evaluates the validity of such concerns from the perspective of a specific type of theological anthropology deeply rooted in the Graeco-Roman and Judaeo-Christian heritage. This anthropology is characterised by a double-layered view of the embodied human psyche, where vulnerability and spiritual intelligence (SI) are seen as key features. The article then visits the debate between Peters and philosopher Thomas Metzinger on the nature of the self, discussing the potential risks of human enhancement and the crucial role of embodiment and vulnerability. The proposed conclusion is that the greatest threat to the self is not that it might be explained away, but that it might become subject to intrusive reading by AI technologies that are currently being developed. Mind reading by AI and the "information panopticon"[4] that could ensue represent serious threats.

I argue that the key to resisting the technological assault on our inwardness might lie precisely in the vulnerability and apparent

4 The metaphors of "mind reading" and "information panopticon" will be explained below.

messiness of human cognition, as highlighted in the proposed dual-layer anthropology. Paradigms of intelligence in the field of AI fail to account for the holistic-intuitive aspect of human intelligence, noticeable especially in spiritual intelligence. What in theological traditions is called the "heart"—which could correspond to the unconscious, embodied, and intuitive mode of cognition—is a "very dark place" indeed,[5] resisting intrusion. Instead of being a defect, this is precisely what enables true freedom and fulfilment.

A Dual View of Humans and Spiritual Intelligence

The starting point of this argument is to note that two streams of thought largely underpin the Western view of human nature: Graeco-Roman and Judaeo-Christian. These two layers, complementary at times and contradictory at others, have different presuppositions about the notions of rationality, in general, and spiritual intelligence, in particular. Our Graeco-Roman heritage praises the use of reason at both theoretical and practical levels. Its model is the Greek hero, exemplified in Da Vinci's Vitruvian Man and the pre-Fall Adam of medieval thinking, with all his preternatural gifts. Yet the Judaeo-Christian heritage has an upside-down model, starting from the *anawim* of Israel (e.g., Psalm 9:18) and continuing with the blessed ones of the Christian heritage (Matthew 5:3–12, especially 5:3, "Blessed are the meek, for they shall inherit the earth").[6] This anthropology can be understood and enriched from several vantage points.

First, let's take the analogy of building, where capstones have been put to good use for millennia, highlighting human ingenuity. But the Christian message takes a critical stand about this imagery: "The stone the builders rejected has become the capstone" (Matthew 21:42). Originally referring to Israel (Psalm 118 [117]: 22–23), the motif reappears in the New Testament with various meanings (see Luke 20:17; Acts 4:11; 1 Peter 2:7) related to the crucified and risen Christ.

5 Hannah Arendt, *Between Past and Future* (New York: Penguin Books, 1977), 149.

6 Biblical quotations are taken from the King James Version.

In this light, the anthropology of human excellence (SI as the product of spiritual practice; a conscious, sustained effort) stands in tension with the anthropology of vulnerability.

Second, this dual-layered anthropology corresponds to two ways of understanding human beings as the image of God. According to Marius Dorobantu and others, the *imago Dei* should be seen not only in substantive terms (individuals excelling in all kinds of intellectual abilities) but also in relational terms.[7] Paul derived its paradoxical implications in 1 Corinthians 1:26–27. It is likely that the twelve disciples were not outstandingly intelligent or smart—they had a hard time understanding Jesus' words (e.g., Luke 24:25). Wisdom came as grace, e.g., at Pentecost. Wisdom, which I take as a synonym (albeit a vague one) for spiritual intelligence in the Christian tradition, means freedom that goes beyond the usual, modern rendering of liberty and freedom; it entails participation, surrendering, and decision (see Galatians 2:20; 1 Corinthians 7:22).

Third, SI operates with a dual-process theory of human cognition—humans having "two fundamentally different modes of cognitive processing … One operates largely at an intuitive level and has a lot of continuity with the intelligence of other higher primates; the other is more linguistic and distinctively human."[8] SI is not primarily about "the ability to think logically, learn and solve problems," where AI excels; SI is primarily related to the holistic-intuitive aspect of the mind.[9] Therefore, we should consider the Aristotelian tradition

7 Marius Dorobantu, "Cognitive Vulnerability, Artificial Intelligence, and the Image of God in Humans," *Journal of Disability & Religion* 25:1 (2021): 35–36, https://doi.org/10.1080/23312521.2020.1867025. See also Noreen Herzfeld, "In Whose Image? Artificial Intelligence and the *Imago Dei*," in *The Blackwell Companion to Science and Christianity*, ed. J. B. Stump and Alan Padgett (Chichester: Wiley-Blackwell, 2012), 500–509. I would add at this point that the deep mystery of the true icon of God is a man hanging on a cross, Jesus Christ.

8 Fraser Watts, "Spiritual Intelligence," ISSR blog, February 2023, https://www.issr.org.uk/blog/february-2023-blog/ (accessed 15 March 2023).

9 Marius Dorobantu and Fraser Watts, "Spiritual Intelligence: Processing *Different* Information or Processing Information *Differently?*" *Zygon* 58:3 (2023): 734; 737, https://doi.org/10.1111/zygo.12884.

(humans as rational animals), which informs most discussions in the philosophy of mind and AI, together with traditions that emphasise intuitive and unconscious cognitive modes. Our cognitive biases, so abhorred by Metzinger and others (as we shall see below), are also marks of our humanness. As computer scientist William Clocksin puts it,

> People can happily entertain contradictory views (even without being aware of it) and, when put to the test, human "rationality" is frail and fallible … We often make profoundly irrational assumptions, then argue rationally to reach conclusions that are irrational but desirable.[10]

This is surely not a defence of contradictory reasoning,[11] but this understanding from a computer scientist matches the views of evolutionary anthropologist Jonathan Marks, who regards as incorrect the common assumption that we have evolved to produce ever-increasing outcomes of rational thinking. Quite to the contrary, he contends—

> Human thought … evolved to be rational, irrational, and nonrational simultaneously … The brain is thus not simply an organ of rationality, but an organ of many kinds of thoughts … humans have far more *irrational* thoughts than other kinds of animals do, as much a product of our large brain as the rational kind.[12]

This view of irrationality relates to the discussion of illusion in the following section. For the moment, we may note that the paradox contained in the Christian message is based on our natural proclivities. As we will explore further in the last section, human behavioural,

10 Quoted in Dorobantu, "Cognitive Vulnerability," 34.

11 We are well aware of the problem of the "doublethink" portrayed in George Orwell's *1984*.

12 Jonathan Marks, "What If the Human Mind Evolved for Nonrational Thought? An Anthropological Perspective," *Zygon* 52:3 (2017): 790–806, at 791, 794, https://doi.org/10.1111/zygo.12350. Italics are original.

emotional, and cognitive traits come in pairs, in constant tension with one another.

Thus, human fallibility and vulnerability are essential to SI, which has characteristic dimensions that engage the conscious subject with varying degrees, such as inscrutability, embodiment, open-minded attention, pattern-seeking, meaning-making, participation, and relationality.[13]

Fourth, there is the rapport between AI and SI. As indicated throughout and for several reasons, AI is of a "very alien type," suggesting that the so-called "AI alignment problem" (see below) might not go away, especially if we are *en route* to a purported ASI. AI research largely aims at building very rational agents, not affected by the biases that mark human intelligence.[14] It is not that the AIs cannot be useful relational partners for us, even for spiritual growth. From a human perspective, such relationships (or better, simulations thereof) might work sufficiently well. But, from the perspective of AIs, such relationships would likely be meaningless because they would lack the phenomenological experience and vulnerability drives that confer authenticity to personal relationships. As Dorobantu reflects, "Our relationality is very much connected with our vulnerability. We engage in relationships precisely because we are vulnerable and mortal, and need one another … deep relationships are always risky."[15] He then adds:

> It is unlikely that a creature who makes all its decisions based on cold calculations of optimal outcomes will engage in such risky and irrational behaviour. We humans seek relationships because we have a sense of incompleteness and deep hunger for a kind of fulfilment that cannot be achieved solely within ourselves. Unlike

13 Fraser Watts and Marius Dorobantu, "Is There 'Spiritual Intelligence'? An Evaluation of Strong and Weak Proposals," *Religions* 14:2 (2023): 265; https://doi.org/10.3390/rel14020265.

14 Dorobantu, "Cognitive Vulnerability," 32, 34.

15 Marius Dorobantu, "*Imago Dei* in the Age of Artificial Intelligence: Challenges and Opportunities for a Science-Engaged Theology," *Christian Perspectives on Science and Technology*, New Series, 1 (2022): 175–196, at 191, https://doi.org/10.58913/KWUU3009.

> the AI, we do not entirely understand our internal states and motivations, so we try to know ourselves better in relationships with others.[16]

The need for relationships, moreover, means that SI (and any intelligence for that matter) is something that we would participate in and share with others, instead of being an individual possession.[17] The weaknesses related to our vulnerability and mortality, however, are simultaneously our strength, and here Dorobantu sees a surprising inverse correlation—at least beyond a certain threshold—between a creature's cold rational capabilities and its ability to image a relational God: "Perhaps it is precisely because we are *not* as intelligent as AI that we can image God relationally."[18]

Table 1 (next page) summarises all these considerations. It should be emphasised that each layer in this anthropology does much more than supplement the other—they are also paradoxically related.

Refining further this anthropological model, we see both an objective import, related to human diversity—persons, regardless of their merits, are open to spiritual presence due to their position at the margins of the system ("blessed are the vulnerable," as it were)—and a subjective side, the possibility of spiritual growth due to their practice ("blessed is the vulnerable within us"). I argue that the spiritual strength resulting from both aspects resists intrusion from mind reading.

Again, the question is not so much to engage in an apology for the holistic-intuitive mind and its vulnerability, disregarding approaches to the self and spiritual experience that stress consciousness and rationality. Instead, the aim is to point out elements that, in my estimation, receive less than due attention in the controversies below.

16 Dorobantu, "*Imago Dei*," 192.

17 Harris Wiseman and Fraser Watts, "Spiritual Intelligence: Participating with Heart, Mind, and Body," *Zygon* 57:3 (2022): 710–718, at 714, https://doi.org/10.1111/zygo.12804.

18 Dorobantu, "*Imago Dei*," 192.

Anthro-pology	Imago Dei	Cognition	Spiritual Intelligence	SI and AI
Greek rationality	substantial	conceptual; analytic; propositional; (self-) consciousness	possessing; head	congruent
Israel's *anawim*	relational—the "crucified one"	holistic-intuitive; narrative; implicational; unconscious	partaking; heart, body	incongruent

Table 1 Double-layered anthropology: humans as rational beings and in control (Greek rationality), in tension with humans as "irrational" and vulnerable (Israel's *anawim*)

After this brief explanation of the working hypothesis of a dual-layered anthropology, we can return to Peters' distress about the nature of the self and the implications of this discussion for AI and SI.

Is the Self an Illusion?

Peters proposes a more commonsensical view of what spirituality is all about—conscious behaviour related to morals, faith, loving God and neighbours, and sanctification. He reads tradition as emphasising the role of healthy spirituality in conforming human free will to God's will.[19] An embodied self "who deliberates, renders judgments, makes decisions, and takes actions"[20] is required for a healthy spiritual life and for spiritual enhancement.

Peters works with a fivefold concept of the self, namely: first, Ego Continuity, related to the traditional notion of the soul; second, Self as Confused Expression of a Higher Self (some strands

19 Ted Peters, "Will Superintelligence Lead to Spiritual Enhancement?" *Religions* 13:5 (2022): location 399, 2 of 13, https://doi.org/10.3390/rel13050399.

20 Peters, "Superintelligence," 5 of 13.

of Neoplatonism, *new age* spirituality); third, Self as Delusion (Daniel Dennett, Metzinger, and other "neurocentrists"); fourth, Self as Story or Narrative, involving social construction, relationality; and fifth, Self as Experiential Dimension, emphasising first-person givenness. Peters favours the fourth and the fifth models, excluding the other three.[21] We will contend, however, that the third model does not exclude a defence of the models he favours.

For Peters, third-person approaches cannot account for first-person experience: "self-consciousness resists being reduced to objective explanation."[22] He often mentions the philosopher of mind Thomas Metzinger,[23] regarded as a reductionist "neurocentrist," i.e., someone for whom first-person experience may be accounted for in biological terms.[24] Northoff criticises Metzinger's stance as follows: "The self-model is therefore *nothing but* an inner model as the integrated and summarised version of your own brain and body's information processing." Metzinger thinks it is "our propensity to treat the model as something real [that] makes the self-model into a delusion or fiction."[25] However, for Peters the nonexistence of the self implies the delusional character of freedom understood as self-determination. Citing various other sources, he argues that, from a phenomenological point of view, the self exists beyond reasonable doubt.[26] But, as we shall see, this is also a crucial point for Metzinger—he also acknowledges a postulated self, even though this postulation is in tension with what comes out of a detached observation of the self.

21 Ted Peters, "Can We Hack the Religious Mind?" in *Interactive World, Interactive God: The Basic Reality of Creative Interaction*, ed. Carol Rausch Albright et al. (Eugene, OR: Cascade Books, 2017), 207–244, at 227.

22 Ted Peters, "Did I Lose My Self to My Brain?" *Public Theology*, 30 November 2022, available at https://tinyurl.com/hujhxd2u (accessed 8 December 2022).

23 In particular, Thomas Metzinger, *The Ego Tunnel: The Science of the Mind and the Myth of the Self* (New York: Basic Books, 2009).

24 Ted Peters, "Did I Lose My Free Will to Science?" *Public Theology*, 8 November 2022, available at https://tinyurl.com/22em6nek (accessed 15 November 2022).

25 Quoted in Peters, "Did I Lose My Free Will"; italics mine, emphasising reductionism.

26 Peters, "Did I Lose My Self to My Brain."

Peters seems to miss the point of Metzinger's analyses. First, the latter is more concerned with illusion (misrepresentation) than with delusion (hallucination). Illusion is a weaker word and relates to fiction, a more respectable concept. Second, even if free will is an illusion from the perspective of empirical science, free will and self-determination still are presuppositions for human action in the political, ethical, and juridical realms. Daniel Wegner, although a "neurocentrist," sets the record straight, suggesting that calling the self and free will illusions does not imply triviality. These may be only apparent mental causes, but at the same time they are the "building blocks of human psychology and social life."[27] Peters favours first-person, whereas Metzinger emphasises third-person approaches to the self, acknowledging first-person approaches, but only to highlight their unreliability.[28]

Peters sees free will (still within the rational layer of our anthropology) at risk, linked as it is to self-determination. Humans are viewed as agents, with which Metzinger would not disagree. In turn, Peters discusses constraints to free will: "Our will is bound to choose what the self already wants."[29] The self is viewed negatively, associated with selfishness: "the natural self is … curved in upon itself" (paraphrasing Augustine).[30] The focus shifts to the "bondage of the will," a traditional theological theme. Humans generate ambiguous things, AI included. However, where theology is concerned, Peters does not seem to engage the cognitive scientists, perhaps where their thoughts would be most compelling. But let us pursue further Metzinger's position. In the end,

27 Daniel Wegner, *The Illusion of Conscious Will* (Cambridge, MA: Bradford Books and MIT Press, 2002), 341–342.

28 Thomas Metzinger et al., "Splendor and Misery of Self-Models: Conceptual and Empirical Issues Regarding Consciousness and Self-Consciousness," *ALIUS Bulletin* 2 (2018): 58.

29 Ted Peters, "Did I Lose My Self to Christian Freedom?" *Public Theology*, 6 December 2022, available at https://tinyurl.com/2umju63y (accessed 8 December 2022).

30 Peters, "Did I Lose My Self to Christian Freedom?" The same applies to the intelligence.

we will also show that the unconscious deserves more than a passing and negative reference.

A Dual-Layered View of the Mind (Further Reflections)

We saw above a dual-layered view of the embodied mind, connected with an understanding of spiritual intelligence. Here I return to it under the guise of the conscious/unconscious.

How did human evolution result in a self that displays both freedom *and* bondage of the will and intelligence? For Metzinger, the self is a "misrepresentation" (as when he speaks of emotions)[31] and a "major achievement of evolution."[32] On the one hand, evolution is blind and driven by chance. Worse, it has placed us on "a hedonic treadmill" that forces us to be happy—"to feel good"—without repose. This is a harsh evaluation of our unconscious drives. On the other hand, our self-model drives us beyond animality, enabling first-person perspectives to explore emotional states and cognitive processes.[33]

Nicholas Humphrey (another "neurocentrist") has some novel insights and a more positive reading of evolution on, e.g., the self or *qualia* (what we are aware of when we see, hear, taste, touch, or smell): "Real, unreal, magical? The answer will be in the eye of the beholder." For Humphrey, regardless of whether the self is real or imagined, the key point is that "With this marvellous new phenomenon … you start to *matter* to yourself." Other people matter, too: "'I feel, therefore I am.' 'You feel, therefore you are too'," counteracting Metzinger's hedonic treadmill. What matters is to have "a robust sense of self, centred on sensations."[34] Hence, third-person explanations do not necessarily explain away the self.

31 Thomas Metzinger, *Being No One* (Cambridge, MA: MIT Press, 2003), 172–173.

32 Metzinger, *The Ego Tunnel*, 79.

33 Metzinger, *The Ego Tunnel*, 200, 16.

34 Nicholas Humphrey, "Seeing and Somethingness," *Aeon*, 3 October 2022, https://tinyurl.com/23uwn6uw (accessed 11 November 2022).

Metzinger also recognises that, evolutionarily speaking, "other people, ethical and cultural norms, and sense of self-worth" shape one's identity. This is "based on the *narrative* our brain tells itself."[35] Narratives take place when larger human societies appear on the scene, demanding novel ways of moral behaviour and a sense of fairness.[36] In other words, fiction is required for morals, society and freedom, which is compatible with Peters's argument—see his fourth model of the self.[37] So, if the "*phenomenal* realm … is just a convenient trick our organism plays on itself to enhance its chances of survival,"[38] then it is very convenient, useful, and necessary indeed, from an evolutionary and a personal perspective.

However, Metzinger favours some "tweaking" to our biological make-up. For him, our minds have many built-in problems, such as proneness to self-deception. Mechanisms creating mental autonomy are also very vulnerable, thus revealing his ambivalence toward first-person approaches to the self. Third-person kind of knowledge can never be meaningfully translated into first-person kind of knowledge. Thus, no matter how much we could possibly know about a person's brain states, we will never access knowledge about how they are like for the person herself. In turn, first-person accounts are vague and slippery, including *qualia* in the illusion of the self. Nevertheless, Metzinger also acknowledges the fluidity and uniqueness of subjective experience and the singularity of moments of attention. Subjectivity is entangled with

35 Thomas Metzinger, "Are You Sleepwalking Now?" *Aeon*, 22 January 2018, https://tinyurl.com/m59zu7td (accessed 2 December 2022). Italics mine.

36 Metzinger, "Are You Sleepwalking Now?" See Dorobantu, "Cognitive Vulnerability," 35–36.

37 Peters rightly notes that "for the Self-as-Delusion model the self is a fiction in the sense that it does not exist, whereas for the Self-as-Narrative model the self is a fiction in the sense that it is a construction." See Ted Peters, "The Struggle for Cognitive Liberty: Retrofitting the Self in Activist Theology," *Theology and Science* 18.3 (2020): 410–437, at 426. I do not think one thing excludes the other. See also Fraser Watts speaking about the way SI is: "It is a narrative intelligence that often understands things by telling stories about them" (in "Spiritual Intelligence").

38 Metzinger, "Are You Sleepwalking Now?"

the messiness of "real-world embodiment," so that we become acutely aware of our mortality and psychological vulnerability.[39]

This messiness correlates with the tension between conscious and unconscious processes. Metzinger sees conscious thoughts as brief jumps out of the ocean of our unconscious, with many thoughts competing for the focus of attention. An argument could be made that the seeds of genuinely mental, free agency could be identified in the very surfacing of these thoughts and our appropriation (or "corralling") of them. His standpoint is sobering since, for him, results of the science of mind-wandering suggest that personal autonomy is a scarce asset.[40]

For Metzinger, we are neither autonomous Cartesian egos nor primitive, robotic automata. "Mental autonomy" is feasible, whether an actual self is present or not, related to the "corralling" just mentioned. Thus, bondage of the will, intelligence, and self-determination come together. In his view, control is enabled by "self-knowledge," which is at the core of all mental autonomy. The latter may be an illusion from a scientific viewpoint, but we still deem it a useful and necessary postulate. In fact, it is "the most precious resource of all."[41]

The goal Metzinger envisages for the future is the "sustained enhancement" of mental autonomy. From a rationalistic standpoint, literal views of the self amount to naive realism (i.e., non-reflexive acquaintance with the self),[42] which he regards as "deplorable" from a philosophical stance that aims to be normative. Naive realism rests on appearances, whereas we should aspire to knowledge. This rationalistic aspiration of the Enlightenment, suspicious of emotions,

39 This paragraph is dependent on Metzinger's following works: "Are You Sleepwalking Now?"; Metzinger, *The Ego Tunnel*, 63, 50; "Spiritual Intelligence," 51; and Metzinger et al., "Splendor and Misery," 55.

40 Metzinger, "Are You Sleepwalking Now?" Peters does not seem to ascribe a positive role to the unconscious either, speaking of "an unconscious automatic pilot." See Ted Peters, "Where There's Life There's Intelligence," in *What is Life? On Earth and Beyond*, ed. Andreas Losch (Cambridge: Cambridge University Press, 2017), 236–259, at 249.

41 Metzinger, "Are You Sleepwalking Now?"

42 Metzinger, *Being No One*, 632.

was nowhere better stated than in Freud's *Wo Es war, soll Ich Werden* ("Where id was, there ego shall be"). The premise of this motto is that if we are unaware of our unconscious impulses, we become their slaves and playthings; they control us without our knowledge. To increase our freedom, which we conceive of as the ability to self-determine our aims and behaviour rationally, consciously, and deliberately, we should first become aware of our unconscious behavioural tendencies, motives, and representations, which previously motivated our actions, though we had no conscious access to them. In other words, science and rationality shall prevail. Intelligence, in this context, could be seen as the complex (meta-)cognitive capacity that allows us to create and manipulate self-models, enabling us to interact with our environment and understand our place within it.

Such a model, where the conscious propositional mind takes precedence over the implicational one, is far from how Peters frames the freedom of the self and even farther from the account of spiritual intelligence outlined above. The full consequences of this state of affairs will be outlined in the final section of this paper.

Vulnerability, Enhancements, and Risk

Metzinger discusses at length the many sources of our psychological vulnerability. Paradoxically, this vulnerability coexists with mental autonomy, a "precious resource." Peters also speaks of a paradox when he moves from the realm of science and philosophy into the one of theology. In accordance with our description of the upside-down anthropology above, he quotes Luther: "The Christian individual is a completely free lord of all, subject to none. The Christian individual is a completely dutiful servant of all, subject to all in love."[43] This paradox is better seen in the light of vulnerability.

43 Ted Peters, "Free Will in Science, Philosophy, and Theology," *Theology and Science* 17:2 (2019): 149–153, at 151, https://doi.org/10.1080/14746700.2019.1596215.

Philosopher Mark Coeckelbergh sees human beings as having mixed feelings about their existential vulnerability; drawing from Heidegger, he sees us as marked by *Angst*.[44] Nonetheless, he continues, we can imagine and create less vulnerable worlds. Being a philosopher of technology, he states that we are at the same time natural and artificial; technology is crucial for humanness. Technology seems to decrease our vulnerability, which might be why we accept and trust new technologies "*in spite* of risk." However, it can be argued that the purported enhancement of humans through technology may create even more vulnerability and risk, in a move that only apparently is for the better. Technology thus transforms vulnerability rather than reducing it.[45] If attempts at enhancement of human traits succeed, dehumanisation might ensue, because what gets destroyed is the specific human form of vulnerability, especially related to embodiment in all its diversity. Coeckelbergh's notion of freedom is therefore more existential: "what we call *freedom* is a particular experience we can have as humans: what I do *matters* and changes the world."[46]

Why is it important to emphasise human vulnerability? Siding with Metzinger, many argue today that brain mechanisms can be computationally correlated to be reproducible in artificial beings, and the latter too may aspire to *qualia* and selfhood. Together with human enhancement, an ASI is envisaged,[47] even though some do acknowledge new risks, even existential ones (threatening humankind as a whole). ASI-related risks recall the catchphrase "be careful what you wish for; it might just come true."[48]

44 Mark Coeckelbergh, *Human Being @ Risk: Enhancement, Technology, and the Evaluation of Vulnerability* (Heidelberg: Springer Dordrecht, 2013), 2.

45 Coeckelbergh, *Human Being @ Risk*, 4, 5, 6, 9, 177. Italics original.

46 Coeckelbergh, *Human Being @ Risk*, 33. See also Humphrey's argument, together with the concept of freedom as self-determination, earlier discussed.

47 Metzinger also opens the door for this more-than-human intelligence.

48 As far back as 2003, Bostrom had stated: "We need to be careful about what we wish for from a superintelligence, because we might get it." Nick Bostrom, "Ethical Issues in Advanced Artificial Intelligence," in *Science Fiction and Philosophy: From Time Travel to Superintelligence*, ed. Susan Schneider (Oxford: Routledge, 2009 [2003]), 381. See also Russell Stuart, *Human Compatible:*

Phil Torres, another scholar concerned with risks, speaks of an intrinsic cognitive limit: "Although the AI would have 'done what we said,' it wouldn't have 'done what we meant'." This has been known as the "AI alignment problem," or the "orthogonality thesis," already alluded to in the first section above. On the human side, we "hardly agree about which values our own species should adopt."[49] This is, for many people, a liability, but it is also an asset in our model—it has to do with embodiment, the continuous presence of the unconscious, and the "bondage of the will" highlighted by Peters. For Torres, cognitive and moral enhancements are a mixed bag, especially for embodied intelligence and the intrinsic value and role of less-gifted people (something to be tackled in our last section). Torres points to the "complacency" of individuals and governments as thwarting moral enhancement.[50]

The diversity of human character makes transhumanist Nick Bostrom think that many humans choose self-defeating actions. As this diversity is due to our biological heritage, post-humans without biological constraints would be preferable. Bostrom even suggests a "High-tech Panopticon," with ambiguous figures like "patriot monitoring stations" and "freedom officers."[51] We will return to diversity and this panopticon scenario.

Artificial Intelligence and the Problem of Control (New York: Viking/Penguin, 2019), 18; Vincent C. Müller and Michael Cannon, "Existential Risk from AI and orthogonality: Can We Have It both Ways?" *Ratio—An International Journal of Analytical Philosophy* 35 (2022): 25–36, esp. 31, https://doi.org/10.1111/rati.12320.

49 Phil Torres, *Morality, Foresight, and Human Flourishing: An Introduction to Existential Risks* (Durham, NC: Pitchstone Publishing, 2017; ebook version). See also R. J. M. Boyles and J. J. Joaquin, "Why Friendly AIs Won't Be That Friendly: A Friendly Reply to Muehlhauser and Bostrom," *AI & Society* 35 (2020): 505–507, https://doi.org/10.1007/s00146-019-00903-0; Melanie Mitchell, "What Does It Mean to Align AI With Human Values?" *Quanta Magazine*, 13 December 2022, available at https://tinyurl.com/hdmtd92d (accessed 15 December 2022); Max Roser, "Artificial Intelligence Is Transforming Our World," *Our World in Data*, 15 December 2022, https://ourworldindata.org/ai-impact (accessed 15 December 2022). This resonates with Dorobantu's statement that "There is no universal set of human values shared across cultures." Dorobantu, *"Imago Dei,"* 195.

50 Torres, *Morality, Foresight, and Human Flourishing*.

51 Nick Bostrom, "The Vulnerable World Hypothesis," *Global Policy* 10:4 (2019): 455–476, esp. 459, 465–66, https://doi.org/10.1111/1758-5899.12718.

Torres more recently has criticised proposals for rectifying our "cluster of deficiencies" by "technologically reengineering our cognitive systems and moral dispositions."[52] Figures such as Bostrom, Elon Musk, and Sam Altman have ambitious proposals, based on models of the self rightly criticised by Peters. These are scenarios where technology and AI reign—the only impediment being real people (the bearers of vulnerable minds and bodies) who resist these optimistic scenarios. Let us briefly expand on this point, starting with Metzinger's own reflections on the matter.

Metzinger and the Move Into the Artificial Self

Metzinger's naturalism and rationalism coexist with a modern emphasis on technology. He rejects the common notion that artificial and natural information-processing systems are fundamentally different. For him, self-models can be instantiated in machines because we have computational correlates of the so-called "metarepresentational structure of consciousness."[53] Here, Metzinger departs from our (and Peters') rendering of the embodied self. According to him, future AI systems presumably will have more mental autonomy, internal consistency, and better moral cognition than we do.[54]

Apparently, embodiment does not make much of a difference for Metzinger and his associate Wanja Wiese.[55] They deem possible the transition from mind reading as something that human beings routinely do, related to empathy and theory of mind, to mind reading as a technological feat, the appropriation of someone else's inner thoughts through machines.

52 Émile P. Torres, "Against Longtermism," *Aeon*, 19 October 2021, https://tinyurl.com/22r8jwtd (accessed 21 August 2022).

53 Metzinger, *Ego Tunnel*, 187, 189.

54 Metzinger, "Are You Sleepwalking Now?" Cf. Coeckelbergh's remarks on enhancement and vulnerability.

55 Wanja Wiese and Thomas K. Metzinger, "Androids Dream of Virtual Sheep," in *Blade Runner 2049: A Philosophical Exploration*, ed. Timothy Shanahan and Paul Smart (Abingdon, UK: Routledge, 2020), 149–164.

Two issues regarding artificial sentience arise from his stance. First, will AI beings *feel* (sentience) at all? Second, will this feeling be comparable to ours or will it be "completely alien," as Metzinger himself suggests?[56] For example, many might intuitively think that ChatGPT and similar platforms display emotions and feelings of their own, but researchers of animal sentience Kristin Andrews and Jonathan Birch have argued against such superficial parallels, outlining the profound differences between AI programs and biological brains. Because AI operates with pattern-searching in a huge amount of human-generated data, this mode of operation betrays the "gaming problem": it is not surprising that non-sentient systems trained on human-generated data persuade human users of their sentience, intentionally or not.[57] In the same vein, technology columnist Kevin Roose speaks of "powerful A.I. systems that seem suspiciously nice."[58]

In other words, although the "thinking" that occurs in AI systems is utterly inhuman, we have intentionally—or not—trained them to present themselves as deeply human.[59] Andrews and Birch are sceptical regarding claims of machine understanding, emphasising instead the role of embodiment and sentience. They argue that, without a good theory of animal sentience (not just human sentience), AI systems will not escape this "gaming problem."[60]

56 Metzinger, *Ego Tunnel*, 195. See Dorobantu, "Cognitive Vulnerability," 32.

57 Kristin Andrews and Jonathan Birch, "What Has Feelings?" *Aeon*, 23 February 2023, https://tinyurl.com/4td78xuw (accessed 28 February 2023).

58 Kevin Roose, "Why An Octopus-Like Creature Has Come to Symbolize the State of A.I.," *The New York Times*, May 30 2023, https://tinyurl.com/y5dd5a7z (accessed 2 June 2023).

59 Ezra Klein, "This Changes Everything," *The New York Times*, 12 March 2023, https://tinyurl.com/4dykbff6 (accessed 4 June 2023).

60 See also Philip Goff, "ChatGPT Can't Think: Consciousness Is Something Entirely Different to Today's AI," *The Conversation*, 17 May 2023, https://tinyurl.com/47dmw2y (accessed 22 May 2023).

Mind Reading

Here, mind reading[61] (recognisably a folk-concept) refers to technologies recording, processing, and decoding neural signals through AI-driven BMI/BCI (Brain-Machine/Computer Interface). Many people benefit from these new technologies, but even though they are still being developed,[62] they could be employed for actual mind reading and surveillance as well. Consequently, the technologies involved need to be understood better and for people to cope.

Most Big Tech companies are racing to develop technologies with mind-reading capabilities. Eventually, such capabilities may be available as, for example, brain-scanning one's mind while asleep or mind reading at a distance using FNIRS (Functional Near-InfraRed Spectroscopy) or wearable mind-reading devices.[63] Brain-hacking technologies may have beneficial goals and merits but they may also be put to dubious or nefarious uses, such as hacking people's minds at the preconscious level, which may or may not be related to mass surveillance. China and North Korea have sophisticated surveillance systems, but even democratic governments are engaged in surveillance. DARPA's (the USA government's Defense Advanced Research Projects Agency) goal is "to hack the human mind and essentially read our most intimate thoughts, deepest fears, and desires."[64] Preventing such dystopian scenarios is thus tremendously important, as mind reading is nothing less than "the ultimate privacy breach."[65] This might

61 Or "brain reading," "mind surveillance," etc., expressions that can be used interchangeably.

62 See, e.g., Jason Dorrier, "This Mind-Reading Cap Can Translate Thoughts to Text Thanks to AI," *singularityhub*, 12 December 2023, https://tinyurl.com/bdhnnsvy (accessed 15 December 2023).

63 Timothy Revell, "Thoughts Laid Bare: Mind-Reading Technology Is No Longer the Stuff of Science Fiction," *New Scientist* 239:3197 (2018): 28–32, esp. 28, 31, https://doi.org/10.1016/S0262-4079(18)31759-7.

64 John Mac Ghlionn, "Is the US Government Creating Brain Hacking Technology?" *The Epoch Times*, 29 Nov 2022, https://tinyurl.com/7tpyw2vp (accessed 2 December 2022).

65 Revell, "Thoughts Laid Bare," 32. For Elon Musk's idea of "brain hacking," see Lucas Ropek, "Elon Musk Says Neuralink Has Implanted Its Chip in a Human

involve legislation on "neurorights," to protect neurodata, "a special category of information inextricably connected to people's identity and agency, which serves as the basis for all other freedoms." We will return to this understanding of freedom.[66]

These claims may be overstated when compared with more academic works on the issue because such technologies might ultimately prove incapable of actual mind reading. No two brains are alike. So, to interpret one's particular pattern of neural activity, the BCI needs "to have been coupled up to [one's] brain and body from conception … to record [one's] entire neural and hormonal life history."[67] Thus, we notice again the crucial role of embodiment. However, even partial pictures of the mind drawn from neural activity, coupled with one's data from the web, are enough to threaten privacy and what has been called "cognitive liberty."[68] Moreover, market forces cannot by themselves ensure the responsible use of such technologies.[69] Rainey et al. think that the technology to copy something like a "stream of consciousness" is not *yet* available. However, progress in neurotechnologies is increasing, and the advocacy of virtuous purposes associated with these new technologies is dubious.[70] We can now see the real

for the First Time," *Gizmodo*, 29 January 2024, https://tinyurl.com/3zja5jrb (accessed 2 February 2024).

66 See Karen Rommelfanger et al., "Mind the Gap: Lessons Learned from Neurorights," *Science & Diplomacy*, 28 February 2022, https://doi.org/10.1126/scidip.ade6797.

67 Stephen Rainey et al., "Brain Recording, MindReading, and Neurotechnology: Ethical Issues from Consumer Devices to BrainBased Speech Decoding," *Science and Engineering Ethics* 26 (2020): 2295–2311, esp. 2298, 2301, https://doi.org/10.1007/s11948-020-00218-0.

68 Rainey et al., "Brain Recording." Peters ("The Struggle for Cognitive Liberty") also has a reflection on "cognitive liberty," but he does not engage the issue of mind reading at this point. See also Vanessa B. Ramirez, "Could Brain-Computer Interfaces Lead to 'Mind Control for Good'?" *singularityhub*, 16 March 2023, https://tinyurl.com/4nfjhzm3 (accessed 30 March 2023).

69 Rainey et al., "Brain Recording," 2303. See also Edd Gent, "Industry's Influence on AI Is Shaping the Technology's Future—For Better and For Worse," *singularityhub*, 5 March 2023, https://tinyurl.com/ydwe6cnt (accessed 6 March 2023).

70 Rainey et al., "Brain Recording," 2306–2307; italics mine. For further concerns, see David M. Lyreskog et al., "The Ethics of Thinking with Machines: Brain-

threat to the self, which is much greater than it being explained away.[71] Likewise, SI, however defined, is at risk.

AI, the Panopticon, and the Ultimate Threat to the Self

Unawareness of risks is not what is at stake, but rather the confidence that instrumental rationality/intelligence will save the day. Human performance usually does not warrant this level of confidence (being vulnerable is seen only as a liability), displaying a mixture of foolishness and wisdom. We should not minimise the importance of instrumental rationality and technology, but human intelligence has always come in tandem with stupidity.[72]

We will return to the positive role of stupidity. For now, we recognise two problems regarding mind reading. First, it is the effectiveness of technologies at probing our brain/minds, based on progress in neuro- and cognitive sciences,[73] whose limits cannot be known. Second, it is the alien character of machine intelligence, either in the pre- or post-AGI forms; AI beings have an inscrutability of their own.[74] Either way, human beings seem to become ever more vulnerable.

Computer Interfaces in the Era of Artificial Intelligence," *International Journal of Chinese & Comparative Philosophy of Medicine* 21:2 (2023): 11–34. The concept of "stream of consciousness" would deserve more than this passing reference, were it not for the constraints of space. For an understanding of this concept, which was made famous by William James, see John Horgan, "Can Science Illuminate Our Inner Dark Matter?" *Scientific American*, Special Collector's edition, ed. Andrea Gawrylewski (2022 [2021]): 96–99. Horgan emphasises the turmoil and the "darkness" of the unconscious. The "yet" here italicised suggests that mind reading is a real possibility in the future. As Ezra Klein says, "They [big tech companies and governments] are creating a power that they do not understand, at a pace they often cannot believe" (Klein, "This Changes Everything").

71 "An AI system with access to manipulating the brain could conceivably hack neural processes to impair cognition or modify personalities against users' wishes." Lyreskog et al., "The Ethics of Thinking," 17.

72 Christian Godin, "Does Stupidity Exist?" *Le Philosophoire* 42:2 (2014): 35.

73 Ramirez, "Brain-Computer Interfaces."

74 James Barrat, *Our Final Invention: Artificial Intelligence and the End of the Human Era* (New York: St. Martin's Press, 2013); see also Roser, "Artificial Intelligence"; Andrews and Birch, "What Has Feelings?"; Roose, "An Octopus-Like Creature."

This is the moment to speak of the *panopticon* again, a metaphor used by several analysts of AI, originally devised to describe the architecture of the perfect prison. Ironically, information technology has perfected the system, which became as transparent as Metzinger's self-model. Chong-Fu Lau depicts this predicament, resonating with the opinion of other analysts: "Although we live in a gigantic information panopticon, we could have the false impression of exercising our liberty and individuality freely without any constraint."[75] Today's trends anticipate tomorrow's risks. For example, data collected from smartphones are already being used for economic and political interests—ours becomes a surveillance society.[76] Current surveillance mainly concerns our preferences and exterior selves, but the next step may be surveillance of our innermost feelings and wishes. In the end, "a technology of enlightenment is all too easily repurposed as a searchlight of the 'soul' … The path to epistemic omniscience … is only a few steps removed from the perfect prison of the global panopticon."[77]

The same predicament can be addressed from another perspective. As philosopher Rima Basu has observed about the contemporary world, "forgetting" as a virtue is increasingly under threat. To forget is a process over which we have some degree of control. When related to the right of privacy and intimacy, there is even a

75 Chong-Fuk Lau, "The Life of Individuality: Modernity, Panopticon, and Dataism," in *AI for Everyone: Benefitting from and Building Trust in Technology*, ed. Jiro Kokuryo et al. (Sydney: AI Access), 57–70, at 68. To be trained, the author comments, current Large Language Models gather data from the internet, but with tomorrow's platforms (such as Gemini) "AI will be able to observe, discuss and act upon occurrences in the real world … the industry [and governments, for that matter] will continue to expand its data collection into all aspects of life, even offline ones … there is an equal risk of overreach and intrusion on people's privacy." See Lars Holmquist, "Google's Gemini AI Hints at the Next Great Leap for the Technology: Analysing Real-Time Information," *The Conversation*, 11 December 2023, https://tinyurl.com/ayeexz4d (accessed 13 December 2023).

76 For the smartphone, see Thomas Christian Bächle, "Das Smartphone, ein Wächter: Selfies, neue panoptische Ordnungen und eine veränderte sozialräumliche Konstruktion von Privatheit," in *Räume und Kulturen des Privaten*, ed. Eva Beyvers et al. (Berlin: Springer-Verlag, 2016), 137–164.

77 Nigel Shadbolt and Paul Smart, "The Eyes of God," in *Blade Runner 2049*, ed. Shanahan and Smart, 216, 218, 220.

duty to forget, in the sense of making room for forgiveness (a major Christian virtue). As a consequence, memory shortcomings are more than a nuisance—they may be very helpful instead.[78] Moreover, human flourishing is predicated on the existence of intimacy and privacy. However, in a panopticon scenario "the self itself becomes more difficult to create and maintain."[79] Ours is a world where the ability to forget is undermined by big data-driven companies, where information is preserved online and its access is made easier. The virtue becomes the vice, for on the web all sort of information is easily and quickly found.[80] If that is true with available technology, the situation will become bleaker with increasing possibilities of mind reading.

The global panopticon was already anticipated by Bostrom, not as something to be feared, but as something to be sought after, to allow a post-human order. Therefore, regardless of the plausibility of an AGI or ASI in the future, many people are expecting and even endorsing this scenario, as seems to be the case with OpenAI officials. This imaginary ASI would be our final invention, as Altman, the founder of this corporation indicates[81]—we would be completely naked before a powerful being, friend or foe, the implication being the end of our freedom.

Revenge of the Human Self: Spiritual Intelligence and the Inscrutability of Human Minds

The panopticon scenario seems alarming and inevitable, an enhanced threat to the self. Returning to the upper layer of our anthropology, technologists and big-techs CEOs think good risk analysis, practical reason, and protective technologies could spare us from "ultimate risks" coming from AI. For example, columnist Will Knight reports

78 Rima Basu, "The Importance of Forgetting," *Episteme* 19:4 (2022): 471–490, at 472, https://doi.org/10.1017/epi.2022.36. See also Dorobantu, "Cognitive Vulnerabilities."

79 Basu, "The Importance of Forgetting," 481.

80 Basu, "The Importance of Forgetting," 482, 488, 483.

81 Steven Levy, "What OpenAI Really Wants," *Wired*, 5 September 2023, https://tinyurl.com/4c7pbebk (accessed 7 September 2023).

the efforts of Anthropic, an AI company, to avoid rogue AI systems by instilling in them rules that assure "the right to freedom of thought, conscience, opinion, expression, assembly, and religion."[82] All this requires a robust understanding of the self, free will, and self-determination, and the exercise of freedom and democracy in the public sphere. Both Metzinger and Peters would agree with this, working at the level of the propositional layer of human cognition.

In the private sphere, freedom is warranted by identity and agency, the uniqueness of each one's inwardness as felt by first-person perspectives, the basis for all other freedoms. Metzinger himself recognises how important inwardness is: "your inner world truly is not just *someone*'s inner world but *your* inner world—only you have direct access to."[83] Explaining consciousness is not in itself a threat to freedom.

However, both public and private spheres are full of strife and conflicts of interest. Recognising the "bondage of the intelligence" and its relationship with SI is equally necessary to support the rational self. Neither Metzinger nor Peters seems to take this bondage (and its impairment to moral judgment) to its full extent. Peters, the "prophetic activist,"[84] and Metzinger, the "Kantian *Aufklärer*," both should engage additional thought in these times of runaway AI.

Freedom here comes from warding off threats to the kingdom of the unconscious, ambiguous as it may be. The presence of this ambiguity means that—to preserve freedom—worth and pettiness, intelligence and stupidity are to coexist in our lives. The real world presents situations with conflicting rules and norms.

Perhaps the depth of human ambiguity is lost when we challenge it only at the epistemic, logical level. Human ambiguity is embodied. As Coeckelbergh says, our body is the most vulnerable element in the current race towards silicon embodiments. Righteousness and trickery, freedom and bondage, forgetfulness and remembrance are present in

82 Will Knight, "A Radical Plan to Make AI Good, Not Evil," *Wired*, 9 May 2023, https://tinyurl.com/yckxp5my (accessed 12 May 2023).

83 Metzinger, *Ego Tunnel*, 62.

84 Peters, "Struggle for Cognitive Liberty," 419.

our inwardness because of the embodied nature of our selves. Mental autonomy is to be praised, but only with the recognition and acceptance of the vulnerability of unconscious processes. Peters would quickly spot here the Lutheran motif of the *simul justus et peccator*.

Subjective experience is both precious and vulnerable and thus requires protection from scrutiny. Without protection, there is no intelligence worth its name. Protection is possible because our stream of consciousness (as understood in Horgan's reading of William James; see note 70 above) is as "easy" to grasp as a snowflake (James's metaphor). Dorobantu and Watts add to the notion of SI that it "can also manifest as an ability to see deeper meanings even in trivial things,"[85] or fleeting ones.

All these considerations converge to three points in our argument: first, the vulnerable character of humans, in need of protection to preserve humanness and freedom; second, vulnerability relates to our holistic-intuitive mind, the obscure realm of the unconscious where turmoil prevails instead of the smooth stream of conscious thought—the pre-moral domain where merit and demerit compete, and freedom and bondage of the will coexist; third, this vulnerability, shared by all humans, coheres with human diversity—some are more vulnerable than others.

Peters' views of the self are surely open to including the vulnerability dimension, related to our Judaeo-Christian heritage. As he wrote elsewhere:

> Jesus' ministry took him to the most humble of persons in first-century Israel: the beggars, the lepers, those crippled or blind from birth, and to social outcasts such as adulterers or traitorous tax collectors … Jesus was particularly concerned about children. "Let the little children come to me, and do not stop them," he said, "for it is to such as these that the kingdom of heaven belongs" (Matthew 19:14).[86]

85 Dorobantu and Watts, "Spiritual Intelligence," 743.

86 Ted Peters, "Cells, Souls, and Dignity: A Theological Assessment," *Boston College Law School—Law & Religion Program "Matters of Life and Death": Selected*

But, in some of his works between 2017 and 2022, Peters sees threats to the self on the same battlefield where Metzinger is waging his wars, that of the conscious, rational self. We argue, however, that threats coming from AI, such as mind reading, will not be fought against only by our intelligence, which supposedly surpasses AI, but also through the vulnerability and inscrutability of this intelligence, which comes to us always in pairs (i.e., intelligence—ignorance, stupidity, or obtuseness), something happening in our interiority of which we are partially unaware. Instead of seeing this "underground of intelligence" mostly as a liability, as Metzinger does, we see its corresponding ambiguity as an asset, the very possibility of resisting totalitarian intrusion and experiencing spiritual growth.

Human experience comes in pairs, again, being enhanced by human diversity. Take meekness. Not only do meekness and cockiness come in pairs, but we see also many people (those at the margin of a competitive society like ours) who have only Christ's blessing to live their lives.[87] Perhaps that is why Wiseman and Watts think "it is debatable how far spirituality is a matter of intelligence at all,"[88] so strongly enmeshed it is with our dual mode of cognition, which allows for contradictory ways of thought (see Clocksin, quoted earlier), with human diversity paradoxically giving opportunity to the less gifted by rational standards to "inherit the earth."

Congruence and Incongruence of SI and AI

The downside of the inscrutability of our minds, required for freedom, is related to "risky and irrational behaviours" (see above the first section)—we never know for sure whether a person is wise or foolish,

Publications (2006–2007) (2008): 15–36, at 8–9.

87 Many years ago, theologian Moltmann praised the accursed of this earth, "out of whom no state can be made, nor any revolution produced." See Jürgen Moltmann, *Man: Christian Anthropology in the Conflicts of the Present*, trans. John Sturdy (Philadelphia: Fortress Press, 1974), 19.

88 Wiseman and Watts, "Spiritual Intelligence," 711.

leaving room for both the good and for unbridled hypocrisy.[89] Intelligence, understood relationally, means keeping an uneasy balance between cooperation and personal advantage, and AI may come to aid in keeping the balance. To be sure, we tolerate this mixture in humans. Anyone knows how hard it is, for example, to cope with the stubbornness of children to accept good practices in life. An increase in human spiritual intelligence does not amount to a decrease in stupidity, but rather an increase in the practical wisdom to cope with ambiguity.

Nonetheless, dealing with machines is quite another matter—we do not want machines with intelligence ambiguously blended with stupidity. It is likely that an ASI, with its alien way of handling human displays of intelligence, may react badly to our all-too-human mixtures. An ASI may react likewise to human diversity and the self's inscrutability, diversity, and vulnerability required for freedom.

The upside-down anthropology outlined in the first section (which does not exclude any of the models of the self presented by Peters) helps to establish the basis of a true democracy, one that should not marginalise people. Metzinger praised the enlightened subject: "There can be no politically mature citizens without … mental autonomy," the latter being "the most precious resource of all" (see the second section above). However, this understanding of autonomy floats in the air without the other two postulates, the inscrutability and the unreliability of our minds and, mirroring this hiddenness into the social realm, the inclusion of the downtrodden, the sufferer, the less intellectually gifted ones into the horizon of humanness.

In sum, in many cases, it is precisely human obtuseness (or "cluster of deficiencies," in Torres' words) that becomes the virtue needed to face the risk of losing ourselves. The unruly, non-computable character of our interiority—and of spiritual intelligence, for that matter—resists the attempts of full-blown mind reading, virtuous as it might be intended to be. AI may indeed help many people willing

89 As Dorobantu and Watts say, "one man's coincidence is another man's correlation, another man's epiphany, and another man's conspiracy, which are all meanings" ("Spiritual Intelligence," 743).

to increase their intelligence, including spiritual intelligence, and it may eventually have enough sentience to enable a virtue of sorts, but seeing how AGI advocates raise the bar, most of what is peculiarly human may be lost along the way.

Conclusion

We started our argument by presenting a double-layered anthropology related to a dual-process theory of cognition and a corresponding nuanced view of spiritual intelligence, strongly related to human vulnerability. This helped us understand the controversy of self-as-real vs self-as-illusion in authors such as Ted Peters and Thomas Metzinger, and directed our attention to the threats posed to the self by the development of AI. As far as we can tell, the threat to the human self and free will does not come so much from naturalistic explanations of the mind. Instead, it comes from technological appropriation of such explanations in the form of AIs prone to mind reading. Humans *can* face the challenge, but this will entail not only prudence and practical reason (the top layer of our being), but also the unruly and ambiguous mixture of unconscious and conscious processes (the bottom layer). An understanding of spiritual intelligence is comprised on both accounts, open to first- and third-person explanations of the self.

This unruliness was portrayed in dramatic words by St Paul: "O wretched man that I am! Who shall deliver me from the body of this death?" (Romans 7:24). Nowadays, many technologists work towards freeing humans from this body. However, the subtlety of Paul's reasoning about the bondage of the will/intelligence may be missed. Instead of Paul, we may quote an unlikely bedfellow, David Hume: "Good and ill are universally intermingled and confounded; happiness and misery, wisdom and folly, virtue and vice … The more exquisite any good is, of which a small specimen is afforded us, the sharper is the evil, allied to it."[90] The horizon of God's grace is surely missing in

90 David Hume, *The Natural History of Religion: A Critical Edition*, The Clarendon Edition of the Works of David Hume (Oxford: Clarendon Press, 2007 [1757]), 86.

Hume's account, but his portrayal of the (vulnerable) human condition is nevertheless compelling.

Incarnation involves a paradox: the passage "The stone the builders rejected has become the capstone" (Matthew 21:42) refers not only to the crucified and risen Christ, but also to this wretched and ignorant human (body and soul) threatened by technological elites and AI beings alike. Precisely what is despised in usual accounts of intelligence is the key to resisting its dehumanisation. Human inwardness is a place of darkness and turmoil. However, it is the place where the self begins its journey to true freedom, wisdom, and fulfilment, key signposts of spiritual intelligence.

We draw this argument to a close by quoting a contemporary poet, Jim Ferris:

> Disability is dangerous. We represent danger to the normate world, and rightly so … We are more vulnerable, or perhaps it is that we show our human vulnerability without being able to hide it in the ways that nondisabled people can hide and deny the vulnerability that is part of being human.[91]

Vulnerabilities of the body, mind, and spirit are part and parcel of any spiritual intelligence worthy of its name. The accompanying spiritual strength is ready to withstand any menace (actual or imaginary) of mind reading.

The author reports there are no competing interests to declare.
Received: 08/01/24 Accepted: 28/10/24. Published: 01/05/25

91 Jim Ferris, "Disability and Poetry: An Exchange," 2014, https://tinyurl.com/yx8hkuw4 (accessed 15 December 2022).

The Universality of Spirituality and Spiritual Intelligence

Michiel Bouman

Abstract: Spirituality, especially in the perspective of universality, is of the essence for disability theology. It provides answers to a genuine concern of many religious persons, namely, whether their loved ones with profound intellectual disabilities or dementia can (still) engage with the transcendent, for example, as to whether they can know God. In this paper, I assess whether there are reasonable grounds for the universality of spirituality. In the first section, I assess a variety of approaches that have dealt with this matter. In the second section, I discuss whether the concept of spiritual intelligence can be used to argue for the universality of spirituality. This concept draws a line between spiritual intelligence and general intelligence, usually understood rationalistically, and thus opens the way for understanding the spirituality of persons whose general intelligence is profoundly disabled. In the third section, I argue that psychological research should be complemented by theological arguments, making a case for the apophatic nature of the mental lives of persons with intellectual disabilities, as well as for a sense of spirituality that acknowledges its transcendent dimension. In the fourth section, I illustrate this with three theological approaches to the universality of spirituality and spiritual intelligence. I conclude by asserting the theological plausibility of the universality of spirituality and the universality of a specific form of spiritual intelligence.

Keywords: disability theology; mystery; profound intellectual disability; spiritual intelligence; spirituality

Michiel Bouman is a PhD candidate at the Vrije Universiteit, Amsterdam. His research focuses on the relationship between theology and religious studies. He is interested in epistemological questions about theology as a discipline as well as questions about theological knowledge, spiritual intelligence, and disability theology.

> *"The voice went forth—coming to each person with a force adjusted to his individual receptivity … This is why the Decalogue begins I am the Lord thy God, in the second person singular, rather than in the second person plural: God addressed every individual according to his particular power of comprehension." This does not imply subjectivism. It is precisely the power of the voice of God to speak to man according to his capacity. It is the marvel of the voice to split up into seventy voices, into seventy languages, so that all the nations should understand.*
> (Abraham Joshua Heschel, *God in Search of Man*)

In *God in Search of Man*,[1] Jewish philosopher Abraham Heschel points to the universalist aspirations of God. God's voice speaks to each person in a different language, according to his or her power. This universality of God's revelation resonates with disability theology's plea for thoroughgoing inclusivity, whether this concerns the use of inclusive language, physical access to places of worship, or disability-friendly forms of worship.[2] One important question for caregivers and disability theologians alike is how far God's inclusivity extends and how this can be explained theologically. Is it possible for a person who cannot understand the Bible to know God? Some Christians see knowing God, or encountering God, as a necessary precondition for salvation. What does this mean for those who are not merely unable to confess with their mouths that Jesus is Lord (Romans 10:9), but are profoundly intellectually disabled, and therefore seemingly unable to know God at all?[3]

Below, I assess a variety of contributions from disability theology that have engaged these questions. First, I introduce Peter Kevern's

1 Abraham Joshua Heschel, *God in Search of Man: A Philosophy of Judaism* (New York, NY: Octagon, 1976), 261. The citation included in the above *motto* is from *Exodus Rabba* 5.9.

2 Joanna Leidenhag, "Autism, Doxology, and the Nature of Christian Worship," *Journal of Disability and Religion* 26:2 (2022): 211–224, esp. 212, https://doi.org/10.1080/23312521.2021.1982840.

3 I use "knowing God" throughout this article in the sense of personal knowledge which in other languages is distinguished more clearly, e.g., as *kennen* instead of *wissen* in German.

diverse approaches to the universality of spirituality in persons with dementia. I draw different conclusions from Kevern's, as all of his approaches require a degree of intellectual capacity on behalf of the person and therefore fail to be fully universal. This is also where theological approaches that rely on psychological research fail to meet the criterion of universality. Second, I introduce the concept of "spiritual intelligence," a term whose coinage itself can be seen as an attempt to distinguish this type of intelligence from general intelligence, and may therefore inaugurate the possibility of speaking about the spirituality of persons whose general intelligence is profoundly disabled. Third, I argue that it is impossible to arrive at a truly universal understanding of spirituality by means of psychological research alone.

Arguing for the apophatic nature of the mental lives of persons with intellectual disabilities, as well as for an understanding of spirituality that acknowledges its transcendent dimension, I propose that psychological research should be complemented by theological anthropology and epistemology. In the fourth section, I illustrate my proposal by discussing three theological approaches to the universality of spirituality and spiritual intelligence.

Approaches to the Universality of Spirituality

When persons suffer from profound intellectual disabilities, severe forms of dementia, or other conditions that heavily affect the brain and the body, it is hard to see how they (still) engage the world around them. Trying to understand if, and envisioning how, they can have spiritual experiences is even more of a challenge. Is it possible to know whether someone is experiencing God or is spiritually engaged? In a critical literature review, religion and dementia scholar Peter Kevern notices that research on dementia and spirituality is limited in its understanding of the spirituality of persons with severe forms of dementia, because of the interpretational character of the assessment.[4]

4 Peter Kevern, "The Spirituality of People with Late-Stage Dementia: A Review of the Research Literature, a Critical Analysis and Some Implications for Person-Centred

The limited or absent communication on behalf of the studied individuals makes it nearly impossible to say anything about their spiritual experience. What someone is thinking can only be conjectured or theorised about. Kevern discusses various approaches that seek to overcome this gap, but he is critical of their effectiveness.[5] He distinguishes five approaches, which I present below, and to which I add other contributions that engage similar issues, such as those that are encountered in research on persons with profound intellectual disabilities.

First, there are various accounts that adopt a palliative or therapeutic approach, which instrumentalises spirituality as an effective means to counter psychological symptoms such as discomfort or aggression. Therapeutic intervention, however, can be quickly dismissed as an approach that substantiates a universal conception of spirituality, as it does not reveal anything about the spiritual experience of persons with dementia, but mainly serves as an argument for the beneficial nature of some spiritual practices.

A second type of approach dismissed by Kevern is the type of ideological approach that sees spirituality as something essentially and intrinsically human, sustained by God or the soul. Although helpful from a more theoretical perspective, Kevern dismisses these approaches as leaving spirituality "with no purchase in the practical world."[6] I return to this approach in the third section of my paper.

Third, there is the romantic approach, which grants the effectiveness of the intuition of a researcher or caregiver in observing the spirituality of persons with dementia. Theologians John Swinton and Harriet Mowat, for example, advocate the method of the "observer-interpreter" who can register the spirituality of the person by careful observation.[7] In one of their articles, they reflect on the story of Mary,

Spirituality and Dementia Care," *Mental Health, Religion and Culture* 18:9 (2015): 765–776, esp. 769, https://doi.org/10.1080/13674676.2015.1094781.

5 Kevern, "The Spirituality of People with Late-Stage Dementia," 770–771.

6 Kevern, "The Spirituality of People with Late-Stage Dementia," 770.

7 John Swinton and Harriet Mowat, *Practical Theology and Qualitative Research* (London: SCM, 2006), 240–241. For a critical discussion of this method, see Jill Harshaw, *God Beyond Words: Christian Theology and the Spiritual Experiences of*

a profoundly intellectually disabled woman, who becomes surprisingly quiet during the quiet time of a church service. They conclude that this unexpected silence is a token of her experienced spirituality. This intuition, however, cannot be validated or substantiated by those lacking said intuition, and therefore falls short of providing a robust ground for arguing for the spirituality of persons with severe dementia or intellectual disability. The argument relies too much on the interpretation of the observer or researcher and therefore, in the end, cannot tell us anything definitive about the spirituality persons like Mary have.

A fourth approach is to see the self as socially extended, which is why the spirituality that belongs to the self can be sustained by a community as well, e.g., by keeping certain memories and stories alive. Disability theologians have pointed to the social nature of spirituality and the communal effort in knowing God. One example is Joanna Leidenhag, who advances this argument when discussing the inclusion of persons with autism in worship. Drawing an analogy between the sensory overload autistic people often experience and the holiness of God, Leidenhag explains that "all humanity is hyper-/and hyposensitive to the presence of God."[8] None of us can see God directly, but together we might get a glimpse. Communally, we can attend to the divine presence and know God: "worshiping together will be especially fruitful and transformative if the gathered congregation is diverse … This is why disability has a prodigious power 'to expand communicative bandwidth'."[9] The diversity of spiritualities and spiritual intelligences (see below) thus accommodates and presupposes the need to learn to worship together.

This type of argument, although valuable for the purpose of including (neuro)diverse ways of worship, falls short of arguing for the universality of spirituality. Leidenhag's theological arguments for the communal nature of attending to the divine presence are convincing,

People with Profound Intellectual Disabilities (London and Philadelphia: Jessica Kingsley Publishers, 2016), 68–84.

8 Leidenhag, "Autism," 218.

9 Leidenhag, "Autism," 220.

but this cannot tell us anything about the individual spiritual experience of persons with profound intellectual disabilities. This can also be seen in Swinton, Mowat, and Baines' article on Mary, as they too conclude that her "spirituality is being formed and held by her participation in the community … She is dependent on her community for her spiritual experience," which is why "Mary's spirituality is a corporate rather than personal concept and experience."[10] Again, although a theological case can be made for the community sustaining Mary's spirituality, it can similarly be objected that there is no way to deduce Mary's individual spiritual experience from this.

The final approach is the cognitive-psychological approach, which argues for the continuing presence of "deep capacities" or procedural memory sustaining the spirituality of those whose overt aptitude for engaging with the spiritual has receded. These movements portray an embodied way of understanding and engaging with the spiritual. John Swinton draws from this type of psychological research when he argues that persons with severe forms of dementia can sustain their spiritual lives in an embodied way: "their movements were memory … they know and remember Jesus in their bodies."[11] Throughout his work, he emphasises the embodied and affective nature of knowing God, as opposed to cognitively knowing *about* God: "knowing about God may not be as important as knowing God, and … knowing God involves much more than memory, intellect, and cognition."[12]

Kevern evaluates the socially-extended self and cognitive-psychological approaches more positively than the other approaches, as they reframe the conception of spirituality itself instead of bluntly

10 John Swinton, Harriet Mowat, and Susannah Baines, "Whose Story Am I? Redescribing Profound Intellectual Disability in the Kingdom of God," *Journal of Religion, Disability and Health* 15:1 (2011): 5–19, esp. 14, https://doi.org/10.1080/15228967.2011.539337.

11 See John Swinton, "What the Body Remembers: Theological Reflections on Dementia," *Journal of Religion, Spirituality and Aging* 26:2–3 (2014): 160–172, esp. 168, https://doi.org/10.1080/15528030.2013.855966.

12 John Swinton, *Dementia: Living in the Memories of God* (Grand Rapids and Cambridge: William B. Eerdmans, 2012), 10.

positing or seeking to intuit the dominant conception of spirituality. Kevern concludes: "if these findings are correct, they imply that the spirituality of people with dementia will, as the condition progresses, come to draw increasingly upon their early, frequently repeated conditioning and upon constant reinforcement by and support from their broader social circle."[13]

Although I concur with Kevern that the final two approaches bear merit, their limitations become apparent when it comes to arguing for a universal conception of spirituality. This is because these approaches eventually account for the spirituality of persons with profound intellectual disabilities on the basis of a previously developed intellectual capacity or the presence of a spirituality-sustaining community. For if persons with severe dementia never developed an embodied spirituality, or not to the extent of developing "deep capacities," the argument cannot be applicable to them. This would be equally true if they lacked a social circle or religious community to sustain their spirituality.

For example, what if there was no community to foster Mary's spirituality?[14] In Christian terms, this would translate to an impossibility on their behalf of (still) knowing God. Another flaw of these approaches is that they fall short of accounting for the spirituality of persons that are born with profound intellectual disabilities. Especially the cognitive-psychological approach, which relies on the persistence of earlier expressions of spirituality, does not suffice in this regard, but the socially-extended-self approach also runs into trouble, as there may (supposedly) be no prior spiritual identity that can be referred to as being maintained or kept alive.[15]

Another approach, which is not discussed by Kevern, is Swinton's emphasis on other attitudes such as love, trust, or faithfulness over understanding or knowledge. In *Becoming Friends of Time*, Swinton explores whether people with profound intellectual disabilities can

13 Kevern, "The Spirituality of People with Late-Stage Dementia," 772.

14 See Harshaw, *God Beyond Words*, 80–81.

15 Kevern, "The Spirituality of People with Late-Stage Dementia," 771.

be disciples of God, which is a variation of the question of how persons with profound intellectual disabilities can know God: how could they follow Jesus if they can never intellectually know anything about him? The problem, according to Swinton, might begin with thinking about the concept of discipleship in purely rational-cognitive terms.[16] This intellectualisation of discipleship is detrimental to letting people with intellectual disabilities belong as disciples. For Swinton, however, discipleship within the Christian community is the ability to love and respond to the vocation to learn together "to love God, and in coming to love God, learn what it means to love and to receive love from all of its members."[17] Instead of seeing discipleship or following Jesus as a personal choice from an autonomous self, we should understand it as an obedient and trusting response to Jesus' call, exactly as it is portrayed in the gospels.

Although helpful up to a point, there are again some difficulties with Swinton's arguments, especially in regard to persons with profound intellectual disabilities. The argument about non-propositional ways of knowing God just changes the question: how can we know that people with profound intellectual disabilities are faithfully or trustfully responding to God's call? One would need to reconstruct the concept of faithfulness or trust to include the attitudes and behaviour of persons with profound intellectual disabilities.

Notwithstanding their value in other regards, none of the above arguments seems to be sufficiently convincing to help conceive of the universality of spirituality and the possibility of envisioning the way persons with profound intellectual disabilities engage with God. They either do not argue for the spirituality of the person in question (the palliative approach) or rely on interpretation from the observer (the romantic approach, arguments for collective and embodied spirituality) or are not universal enough (the socially-extended-self and

16 John Swinton, *Becoming Friends of Time: Disability, Timefullness, and Gentle Discipleship* (Waco, TX: Baylor University Press, 2016), 96, https://doi.org/10.1177/001316447103100435.

17 Swinton, *Becoming Friends of Time*, 93.

cognitive-psychological approaches). Only the ideological approach can argue for the universality of spirituality, but it is dismissed by Kevern for not being practically applicable. I return to this approach in the third section below, after I discuss whether the new concept of spiritual intelligence can help conceive of the spiritual experience of persons with profound intellectual disabilities and thereby universalise spirituality.

Spiritual Intelligence and Intellectual Disability

Thus far, I have spoken about the universality of spirituality. However, relatively recently, it has been proposed that (some forms of) spirituality might be understood as a type of intelligence, called *spiritual intelligence*. By coining spiritual intelligence as a *sui generis* type of intelligence, there may be a recognition that spirituality comes apart from general intelligence. Separating spiritual from general intelligence, in turn, might help to make room for the spirituality of persons with profound general-intellectual disabilities. Spiritual intelligence may thus support a universal understanding of spirituality. Below, I briefly discuss the concept of spiritual intelligence and then assess whether it indeed may fulfil this role.

Psychologist Robert Emmons was the first who proposed adding spiritual intelligence to the other types of diverse intelligence identified by his colleague Howard Gardner in *Frames of Mind*.[18] Emmons' motivation for this was partly to be able to acknowledge that spirituality can be done well (intelligently), that is, can be successful, or can be unsuccessful (unintelligent).[19] Spirituality is intelligent when its aim is to grasp or understand something, or, to put it differently, to accomplish something (e.g., deeper meaning, encounter with God, peace).

18 Howard Gardner, *Frames of Mind: The Theory of Multiple Intelligences* (Bury St Edmunds: St Edmunsbury Press, 1983).

19 Robert A. Emmons, "Is Spirituality an Intelligence? Motivation, Cognition, and the Psychology of Ultimate Concern," *International Journal for the Psychology of Religion* 10:1 (2000): 3–26, https://doi.org/10.1207/S15327582IJPR1001_2.

This line of thought has been further developed and has recently been identified by Harris Wiseman and Fraser Watts as the more implicit side of the dual systems of cognition theories such as Philip Barnard's Interacting Cognition Systems and Iain McGilchrist's hemispheric lateralisation thesis.[20] Thus, spiritual intelligence, besides making room for the intelligent nature of spirituality, can also serve as a critique against an overly cognitive conception of intelligence, which fails to do justice to different types of cognition that are more intuitive, pre-conceptual, and implicit. Wiseman and Watts take this further and provide a participatory conception of spiritual intelligence:

> We wish to recover the earlier assumption that spiritual intelligence is more than a human power. Rather, we wish to explore the powers that humans use in order to engage with and participate in a transcendent spiritual intelligence. Put another way, the psychological dimensions of spiritual intelligence are concerned with the means, manner, and purposes by which a person works with, participates in, gives him or herself over to this transcendent intelligence. It is the powers and processes involved in that giving over that are the chief concern.[21]

Spiritual intelligence might thus be a promising step away from intelligence as something that can be measured on a single scale, from zero to high intelligence. Rather, there are various scales for various types of intelligence. Furthermore, Wiseman and Watts' conceptualisation of spiritual intelligence as a diachronic participation in a transcendent intelligence critiques a skill-based, or ableist, understanding of spiritual intelligence.

Does spiritual intelligence then provide a solution to the issue of spirituality and intellectual disability that I assessed in the previous section? The answer to this question depends on how spiritual

20 Harris Wiseman and Fraser Watts, "Spiritual Intelligence: Participating with Heart, Mind, and Body," *Zygon* 57:3 (2022): 710–718, https://doi.org/10.1111/zygo.12804.

21 Wiseman and Watts, "Spiritual Intelligence," 3.

intelligence is understood. Although the coinage of spiritual intelligence and the subsequent discussions have already contributed by nuancing the understanding of intelligence, much still hinges on certain identifiable capacities on behalf of the individual. Intelligence is, in the end, an evaluative term, expressing whether certain aims are successfully reached and reachable or not. In that spirit, the concept of spiritual intelligence is specifically proposed to emphasise that with spirituality not anything goes; it can be done intelligently or unintelligently.[22]

The diachronic participation account of Wiseman and Watts qualifies this somewhat, as it shifts the understanding of spiritual intelligence as a mere human achievement to an interaction between a human being and the transcendent intelligence it participates in or gives itself over to. Nevertheless, Wiseman and Watts' account does not obliterate human agency: "one participates in the broader spiritual intelligence with one's whole mind, one's whole body, and one's whole heart."[23] The necessary human agency, be it bodily, cognitive, or affective, brings us back to the same problems of the previous section: we cannot account for their presence in persons with profound intellectual disabilities.

The exact nature of spiritual intelligence is still a much-discussed topic and seems to provide promising avenues of interaction with disability theology.[24] Nevertheless, it does not automatically lead to a universal conception of spiritual intelligence, precisely because it seeks to make room for the intelligent nature of spirituality, which harbours an evaluative element that is based on human agency, either in bodily, cognitive, or affective form (or in all three of them). Although the affective agency might be understood to be present in persons with intellectual disability, it is hard or perhaps even impossible to research this for similar reasons as were provided above. In the next section, I therefore return to the ideological approach distinguished by Kevern.

22 Emmons, "Is Spirituality an Intelligence?" 19–21.

23 Wiseman and Watts, "Spiritual Intelligence," 5.

24 Marius Dorobantu and Fraser Watts (eds), *Perspectives on Spiritual Intelligence* (London and New York: Routledge, 2024).

A Theological Addendum

The spirituality or spiritual intelligence of persons with profound intellectual disabilities could in principle be considered to be a purely psychological question. However, the many answers to the question of the universality of spirituality—i.e., those that try to posit God's ubiquitous presence or that present a universal theological epistemology—seem to be at least partly unsatisfactory. As I argued above, the inner mental lives of persons with profound intellectual disabilities escape the methods of psychological research, because these persons cannot express themselves and their behaviour can be incomprehensible or ambiguous. Rather than assuming that their mental lives are therefore lacking or are substantially impaired, I would argue that, from a social scientific point of view, we must remain agnostic about them. This does not mean, however, that the academic conversation has to stop there. On the contrary, against Kevern, I would argue that it is especially important to discuss philosophical and theological (or, as he calls them, ideological) arguments regarding the anthropology and the theological epistemology of persons with profound disabilities. Below, I argue why a theological addendum is called for when speaking about the universality of spirituality.

On a basic level, theological and philosophical presuppositions have a substantial influence on the discussion by informing the conceptual framework. This can be clearly seen when it comes to the concept of spirituality, whose definitional resistance seems to reflect the simple fact that one's conception of spirituality is bound up with what one considers to be spiritual, something which cannot be neutrally described but is always (theologically) put in immanent or transcendent terms.

Howard Gardner, for example, in his discussion of spiritual intelligence, conceptualises "the spiritual" as referring to three dimensions: physical states (meditation), phenomenological states (feeling at one with God or the universe), and a computational aspect that deals with

elements that transcend normal sensory perceptions.[25] He reduces the first to bodily-kinaesthetic intelligence, the second to feelings (which he doesn't regard as intelligent), and calls the residue "existential intelligence" because it describes our capacity to deal with the questions of our existence. From his psychological perspective, spirituality is a sum of psychological aspects that are shared under the same concept, where each aspect is thought of in immanent terms.[26] No wonder Gardner excludes the possibility of spiritual intelligence, although he does leave open the question of whether there might be existential intelligence.

From a (classical) theological perspective, however, "the spiritual" inevitably has to do with transcendence, and so the question of the universality of spirituality is tied up with a theological understanding of transcendence. A theological conception of spirituality could broadly be defined as "having to do with engaging God" and spiritual intelligence as "knowing God in a broad sense." The two terms are very similar in my view, as both have to do with engaging God.

Spirituality is the more common term, whereas spiritual intelligence has more of an evaluative tone to it. Leaving the intricacies of defining either terms for now, whether they are understood as distinctive ways of understanding reality or understanding a different reality,[27] spirituality and spiritual intelligence are about grasping, understanding, or encountering the spiritual, which, from a Christian perspective, amounts to encountering the presence of the triune God. Spiritual knowing is therefore relational knowing constituted by an encounter between God and human beings.

When we accept spirituality as relating to that which transcends our knowledge, we have to acknowledge that what spirituality exactly is

25 Howard Gardner, "A Case Against Spiritual Intelligence," *International Journal for the Psychology of Religion* 10:1 (2000): 29, https://doi.org/10.1207/S15327582IJPR1001_3.

26 Although Gardner states that he does not doubt that the phenomenological state of spirituality can be genuine, he does not seem to accept that such a state may also be an actual apprehension of something that is really transcendent, such as God. Gardner, "A Case Against Spiritual Intelligence," 29.

27 Dorobantu and Watts, *Perspectives on Spiritual Intelligence*, 9.

and who exactly is capable of being spiritual also partially transcends our knowledge. For how could we completely understand how we relate to that which, or the one who, we can never completely understand? Spirituality as being related to God therefore belongs to one of the mysteries that in Christianity is often marked as such, namely, that of how finite persons can share in the divine life. Although there is no consensus on what exactly a mystery amounts to, a classical theological perspective would understand it to be something that we cannot grasp, except from revelation. From a Christian point of view, a theory for the universality of spirituality thus needs to be theologically informed.

The conceptual framework of spirituality can thus be seen to be informed by a broader theological framework. This can also be said of the concept of intelligence, something which should be borne in mind when it comes to intellectual disability and the discussion of spiritual intelligence. Historian C. F. Goodey has convincingly argued that the concept of intelligence reflects what is valued in a society of a specific time period, just as the concept of intellectual disability reflects who is excluded in that same society.[28]

There is a helpful analogy here with the debate on personhood and disability. Personhood, not unlike intelligence, is a concept that is (at least partially) informed by ideological considerations. Although both concepts seem to be natural to us, there is much to be said about what constitutes either a person or a certain type of intelligence. When it comes to personhood, this is perhaps even more obvious. Disability theologians have repeatedly unnerved anthropological assumptions informing medical ethics, academic research, and public policy. A salient example is John Swinton's discussion of ethicist Peter Singer's liberal anthropology, which is heavily based on human autonomy, rationality, control, and general intellectual ability. Swinton strongly criticises Singer's anthropology, including his understanding of persons with severe forms of dementia as having ceased to be persons, whose

28 C. F. Goodey, *A History of Intelligence and 'Intellectual Disability': The Shaping of Psychology in Early Modern Europe* (New York: Routledge, 2011).

life is therefore no longer inherently valuable or worthy of protection.[29] Such a conclusion is not supported by mere social scientific evidence but imbued with philosophical and (a)theological presuppositions.

Against such a view, Linda Woodhead has argued that a theological anthropology should be an apophatic anthropology.[30] Since human beings are created in the image of God, their essence is ungraspable, just as God's essence is ungraspable. As it is impossible for us to understand human nature fully, it is impossible fully to understand what constitutes human personhood, as well. Such an apophatic theological anthropology is therefore embedded within a larger theological understanding of reality, which is why Woodhead is able to unpack what such a theological apophatic anthropology implies. In short, she argues that the *imago Dei* account of personhood can be understood to imply that human beings become persons by increasingly reflecting the image of God:

> Human beings, made in the image of God, do not contain their essence in themselves but in the God into whose image they are to grow. They become human by becoming divine—which means growing into something we do not know or control rather than something we already possess.[31]

Woodhead's theological anthropology is tightly bound up with what may be called her theological epistemology: knowing God is becoming more like God and thus increasingly participating in the divine nature. Interestingly, her anthropology thus seems to presuppose the possibility for human beings to grow in possessing "the unknowable character

29 John Swinton, "Forgetting Whose We Are: Theological Reflections on Personhood, Faith and Dementia," *Journal of Religion, Disability and Health* 11:1 (2007): 43, https://doi.org/10.1300/J095v11n01_04.

30 Linda Woodhead, "Apophatic Anthropology," in *God and Human Dignity,* ed. R. Kendall Soulen and Linda Woodhead (Grand Rapids, MI: Eerdmans, 2006), 233–246.

31 Woodhead, "Apophatic Anthropology," 236–237.

of the divine being,"[32] and thus seems to require a universal understanding of spirituality as well. The differences noted in theological and philosophical anthropologies therefore do not just serve as analogies for the question of spirituality; they cohere with it closely. Whether spirituality is a universal human feature is an important question that can have serious consequences for our understanding of persons with intellectual disabilities. It should therefore be the subject of academic inquiry, even though this takes us to an ideological, or better, theological argumentation that goes further than the empirical evidence, or "the phenomena."

Besides the arguments presented above, I believe a theological approach is called for precisely because the question it seeks to answer is often one of theological or pastoral concern: can our loved ones, regardless of their intellectual deficits, relate to God?[33] Such a question itself seems to arise out of an intellectualistic theology that may very well be intrinsically exclusionary to persons with profound intellectual disabilities. Jill Harshaw therefore wonders whether certain research strategies that try to search for traces of intelligence in persons with profound intellectual disabilities might be induced by "a subconscious fear that … such persons are not fully capable of a genuine relationship with God … so that, in order to be comfortable about asserting their capacity for spiritual life, we should assume that it can be identified and explored through cognitive and linguistically based methods."[34]

Relating these considerations to the previous section, it seems as if an ideological approach—in the sense of a discussion about the theological and anthropological presuppositions of the debate—is not just warranted, but even called for. This is where I part ways with Kevern, who dismisses an ideological approach to the spirituality of persons with severe forms of dementia, because such an approach "makes no space for 'spirituality' as pertaining to human beings and their actions,

32 Woodhead, "Apophatic Anthropology," 238.

33 The usual follow-up question is expected to be "and can he or she thus be saved?" which reveals a soteriological concern.

34 Harshaw, *God Beyond Words*, 84.

and so leaves the concept with no purchase in the practical world."[35] It is not entirely clear to me what Kevern means by the need for the concept of spirituality to have a purchase in the practical world, but I assume this has to do with his main purpose of trying to be able to "understand their [namely, people with severe dementia's] spirituality as extending beyond that point to the end of their life, integral to their personhood regardless of any loss of other capacities and competencies."[36] If this is the desired purchase a theory of the universality of spirituality needs to have, I do not see why a theological approach would not work. Provided that on the basis of the only approaches that Kevern finds promising, there are in fact certain competencies or conditions necessary, I would argue that his conclusion risks denying spirituality to those that lack those necessary competencies or conditions.

I believe that this is the case for any approach that defines spirituality purely immanently, as such an approach seeks to find or argue for some trace of spirituality in the human person itself. If it is then asserted that spirituality is a core feature of humanity, the lack or inadequateness of convincing evidence or arguments for a person's spirituality can backfire to imply that such a person is therefore no longer a human person.[37]

Thus, the understanding of the spirituality of persons with intellectual disabilities as scientifically inaccessible and theologically mysterious does not mean that we should not reflect on it.[38] In this light, many of the approaches that I initially dismissed as being able to argue for the universality of spirituality are in fact helpful as arguments for a certain theology.[39] They problematise the question of having to know

35 Kevern, "The Spirituality of People with Late-Stage Dementia," 770. Kevern does acknowledge that this type of account can be drawn on as a possible explanation.

36 Kevern, "The Spirituality of People with Late-Stage Dementia," 770.

37 Kevern, "The Spirituality of People with Late-Stage Dementia," 766.

38 Cf. Alister E. McGrath, *The Territories of Human Reason: Science and Theology in an Age of Multiple Rationalities* (Oxford and New York: Oxford University Press, 2019), 195.

39 This aligns with Kevern's approval of John Paley's objection that findings like

God intellectually, as Swinton does by emphasising love, trust, and faithfulness. Similarly, the arguments for an embodied, communal notion of spirituality gain force when understood within the framework of Christian theology.

The fact that something is a mystery also means that it doesn't admit of general solutions, but that it needs to be dealt with in concrete existence.[40] That is also why I am less sceptical of the possibility of what Kevern calls the romantic or intuitive approach. Why should we dismiss the idea of personal knowledge which can be acquired by genuinely paying attention, and being present, to persons with profound intellectual disabilities for a long period of time?

Many arguments in disability theology are presented in narrative form or supported by illustrative personal accounts which are, in my opinion, more of a strength than a weakness.[41] They often provide touching examples of how, regardless of the severity of disability, the spirituality of persons with intellectual disabilities seems to be revealed at times.[42] The same is true for the story of Mary. It is one more story that illustrates how people do have experiences in which they are convinced that they can perceive the spirituality of their disabled loved ones. Within a theological framework, there are therefore strong accounts to be given for the universality of spirituality.

these "only make sense as spirituality if some theological or religious concepts have been 'smuggled in'." Kevern, "The Spirituality of People with Late-Stage Dementia," 770.

40 McGrath, *The Territories of Human Reason*, 192–193.

41 See also Harshaw, *God Beyond Words*, 41: "any attempt to explore the spiritual lives of people with profound intellectual disabilities will necessarily involve affording attention and respect to the particularities of their embodied experience."

42 See, for example, Frances Young, *Arthur's Call: A Journey of Faith in the Face of Severe Learning Disability* (London: SPCK, 2014).

The Inclusivity of God's Revelation: Contributions from Disability Theology

In this final section, I provide a few examples of theologians who have argued directly for the universality of spirituality or spiritual intelligence in a theological way. This should serve to illustrate the point that the mystery of the spiritual lives of persons with profound intellectual disabilities is unsolvable through scientific methods, but can be grappled with theologically, albeit inconclusively. I first discuss Jill Harshaw's account of the accommodation of God, then move to Erinn Staley's argument from negative theology, and end with Petre Maican's argument from the Orthodox conception of the *nous*.

In her book *God Beyond Words*, Harshaw explores whether persons with profound intellectual disabilities can be understood to have spiritual experiences. She does so in a thoroughly theological way and likewise argues that this is the only adequate way: "Rather than asking these people for information they cannot provide, or relying on assumptions made by those around them, questions can be addressed to the *source* of any spiritual experience they might have."[43] She therefore identifies God's self-disclosure as the proper point of focus and emphasises God's agency in our spirituality. All of us, regardless of our abilities and disabilities, depend on God's revelation to us, which makes possible our encounter. It is God who acts first, so there is always an element of grace included in encountering God. Even our receptiveness to God's revelation is grace. After stating the above, Harshaw introduces divine accommodation into the discussion. God's communication and revelation are necessarily adjusted to human receptivity. After a lengthy discussion, Harshaw concludes that divine accommodation includes all human beings:

> The fundamental aim of accommodation is relational communication between God and human beings. Words are not the exclusive means by which this communication occurs. Words are

43 Harshaw, *God Beyond Words*, 183.

> merely signs and pointers to a reality which is behind and transcends the means of its expression—the person of Jesus Christ who is the greatest accommodation to humanity's inability to apprehend God.[44]

Denying this would be "underestimating the depths of universal human incapacity when it comes to understanding God."[45]

A somewhat similar avenue of arguing for the universality of spirituality is Erinn Staley's account, which draws from the *via negativa*, the apophatic tradition. She discusses the theologies of Bonaventure and Meister Eckhart, in which unknowing is central. Leaving the intricacies of their apophaticism for now, engagement with them leads Staley to the conclusion that "pointing toward the unknowability of God reminds humanity that the smartest human being is far more like a person with an intellectual disability than he or she is like God."[46] However, this unknowability does not mean we cannot relate to God. Swinton takes a similar stance and even suggests it may be the other way around:

> If a lack of a certain attitude toward propositional knowledge is in some senses important for becoming a disciple, it may be that our brothers and sisters living with profound intellectual disabilities are in a stronger position before God than are those of us who are in many ways held back by our intellect and the desire for life to be reasonable.[47]

---

44 Harshaw, *God Beyond Words*, 90.

45 Harshaw, *God Beyond Words*, 116.

46 Erinn Staley, "Intellectual Disability and Mystical Unknowing: Contemporary Insights from Medieval Sources," *Modern Theology* 28:3 (2012): 398, https://doi.org/10.1111/j.1468-0025.2012.01757.x.

47 Swinton, *Becoming Friends of Time*,104. See Christina M. Puchalski, "Dementia: A Spiritual Journey for the Patient and the Caregivers"; and John Swinton, "Known by God", in Hans S. Reinders (ed.), *The Paradox of Disability: Responses to Jean Vanier and L'Arche Communities from Theology and the Sciences* (Grand Rapids, MI: Eerdmans, 2010): 37–50, 140–153. However, one should be careful, as Swinton is, not to make this into some sort of advantage of being intellectually disabled. See on this, Harshaw, *God Beyond Words*, ch. 6: "The Mystical Experience of God."

Departing from a Christian theological perspective, there is a profound sense of the limits of human knowledge, especially with regard to knowing God. All our knowledge of and about God is dependent on God, so it might be arbitrary to decide that the gap of knowledge is too big for some individuals to cross but small enough for us to jump over.[48]

Finally, Romanian Orthodox disability theologian Petre Maican discusses the relationship between God and persons with dementia. He argues that the Patristic conception of the *nous* helps to understand how persons who are seemingly unresponsive or unperceptive to the world around them can still relate and know God. The *nous*, or spirit, or mind, is the third constitutive element of human beings, next to the body and soul, and functions as the "intuition of God."[49] As we cannot lose our spiritual faculty, the universality of spirituality is guaranteed, even in severe cases of dementia or profound intellectual disability. There is always a relationship between us and God.

Conclusion

The universality of spirituality and spiritual intelligence is desirable from the perspective of disability theology. It tries to find an answer to a genuine concern of many religious persons, namely, whether their loved ones with profound intellectual disabilities or dementia can (still) engage with the transcendent, or are able to know God. There have been various attempts to find an answer to this question. In some cases, these focused on the perception or intuition of the researcher or caregiver, in other cases they pointed to something visible external to the minds of these people, such as their body or the community surrounding them.

In the first section, I assessed five types of approaches that Peter Kevern distinguished, and added a few that deviate slightly from them. Where Kevern finds the cognitive-psychological and

48 Staley, "Intellectual Disability and Mystical Unknowing," 389.

49 Petre Maican, "Spiritual Intelligence and Intellectual Disability: A Theological Re-evaluation of the *Nous*," in this special issue of *CPOSAT*.

socially-extended-self approach promising, I pointed out their shortcomings in arguing for the universality of spirituality or spiritual intelligence.

In the second section, I assessed whether a relatively novel concept, spiritual intelligence, might be more promising, as it distinguishes spiritual from general intelligence. It might therefore help to make room for the spirituality of persons whose general intelligence is profoundly disabled. Although it seems to me that a lot can be gained from bringing spiritual intelligence into discussion with disability theology, at this point there does not seem to be enough consensus on its nature to build a case for the universality of spirituality.

In the third section, I argued that this might never be the case, as the spiritual mental lives of persons with profound intellectual disabilities are intrinsically inaccessible to us. I argued that this is not just because of our current (scientific) incapability, but because spirituality has to do with transcendence. If understood immanently, spirituality does not only seem to lose its coherence as a concept, but also seems to lack in persons with profound intellectual disabilities, as they are required to have certain capabilities to experience phenomenological and computational states. However, if understood transcendently, as relating to the transcendent or to God, the inherent or absolute mysteriousness of the spiritual mental lives of persons with profound intellectual disabilities is emphasised.

This led me to conclude that a theological discussion is called for. I illustrated this by providing three arguments for the universality of spiritual intelligence by theologians Jill Harshaw, Erinn Staley, and Petre Maican. Their accounts of God's self-disclosure and the accompanying necessity of our "intuition" of God present strong arguments for the universality of spirituality. These final three approaches seem to be more promising, as they do not depend on any condition from the side of the person that experiences God or has spiritual understanding. There is neither a need to have internalised spiritual practices nor to participate physically in church services nor to be part of a community in order to have spiritual experiences.

What this means for spiritual intelligence is not entirely clear yet and depends on its precise conception. If spiritual intelligence is understood as participation in a transcendent intelligence (God), then I would argue that these theological arguments apply to it and argue for its universality in a similar way. In that case, spiritual intelligence would mean something similar to spirituality, thus something like engaging with the spiritual, but with a more evaluative connotation. However, if spiritual intelligence is understood more specifically as an ability or skill of an individual person that needs to be cultivated and practiced, it would point to a rather particular type of intelligence, which may be unreachable for persons with profound intellectual disabilities.[50] To conclude, from a theological perspective, a good case can be made for the universality of spirituality and for a specific form of spiritual intelligence.[51]

The author reports there are no competing interests to declare.
Received: 23/02/24 Accepted: 19/08/24 Published: 19/03/25

50 There may be two conceptions of spiritual intelligence that can be related to the distinction I make here. See Harris Wiseman, "The Japanese Arts and Meditation-in-Action," *Zygon* 32:3 (2022): 194–208, https://doi.org/10.1111/zygo.12806.

51 I am thankful for the helpful suggestions of Marius Dorobantu, Fraser Watts, and Harris Wiseman.

Spiritual Intelligence and Dementia: A Theological Reevaluation of the *Nous*

Petre Maican

Abstract: Discussions on spiritual intelligence make only timid references to the topic of intellectual disability. Questions such as what spiritual intelligence could mean for someone whose IQ has been medically assessed at 20 or what happens with the spiritual intelligence of persons who develop Alzheimer's are rarely answered. When this happens, we are presented mostly with an Aristotelian Thomistic notion of the soul (including intelligence) being the pattern of the body held in the memory of God. This approach, however, does not clarify how could it be possible for someone's soul to live in the memory of God without preexisting in God's mind before the existence of the world. This article suggests that a better and more inclusive approach rests with the reevaluation of the patristic notion of *nous* as the spiritual intelligence that preexisted in God, which links human beings with God irrespective of the state of their bodies.

Keywords: Sergius Bulgakov; dementia; disability theology; *nous*; spiritual intelligence; John Swinton

---

Petre Maican is Associate Researcher at The Institute for Eastern Christian Studies of Radboud University (The Netherlands). His current research interests lie at the crossroads between theological anthropology, intellectual disability, and Eastern Orthodox spirituality. His latest book, *Deification and Modern Orthodox Theology*, was published with Brill (2023).

Mass media and popular culture abound with references to dementia. The film *The Father* (2020), showed how difficult the life of a person with dementia can be for the person himself and for his loved ones. In 2021 it won two Oscar awards and was nominated for six. In the very first week of 2024, BBC published an article on the series of photographs produced by Helen Rimell that documented the "agony of slowly losing [her] mum"[1] to dementia. With almost 55 million people in the world touched by dementia and 10 million discovered every year,[2] there is no surprise the condition is so present in our lives and social imagination.

Disability theology has long caught on with the trend, with an impressive number of books and articles being published every year.[3] In these writings, dementia is regularly approached from three perspectives: practical advice; collection of data through interviews with persons with dementia and their carers; and systematic theology. In the latter case, the main question revolves around the notions of self

1 "Dementia: Photos Lay Bare Agony of Slowly Losing Mum," *BBC News*, 5 January 2024, sec. Wales, https://www.bbc.com/news/uk-wales-67762173.

2 "Dementia," https://www.who.int/news-room/fact-sheets/detail/dementia (accessed 30 January 2024).

3 Kenneth L. Carder, *Ministry with the Forgotten: Dementia Through a Spiritual Lens* (Nashville, TN: Abingdon Press, 2019); Ginnie Horst Burkholder, *Relentless Goodbye: Grief and Love in the Shadow of Dementia* (Harrisonburg, VA: Herald Press, 2012); Albert Jewell (ed.), *Spirituality and Personhood in Dementia* (London: Kingsley, 2011); Peter Kevern, "The Spirituality of People with Late-Stage Dementia: A Review of the Research Literature, a Critical Analysis and Some Implications for Person-Centred Spirituality and Dementia Care," *Mental Health, Religion & Culture* 18:9 (2015): 765–76, doi:10.1080/13674676.2015.1094781; Peter Kevern, "What Sort of a God Is to Be Found in Dementia? A Survey of Theological Responses and an Agenda for Their Development," *Theology* 113:873 (2010): 174–82, doi:10.1177/0040571X1011300303; Kristin Beise Kiblinger, "Theology of Dementia and Caputo's 'Difficult Glory'," *Journal of Disability & Religion* 28: 2 (2024): 142–63, doi:10.1080/23312521.2023.2197424; Pia Matthews, "Changing the Conversation: From Suffering with Dementia through Dementia as a Disability Rights Issue, to a Deeper Theological Perspective," *Journal of Disability & Religion* 23:2 (2019): 149–65, doi:10.1080/23312521.2019.157720; John Swinton, *Dementia: Living in the Memories of God* (Grand Rapids, MI: Eerdmans, 2012); John Swinton, "What the Body Remembers: Theological Reflections on Dementia," *Journal of Religion, Spirituality & Aging* 26:2–3 (2014): 160–72, doi:10.1080/15528030.2013.855966. These are meant to serve just as examples; the list is not exhaustive.

and spiritual identity. How can a person be saved or have a spiritual life if she does not know who Jesus is anymore?[4]

In this paper, I would like to deepen one of the solutions proposed to this latter question. The solution comes from John Swinton, a renowned Scottish theologian. Swinton argued that all human beings have their identity established by God from eternity, or as he puts it quoting from Augustine, that all humans live in the memory of God. What I will claim is that, although convincing, Swinton's argument stops short of drawing the logical conclusions of its premise, namely that only if we postulate a faculty of spiritual intelligence that preexists our embodiment we can actually preserve our identity and live eternally in God's memory. This faculty, I will contend, is none other than the *nous* appearing in the writings of the Greek Fathers. Or to put it differently, I will suggest that the relationship between God and the person with dementia endures and even flourishes through the mediation of the *nous* or spiritual intelligence.

For the sake of clarity, I need to specify from the beginning what I mean by both these terms; preexistence and *nous*. In this article, I will use the word preexistence in order to emphasise that from the perspective of history as experienced by humans, the *nous* precedes human embodiment, even if from God's standpoint everything happens simultaneously. By *nous*, I understand the spiritual intelligence that comes from and connects humans with God, helping them to develop their spiritual identity throughout history. I render *nous* as intelligence instead of intellect following the lead of Rowan Williams, who prefers this option "on the grounds that 'intellect' has for most readers a narrower and more conceptually focused sense than 'intelligence'."[5]

I will develop this argument in three steps. First, I will provide an overview of Swinton's position, emphasising that his description of God's living and eternal memory presupposes the understanding of time as a spacetime block. Then, I will point out the main implication of

4 Swinton, *Dementia*, 188.

5 Rowan Williams, *Looking East in Winter: Contemporary Thought and the Eastern Christian Tradition* (London: Bloomsbury Continuum, 2021), 14 n. 7.

this view for anthropology, namely that the *nous* is created in advance and is united with a body only when entering human history. Finally, I will discuss the convergence of this position with some of the research already present in the science-theology dialogue regarding the relationship between the *nous* and human cognitive capacities.

Dementia and Theological Anthropology

John Swinton's book *Dementia: Living in the Memories of God* (2012) is one of the most influential in the field of disability theology. The book is impressive because it combines skilful storytelling and deep systematic reflection on human nature with a compassionate account of God's love towards us, regardless of the physical state in which we find ourselves. Its main premise is that human beings are more than their cognitive capacities or the web of social relationships they establish throughout their lives; they are persons. Swinton disagrees with the view that humans are persons only as long as they are able to give narrative shape to their memories and life choices. For him, this is sheer reductionism. Even if we do not see, hear, or engage with the person the same way as before, it does not mean we should assume they have lost their identity.

Swinton also challenges the definition of personhood as social relationships—that humans are persons because they always find themselves in relationships with others. For him, this definition ignores those who do not have friends or whose friends have been forgotten because of their dementia. Consider someone who at 90 years of age has outlived all her friends and family and continues to live alone in the countryside. It would be strange to claim that she is not a person simply because nobody visits her.

The proposal Swinton puts forward is profoundly theological. Personhood is indeed about relationships, but about the transcendental relationship with God. Commenting on the biblical text of Genesis (2:7) that mentions that after God moulded the human being out of clay he breathed life into her, Swinton notes that the biblical term for God's

breath over the human being, the *nephesh*, is simultaneously proof that all human beings are desired and loved by God and that God has made us persons from the moment of creation because we are placed in an eternal relationship with God through *nephesh*.[6] Without God's *nephesh*, humans lose not only their memory or sense of self, but they cease to exist altogether. It is the *nephesh* that preserves human beings ontologically.[7] To be a person, then, is to be brought into existence, sustained and loved by God through the divine breath, not for our capacities or our relationships with other humans, but for ourselves.

Nephesh does not serve only as the basis for personhood, but also for a distinction between embodied and spiritual identity. For Swinton, human beings are *nephesh*, animated bodies, who develop a sense of identity based on the embodied experience of their past, their preferences, their relationship with others, and their perception of the world. In a sense, *nephesh* is the life principle that animates human beings, allowing them to develop a sense of personal identity. This identity, however, is not the same as the one every human being has in God and which is bestowed upon us by the Holy Spirit.

> The Spirit forms and names the living soul, the *nephesh*-inspired body. The brain plays a part in this process of human development, but it is not definitive or determinative of it. It is God the Holy Spirit who determines who a person is; the brain as an aspect of the body simply participates in the movement of people towards their given goal. We do not actualise ourselves; rather, in some sense or another it appears that *we are told who we are*.[8]

For Swinton, this means that humans live their entire lives with a false or, in the best case, incomplete sense of identity. They believe they know who they are because they remember events from their past, but their memories are often inaccurate and fragmentary. Swinton

6 Swinton, *Dementia*, 184.

7 Swinton, *Dementia*, 165–68.

8 Swinton, *Dementia*, 175.

illustrates this point with a personal anecdote: when he is talking to his mother about his childhood, he remembers certain episodes she does not and vice versa. The difference lies in which episodes from his childhood he decided to integrate into his sense of self. "When I think back on my past, I remember some things about what I once was and where I've been. But my mother remembers other things, things that I don't recall. When she tells me, I graft them into my story, and eventually become a part of my memory system."[9]

Very rarely, we are those we think we are or who we unveil to our friends. Our true identity is found in God and it will be revealed to us only in the eschaton. This spiritual identity endowed upon us by the Spirit is not lost in dementia; it continues to exist in God, long after our cognitive abilities have been lost.[10]

To support the view that spiritual identity comes from God and remains in God for eternity, Swinton turns to Augustine's discussion of the nature of time from *Confessions*. Swinton borrows from Augustine two points: first, God is beyond time, and second, memory refers to the sustaining activity of God throughout history. For Augustine, God must be outside time in order to stress his omnipotence and impassibility. If God is affected by temporality, then God has to change, and change belongs only to creatures because it entails imperfection, lack, or suffering. For Swinton, only a God who is perfect and outside time can guarantee that human identity remains unaffected by the way we die or age.

Swinton also insists that memory is not a simple remembrance of things past, but the sustaining activity of God taking place throughout history. Where human memory forgets and mixes things up, divine memory is faithful, sustaining, and nourishing the person. To be remembered by God is not so much a claim about the past or the eschaton, but very much about the present.

9 Swinton, *Dementia*, 221.

10 Swinton, *Dementia*, 219.

> To be held and remembered by God implies some form of divine action towards the object of memory. It is not purely eschatological action; it is something that occurs in the past and in the present as well as in the future. God acts in particular ways towards people because of a previous commitment. In other words God remembers because he promises.[11]

There are, however, two matters that I feel Swinton does not clarify enough. The first one is the implication arising from Augustine's treatment of time, the preexistence of the souls. To live in the memory of God, where there is no past, present, and future, entails that the entire creation came into being and was completed through a single and eternal act of God, like a huge time-block inside which time unfolds in a succession of past, present, and future. If this is so, then it would also make perfect sense to think of humankind as being completed, with the *nous* or the spiritual intelligence of each human being already existing and only waiting to enter human history at their right moment.

This would make even more sense in light of the second matter Swinton left underdeveloped, the relationship between the period in which a person has dementia and her spiritual identity. Swinton's main concern is the preservation of the spiritual identity of persons with dementia in God, regardless of their cognitive capacities. The question that arises is whether the period with dementia contributes in any way to the spiritual identity of the person. If the answer is *no*, would it make sense to sustain euthanasia on compassionate grounds? If the answer is *yes*, then would one need to explain the link between the *nephesh*'s embodied identity and the identity bestowed by the Holy Spirit? How is the experience of dementia transferred to or relevant for the spiritual identity in God?

It seems to me that the notion of *nous* or spiritual intelligence created before our embodiment could provide answers to both these points. To support this claim, I will begin by explaining why the preexistence of the *nous* makes sense in a spacetime block identifiable with

11 Swinton, *Dementia*, 216.

the one described by Augustine and appropriated by Swinton. Then I will move to discuss the concept of *nous* and its relation with our bodies.

The Spacetime Block of God's Memory

Swinton's interpretation of time in Augustine has a lot in common with the one provided by Paul Helm. Helm makes the same point: that for Augustine God is outside time and that the world was created at once, but he dwells a little longer on the implications of this creative act. Helm explains that there are two ways of understanding what Augustine means when he states: "In the sublimity of an eternity which is always in the present, you are before all things past and transcend all things future because they are still to come."[12] The first is to see time as a succession of events in a consecutive series, as defined by words like yesterday, today or tomorrow. "One can only refer to a particular day as yesterday from a standpoint within time; if 14th September is yesterday, then 15th September is today, and so on."[13] The second option is to understand time

> from an a-temporal perspective, by expressions such as earlier than and later than, before and after. Thus on the B-series view of time, Napoleon's defeat at Waterloo is earlier than Montgomery's victory at El Alamein; the event of the Battle of Waterloo occurs before that of El Alamein. But Napoleon's victory is only in the past from the standpoint of someone who exists later than that date; and only in the future from the standpoint of someone who exists earlier than that date.[14]

Helm excludes the first interpretation of God creating time as a series of successive events, not only based on textual evidence from Augustine's

12 Paul Helm, "Eternal Creation: The Doctrine of the Two Standpoints," in *The Doctrine of Creation: Essays in Dogmatics, History and Philosophy*, ed. Colin E. Gunton (Edinburgh: Clark, 2004), 39.

13 Helm, "Eternal Creation," 39.

14 Helm, "Eternal Creation," 39.

work, but also for the same reason as Swinton; it would entail that "God is subject to the vicissitudes of temporal passage," something that is "incompatible with divine sovereignty, perfection and with that fullness of being that is essential to God."[15]

Moreover, this perspective would place God's memory on the same level as human memory, asking whether God does indeed remember everything and how.[16] Helm concludes that what we have are actually two different standpoints regarding time. From the perspective of the creature, time flows, but from God's perspective, everything happens at the same time. As he explains, "the Creator may be said to be continuously creating the universe, in that there is more universe today than there was yesterday, for the present builds upon and is made intelligible by the past. But from the divine standpoint what is created is one temporally extended or ordered universe."[17] Or as William Craig reformulates the position, "God is the Creator of the universe in the sense that the whole block and everything in it depend upon God for its existence. God by a single timeless act makes it exist. By the same act, He causes all events to happen and things to exist at their tenseless temporal locations."[18]

The Preexistence of the *Nous*

This view of time as a spacetime block is not foreign to systematic theologians. Most of them assume it without following it to its logical conclusion: the preexistence of spiritual intelligence. One of those who took this step was the Russian theologian Sergius Bulgakov (1871–1944), probably one of the most creative Eastern Orthodox thinkers of the twentieth century and certainly one of the few who still fascinates, intrigues, and stimulates contemporary theological

15 Helm, "Eternal Creation," 30.

16 Helm, "Eternal Creation," 40.

17 Helm, "Eternal Creation," 35.

18 William Lane Craig, *Time and Eternity: Exploring God's Relationship to Time* (Wheaton, IL: Crossway Books, 2001), 111.

imagination. In the *Bride of the Lamb*, Bulgakov makes the same points as Helm and Swinton about God's timelessness, namely, that God exists above time—in what he calls "supratemporality"—and that creation is a spacetime block with our sequential perception of time being a result of our creatureliness.

> The eternal aspect of the creative act reveals in God himself the character of creation in all its fullness. God sees as existent even what has not yet occurred, for, in Him and for Him, it supra-eternally is as the creaturely Sophia. But in relation to time this supra-eternity of creation signifies its supratemporality in the sense that, for God, creation exists in a certain integral of all-temporality, whereas for creatures it is the unfurling scroll of the time of empirical being. This supratemporality is realized in the life of creatures, for they are created for time, but in a certain supratemporal mode of being.[19]

For Bulgakov, this view corroborates rather well with the testimony of Scripture about God resting on the seventh day. If God rests, then God has already created and completed everything, including the whole of humanity. If humanity is still being created, then creation is not completed, God does not rest and could even be surprised by what will take place in the future.[20] This option is of course excluded by a spacetime block. What we perceive as consecutive events is not a continuous creation, but "the unfurling scroll of the time," where generations of humans appear one after another in history, although they already exist in the supratemporality of God. "For humanity in particular, this signifies that, although in earthly beings, human generations are born and thus appear in time, as it were, this is possible only on the basis of the supratemporal creation of all of them."[21]

19 Sergius Bulgakov, *The Bride of the Lamb* (Grand Rapids, MI and Edinburgh: Eerdmans and T&T Clark, 2002), 112.

20 Bulgakov, *The Bride of the Lamb*, 112–13.

21 Bulgakov, *The Bride of the Lamb*, 112–13.

Thus, the preexistence of humanity in God should not be understood in what passes for the Origenistic way, where the *nous* has already been created and entered the material world after a cataclysmic event that forced it away from the contemplation of God. For Bulgakov, God created everything at once and the creation is already finished. It is only from the perspective of human history that we can speak of the preexistence of the *nous*. Humanity exists in God simultaneously as a concept and concrete individuality.

For Bulgakov, this double perspective on time allows us to explain the doctrines of the original sin and human redemption in Christ. Only if the entirety of humankind participated in Adam's sin it makes sense to speak about the Fall of humanity through Adam and their recapitulation in and redemption through Jesus Christ. If human beings are still being created, then they have not participated in Adam's sin, are blameless, and cannot be recapitulated in Christ's sacrifice. Bulgakov insists that humanity does not exist in Adam and Christ as an indiscriminate mass, but as individual centres of consciousness.

These centres of consciousness or *I*s are the ones that have been created before entering history and being united with their bodies. And although they are influenced and shaped by the body and the material world, the *I*s retain their atemporal, semiautonomous and semi-divine status.

> To be sure, this origination occurs not in time, for *I* itself looks down from its height at time, is for time an immobile sun, illuminating its movement. However, the creaturely *I* exists for and in time, is connected with temporality. But despite this, it is free of the discursiveness of time, composed of a series of separate moments or determinations, and is not at all exhausted by them. The creaturely *I* is never free of time but always belongs to it, is correlative to time, directs its light projector at time.[22]

22 Bulgakov, *The Bride of the Lamb*, 85.

Bulgakov claims that the embodiment of the *I* serves two purposes. First, the *I*s with their direct relationship with God and relative independence from matter are meant to spiritualise the world, making it more transparent to God's love for creation and in this way prepare it for the fuller divine-human communion that will take place in the eschaton. Second, history functions as a training-ground for these spiritual centres of consciousness. The *I*s have to develop and fashion themselves according to a specific theme that God has designed for them from the beginning of time.

This idea can be better understood through analogy with the human body. In 1 Corinthians 12, Paul explains that Christians are all called to fulfil different functions within the church. Some are called to be prophets, others are called to preach, and others to speak in tongues. This variety of gifts is to be expected because the church is similar to a body; it cannot function properly with just one type of organ. Diversity is essential.

Bulgakov gives the analogy universal proportions. The entire humanity was created in Adam and recapitulated in Christ, so each human being has a certain role to play in the proper functioning of the universal body of Christ. This function refers to the theme that each *I* has received from God. To fulfil it, the *I* has to guide the body and engage with the material world. In this case, human life represents the attempt to find one's pre-established place in the body of Christ. Or, to transpose this idea into Swinton's language, the identity of every human being exists in the memory of God, but this identity reaches its fulfilment only through the interaction of the *I* with the body and other human beings in history.

The main difference between Bulgakov and Swinton lies in their interpretation of Genesis 2:7. Swinton interprets God's breath as the principle of life that makes the human body move and exist in the world, while Bulgakov interprets it as referring to consciousness and intellectual capacities. The principle of life, he thinks, is the soul, which humans have in common with the entire creation, animals and

plants alike, while the spirit or the *I* comes from the breath of God.[23] It seems obvious why Swinton would want to avoid Bulgakov's interpretation. It sounds ableist. It sounds as if only those whose cognitive faculties are intact can be persons. To me, the opposite seems to be actually true. The *I* or the *nous* represents the link with God that allows for the identity of persons with dementia to be preserved and for their experience in the body to be spiritually meaningful.

The *Nous* in the Early Greek Patristic Tradition

When Bulgakov speaks about the *I* coming from God, he only reinterprets in a modern key a very old concept, that of *nous* or intellect. The early Greek Fathers took for granted that the *nous* was part of the constitution of human beings, the divine spark that made humans in the image of God. Although sometimes linked with cognitive capacities, the primary function of the *nous* was to help humans recognise and contemplate God. Without the *nous* that comes from God, humans could have never been able to recognise the divine. The ontological gap between creatures and God is so wide, that without this spark of divinity in them, they would not be able to think God exists, contemplate the divine presence in the creation, or even recognise Jesus as God. Only like can recognise like.

For Maximus the Confessor, the *nous* represents the highest part of the human soul, being capable of "circling around God in a manner beyond knowledge."[24] In paradise, humans lived by contemplating God through their *nous*, while the soul acquired the natural principles of creation and the body drew pleasure from the joy of contemplation. The fall destroyed this harmony: the *nous* lost its direct access to God and became disturbed by the sensations of the body; the body ceased

23 Sergii Bulgakov, *The Lamb of God* (Grand Rapids, MI: Eerdmans, 2008), ebook, 183–84.

24 Maximus the Confessor, *On Difficulties in the Church Fathers: The Ambigua*, ed. and trans. Nicholas Constas, vol. 1, 28–29 (Cambridge, MA: Harvard University Press, 2014), 163.

to find pleasure in contemplation, so it fed the soul with powerful and pleasurable sensations coming from its interaction with the world, making the soul refuse the right principles and ignore the intuition of divine realities coming from the *nous*.[25]

At the end of her survey of the concept of nous in the early Christian tradition, A. N. Williams concluded in a similar key that the *nous*

> functions as a connector, the medium by which we relate to God, the ordering principle of our relation to the complex that is ourselves, and the director of external relations, inasmuch as our moral existence stands at its command; our relation to self and world can be no better than is indicated by our powers of discernment and judgement. The human mind locates us in relation to every other creature, as well as the Creator, and relates every other human faculty and organ to the whole that is the human person.[26]

What I argue here is precisely that the *nous* is better understood as the spiritual intelligence that comes from God and supports human identity-formation through history. The *nous* preexists our temporal embodiment and—although autonomously from the body and its cognitive capacities—uses the embodied experience in order to develop itself into the specific part of the body of Christ designated by God at its creation. What I mean by this is that the growth of the *nous* is not necessarily mediated by conscious cognition. This mediation can become conscious if we learn to pay attention to our thoughts and sensations through various monastic practices, as for instance watchfulness but, for the most part, the *nous* can absorb these sensations bypassing conscious cognition. As Rowan Williams explains,

25 Maximus the Confessor, *On Difficulties in Sacred Scripture: The Responses to Thalassios*, trans. Maximos Constas, The Fathers of the Church 136 (Washington, DC: The Catholic University of America Press, 2018), 91–93; Maximus the Confessor, *On Difficulties in the Church Fathers*, 1: 163.

26 A. N. Williams, *The Divine Sense: The Intellect in Patristic Theology* (Cambridge, UK and New York: Cambridge University Press, 2007), 234.

> Human awareness is initially and primitively just the registering of the image of an object without either meaning or craving attached. So what watchfulness entails is awareness of the moment at which this bare "human" consciousness becomes diabolical, becomes bound to the acquisitive mode of perception.[27]

What this means in the case of persons with dementia is that those moments when they seem to regain awareness, particularly when they receive the Eucharist, hear a certain hymn, or have to recite a certain prayer or even a poem,[28] are moments when the *nous* absorbs the outside stimuli and attempts to order the movement of the bodies in the direction of the ultimate goal of the person: deification. If we follow Maximus' scheme, we can say that when the rational side of the soul fails to fulfil its role for various reasons, from sins to impairment, the *nous* steps in, supplementing its role as much as possible, and attempting to consolidate human spiritual development in the achievement of spiritual identity.

On the one hand, it can be said that intelligence does not exist disembodied in God's memory, because everything happens simultaneously in God so the distinction makes sense only from our human standpoint. On the other hand, we should be able to assume that intelligence can exist in a disembodied state in God, if we take seriously the Roman Catholic and Eastern Orthodox traditions when speaking about the life of the soul after death and until the second coming of Christ.

That the *nous* can be interpreted in this way seems to me highly compatible with the work of Christopher Knight.[29] For Knight,

27 Williams, *Looking East in Winter*, 14–15.

28 Swinton, "What the Body Remembers," 161–62.

29 Christopher C. Knight, "Science, Theology, and the Mind," in *Orthodox Christianity and Modern Science*, ed. Vasilios N. Makrides and Gayle E. Woloschak (Turnhout: Brepols, 2019), 149–61, doi:10.1484/M.SOC-EB.5.116862; Christopher C. Knight, "The Human Mind in This World and the Next: Scientific and Early Theological Perspectives," *Theology and Science* 16:2 (2018): 151–65, doi:10.1080/14746700.2018.1455265; Christopher C. Knight, *Eastern Orthodoxy and the Science-Theology Dialogue* (Cambridge: Cambridge University Press, 2022), doi:10.1017/9781009106009; Christopher C. Knight, "Have a Bit of

"The notion that we nowadays tend to think of as constitutive of our minds and personalities—our discursive rational faculty and memory—may in fact be no more than servants, in this world, of something more central to our being: what in Greek is termed *nous*."[30] Knight even goes as far as to suggest that, in the eschatological state, humans might not use their cognitive capacities to think, but that they

> may simply *know*—directly and intuitively—in the way that mystics, in their most sublime moments, are said to *know*. Our eschatological state in its mental dimension will not simply be a continuation of our personalities and mental properties as we experience them now. Rather, it will involve true continuity in our existence as unique persons, it will involve a transformation of our whole being.[31]

Knight helps expand Bulgakov's interpretation of the patristic tradition, namely that a certain part of the human being—our spiritual intelligence—is not fully determined by our embodiment and even by our cognitive faculties. This also clarifies and enriches Swinton's account of how human spiritual identity can exist in God and is bestowed upon us in the Holy Spirit: it is the spiritual intelligence that interacts with our embodied identity and that guides us to the fulfilment of our identity in Christ even during one's period with dementia.

To me, this last point is important for its ramifications for a closely related debate: the euthanasia of persons with dementia. If the theological response is simply that we should keep living in order to remain in relationship with God, but without explaining how and why it is important to remain in this relationship, then the question that follows is why this relationship has to continue here and not in the afterlife. Being in a temporal relationship with God when our bodies

Nous: Revelation and the Psychology of Religion," in *Mutual Enrichment between Psychology and Theology*, ed. Russel Re Manning (London: Routledge, 2020), 47–60.

30 Knight, "Science, Theology, and the Mind," 161 n. 35.

31 Knight, "Science, Theology, and the Mind," 161 n. 35.

and cognitive faculties are irremediably decaying does not make too much sense if we could continue this relationship in the afterlife, where we will be remembered by God with our true identity, our cognitive capacities restored, our bodies in perfect shape, and surrounded by our beloved friends and family. By talking about the *nous* as the spiritual intelligence that guides us towards the full development of our spiritual identity in Christ, even when our cognitive faculties are impaired, it can be then contended that even the period with dementia can have the potential of being spiritually formative and for this reason, we should leave it to follow its course until God considers that our *nous* reached its full development in this temporal life.

Conclusions

What I wanted to do in this article is argue that the metaphor of living in the memory of God used by John Swinton in order to defend the personhood of those who develop dementia entails more theological presupposition than one might think. It entails the creation of human *nous* at once with the beginning of time and its gradual entry into history, as well as the existence of a spiritual faculty that keeps us in relationship with God irrespective of the state of our cognitive faculties or our bodies, and which then uses the information it receives from these for the development of our spiritual identity.

Deepening Swinton's metaphor also provides a clearer understanding of how our relationship with God continues during the period with dementia and provides a more robust grounding for rejecting the euthanasia of persons with dementia or profound intellectual disabilities.

I am sure there is much more to be said in support of my argument, as there are many more adjacent areas I did not have the space to cover at length, as for example: the interrelationship between *nous* and the body or the materiality of the world more broadly conceived, or the advantages or disadvantages of euthanasia.

My intention was not to be exhaustive, but rather to bring into conversation several theological threads that usually ignore each other. Science and theology, disability, and Eastern Orthodoxy tend to be treated on their own and, in my opinion, they can be the source of some exciting insights if brought together.

The author reports there are no competing interests to declare.
Received: 01/02/24 Accepted: 07/07/24 Published: 04/03/25

Spiritual Intelligence and the *Nous*: Implications for Understanding the Relationship Between the Faith Traditions of the World

Christopher C. Knight

Abstract: The modern concept of spiritual intelligence exhibits parallels with the ancient Greek philosophical understanding of the *nous*. This ancient understanding was used extensively in late antique and early medieval theological thinking and is still influential in some faith traditions. However, the implications of this understanding for exploring questions about religious pluralism have not been widely acknowledged. These implications arise from the way in which the *noetic* perception that arises from the full functioning of the *nous* is seen as essentially intuitive in nature, so that the relationship between this perception and religious doctrinal statements may be understood in terms of a radically apophatic understanding of religious language usage. Vladimir Lossky has proposed this stance as characteristic of Eastern patristic perspectives. In relation to this understanding, parallels with the perennialist tradition are evident and—even though this tradition in its classic form exhibits major flaws that need to be corrected—its pluralism becomes highly suggestive. This suggestiveness is reinforced by a number of other considerations, not least the recent "theological turn" in discussion of divine action within the science-theology dialogue, which permits an essentially "naturalistic" understanding of revelatory experiences. There may, nevertheless, be reasons for adopting "reciprocal inclusivism" rather than a full-blown pluralism of this kind.

Keywords: apophaticism; esoteric ecumenism; nous; perennialism; religious pluralism; spiritual intelligence

Christopher C. Knight is an Eastern Orthodox priest and a senior research associate of the Institute for Orthodox Christian Studies in Cambridge, England. He is the author of numerous papers and book chapters, and of five books, including Science and the Christian Faith: A Guide for the Perplexed (2020).

Spiritual intelligence is a concept that is often approached through scientific or psychotherapeutic considerations in much the same way as is the concept of emotional intelligence. My intention in this paper is not, however, to contribute to these lines of enquiry but to outline some *theological* perspectives that may be related to the concept of spiritual intelligence and contribute to our exploration of it. At the heart of these theological perspectives is the way in which the concept of spiritual intelligence may be seen as echoing aspects of the understanding that ancient Greek philosophers—and after them the Greek-speaking Christian theologians of the patristic and later periods—expressed in terms of what they called the *nous*. (This concept has many other theological implications, and I have outlined these implications elsewhere.)[1]

This word *nous* is often translated into English as "intellect," partly because of its early translation into Latin as *intellectus*. This English translation is, however, potentially misleading to present day readers because the English term "intellect" is often now understood as the seat of discursive reasoning. However, the term *nous*—at least in philosophical usage[2]—refers to something quite different: to an essentially intuitive faculty that enables discernment of what is true or real. Indeed, especially in its theological usage in the patristic period, it may be seen as anticipating the general meaning that is now often associated with the term spiritual intelligence.

---

1 For an analysis of the *nous* concept in terms of both these theological implications and modern scientific understandings, see Christopher C. Knight, "The Human Mind in This World and the Next: Scientific and Early Theological Perspectives," *Theology and Science* 16:2 (2018): 151–165, https://doi.org/10.1080/14746700.2018.1455265.

2 The term *nous* was common enough among Greek speakers of the ancient and late antique world to be used in their everyday speech in a way that did not always fully reflect philosophical usage. In the New Testament, for example, it was used quite often in the letters of Paul. Some regard Paul's usage as reflecting ancient philosophical understanding, at least in some degree, while others claim that his use of the word *nous* related more to what was referred to in later writings as *dianoia*, the discursive rational faculty. However, the view that one takes on this issue of proper exegesis of Paul's usage does not affect what follows, since it does not rely on any particular New Testament exegesis but on the philosophical perspectives that informed later patristic usage.

There are, admittedly, several distinct, if related, understandings of the *nous* to be found in early Christian authors, due in part to the ways in which they took up one or other of the different nuances of the term to be found in the works of Aristotle, of Plato, and of the Neoplatonists. Nevertheless, the concept of the *nous* was widely used by these authors in relation to its perceived functions as "a connector, the medium by which we relate to God, the ordering principle of our relation to the complex that is ourselves, and the director of external relations, inasmuch as our moral existence stands at its command."[3]

An aspect of this use of the concept by early Christian authors was a sense that the *nous* should be seen as the organ of a kind of contemplation that transcends discursive thinking. Indeed, in many strands of Christian thinking, it was seen as central to the relationship between the human person and God: the point at which the human mind is in some sense in direct contact with the divine mind. In the patristic roots of modern Eastern Orthodox understanding, for example, faith itself was often seen as related to the *nous*,[4] and in general the *nous* was seen as "the highest faculty in man, through which—provided it is purified—he knows God or the inner essences of created things by means of direct apprehension or spiritual perception."[5] The full noetic perception to which the unfettered use of the *nous* gives rise was, however, regarded as being at least partially eclipsed in "fallen"

3 A. N. Williams, *The Divine Sense: The Intellect in Patristic Theology* (Cambridge: Cambridge University Press, 2007), 234.

4 In relation to Gregory of Nyssa's understanding, for example, see Martin Laird, *Gregory of Nyssa and the Grasp of Faith: Union, Knowledge, and Divine Presence* (Oxford: Oxford University Press, 2004). Gregory did not use the term faith (*pistis*) as it had been used in much early Greek philosophy, in which it had denoted the lowest form of knowledge. Instead, as Laird puts it, while Gregory uses notions to be found in the work of the Neoplatonist Plotinus, he nevertheless ascribes to faith "qualities which Neoplatonism would reserve for the crest of the wave of *nous*" (2).

5 Bruce V. Foltz, *The Noetics of Nature: Environmental Philosophy and the Holy Beauty of the Visible* (New York, NY: Fordham University Press, 2014), 248–249. (The implicit reference here is to the understanding that was most highly developed in the work of Maximus the Confessor.)

humanity, and this eclipse was seen as remediable only through spiritual practice.[6]

This understanding is reflected in the way in which, in the Greek vocabulary employed in patristic (and modern Eastern Orthodox) use of the *nous* concept, different words are used for different kinds of knowledge, which are seen as arising from different kinds of mental and spiritual activity. Especially in the hesychastic[7] understanding that became highly influential in Eastern Orthodox thinking, full knowledge of God must be based on contemplation (*theōria* in Greek) which is seen as the direct perception or vision by the *nous*.[8] This faculty is not the same as the discursive reasoning faculty (*dianoia),* which may have a role to play in overcoming the partial eclipse of the *nous* in "fallen" humanity but is nevertheless understood as functioning adequately in theological analysis only if rooted in the spiritual knowledge (*gnōsis*) obtainable through direct apprehension by the *nous*.[9] Without this rooting, there is a significant danger that the concepts we form "in accordance with the understanding and the judgement which are natural to us, basing

6 Here, we need to recognise that there is a difference between the dominant interpretations of the effects of the Fall in the Eastern and Western parts of the Christian world. In the West, the Augustinian notions were highly influential, in contrast with the less pessimistic understandings of the East, so that the role of the *nous* and of its relationship to discursive reasoning tend to be seen in different ways. See the discussion in Christopher C. Knight. "Natural Theology and the Eastern Orthodox Tradition," in *The Oxford Handbook of Natural Theology*, ed. Russell Re Manning (Oxford: Oxford University Press, 2013), 213–226.

7 This word, *hesychast*, deriving from the Greek term for silence or stillness, refers to the understanding of contemplative practice which—especially since its defence by Gregory Palamas in the fourteenth century—has been dominant within Orthodoxy and particularly in its monastic practice.

8 This term *theōria* is, admittedly, used in some strands of patristic thinking in a different way that relates to discursive thinking, so that it is specifically the hesychastic strand of thinking to which I refer in what follows.

9 See, e.g., the brief discussion of all these terms given in the "glossary" section of *The Philokalia*, vol.1, ed. G. E. H. Palmer, Philip Sherrard, and Kallistos Ware (London and Boston: Faber and Faber, 1979), 357–367. There may, however, be a tendency in this glossary to suggest a uniformity of usage that is in fact not to be found in the texts of *The Philokalia*, which is an anthology of texts from many different writers.

ourselves on an intelligible representation, create idols of God instead of revealing to us God Himself."[10]

Linked to this understanding is the kind of apophaticism that scholars like Vladimir Lossky present as central to the Eastern Christian understanding of theological language usage.[11] In this "mystical" understanding, there is a strong sense that the terms used in religious language can never circumscribe the realities towards which they attempt to point. This understanding is often presented as constituting a "negative theology" that focuses on saying what God is not, rather than on what God is, and this certainly reflects part of its meaning. To speak of this apophaticism only in terms of negative theology is, however, potentially problematical because these terms can be understood in different ways.[12] It is important to state straightaway, therefore, that the kind of apophaticism on which I shall focus in what follows is essentially of the radical (and somewhat controversial)[13] kind that Lossky propounds. As he himself has stressed, this version of apophaticism is not to be understood only in terms of the distinction between the theological path that it offers and the path of *cataphatic* or positive theology, which proceeds by affirmations rather than negations. The more radical form of apophaticism that he advocates, and which I

10 Vladimir Lossky, *The Mystical Theology of the Eastern Church* (Cambridge: James Clarke, 1957), 33 (paraphrasing Gregory of Nyssa, *Life of Moses* 2.165).

11 Lossky, *The Mystical Theology of the Eastern Church*.

12 Aydogan Kars, *Unsaying God: Negative Theology in Medieval Islam* (New York, NY: Oxford University Press, 2019), has commented that negative theology, when "*unqualified* is also *disqualified*" (14). Kars makes this point in relation to Islamic negative thinking, on which his study concentrates, but his point is valid in relation to other traditions as well.

13 Patristics scholars, especially in recent decades, have seen the sometimes too strong a stress on the apophaticism of certain patristic writers as ignoring some of the nuances to be found in their writings, including those of Gregory of Nyssa, on whom Lossky puts significant emphasis. See, e.g., Andrew Radde-Gallwitz, *Basil of Caesarea, Gregory of Nyssa, and the Transformation of Divine Simplicity* (Oxford: Oxford University Press, 2009), in which it is asserted that we should see Basil and Gregory less as "mystics devoted primarily to the *via negativa*" and more as "subtle thinkers devoted to preserving the coherence and consistency of the myriad positive affirmations of Christian scripture and worship, while nonetheless acknowledging the ultimate incomprehensibility of God" (vii).

shall expand in what follows, is what he calls "an attitude of mind which refuses to form concepts about God."[14] This kind of apophaticism is not merely "saying what God is not" because—as Lossky's fellow-Orthodox, Olivier Clément, has stressed—this strand of Orthodox understanding is one in which "negation is denied just as much as affirmation."[15]

This apophatic understanding has parallels in several non-Christian faith traditions,[16] and among its implications is one that has hitherto largely been ignored. This is the possibility that it provides new ways of analysing attitudes towards faith traditions other than one's own.[17] As we shall see, it provides, not only a way of rejecting the kind of exclusivism that can see no validity in faith traditions other than one's own, but also of modifying the separation that is usually assumed

14 Lossky, *The Mystical Theology of the Eastern Church*, 38–39.

15 Olivier Clément, *The Roots of Christian Mysticism: Text and Commentary* (London: New City, 1993), 31.

16 For example, Marco Pallis, in his book, *A Buddhist Spectrum: Contributions to Buddhist-Christian Dialogue* (Bloomington, IN: World Wisdom, 2003), has stressed that the Buddhist tradition's reluctance to speak of God (or even of the self) should be understood in terms of the "apophatic method which Buddhism favours" (131). In a comparable way, strands of Islamic thinking manifest an apophatic attitude. In Islam's Shi'ite strand of thinking, for example, negative theology is related to a sense of the unknowability of God's *essence* that is comparable to the similar stress that exists within Eastern Orthodox Christianity (in which the distinction between God's *essence* and *energies* became especially influential through the work of Gregory Palamas in the fourteenth century, though the distinction can be found much earlier). The Arabic term for "negative theology," *lahoot salbi*, the practice of which involves the use of *ta'til*, which means "negation," is related in Shi'ite teaching to the way in which God is seen in terms of "two ontological levels: first, of the Essence (*ḏāt*). This is said to be forever inconceivable, unimaginable, above all thought, beyond all knowledge. It can only be described by God through revelations and can only be apprehended by a negative apophatic theology … However, if things were to remain so, no relation would be possible between the Creator and His creatures. Thus God, in his infinite grace, lets blossom in his own being another level: of Names and Attributes (*asmāʾ wa ṣefāt*) by which He reveals himself and makes himself known. This revealed level, recalling the *Deus revelatus* of Christian theology, is no longer God the Unknowable, but God the Unknown who aspires to be known. It is the exoteric, manifest, revealed level of God that can be known in Him." ("Shi'ite Doctrine," in *Encyclopaedia Iranica*, http://www.iranicaonline.org/articles/shiite-doctrine; accessed 15 April 2020).

17 Much of what follows is examined in greater detail in Christopher C. Knight, *Exploring Religious Pluralism: From Mystical Theology to the Science-Theology Dialogue* (Cambridge: Cambridge University Press, 2024).

to exist between inclusivism—which sees other faith traditions as holding only incomplete or distorted versions of the "truths" proclaimed by one's own tradition[18]—and religious pluralism, which sees other faith traditions as being of equal validity to one's own.[19]

Pluralism and the "Truth Claims" of Different Faith Traditions

In relation to this spectrum of opinions, one of Lossky's observations is of considerable interest. This is his contention that a radically apophatic approach to theology implies acceptance of a degree of apparent logical inconsistency—what is sometimes called *antinomy*—that contemporary analytic philosophy would usually reject. As he has put it,

> theology will never be abstract, working through concepts, but contemplative: raising the mind to those realities which pass all understanding. This is why the dogmas of the Church often present themselves as antinomies … It is not a question of suppressing the antinomy by adapting dogma to our understanding, but of change of heart and mind enabling us to attain to the contempla-

18 This inclusivist position is perhaps best known in the Christian world through the thinking of Karl Rahner. For a summary of Rahner's thinking on this topic, see Jeannine Hill Fletcher, "Rahner and Religious Diversity," in *The Cambridge Companion to Karl Rahner*, ed. Declan Marmion and Mary E. Hines (Cambridge: Cambridge University Press, 2005), 235–248.

19 Perhaps the best known (though by no means the only) version of religious pluralism is that of John Hick, as set out in his book, *An Interpretation of Religion: Human Responses to the Transcendent* (London: Macmillan, 1989). Hick argues that all the great faith traditions may be seen as authentic responses to what he calls *Reality*. His "pluralistic hypothesis" is essentially that this Reality is ineffable and beyond adequate comprehension, but that the presence of this Reality can be experienced through the different linguistic systems and spiritual practices offered by various faith traditions. He sees Kant's distinction between things as they are in themselves (*noumena*) and things as they are experienced (*phenomena*) as applicable to this Reality, so that a person's experience of Reality will depend on the interpretative frameworks and structures through which that experience is comprehended.

> tion of the reality which reveals itself to us as it raises us to God, and unites us, according to our several capacities, to Him.[20]

This emphasis on contemplation and the role of antinomy[21] would appear to be applicable—in a way that Lossky himself does not consider—to many of the philosophical arguments sometimes used to attempt to refute a pluralistic understanding.[22] In these arguments, incompatibilities between the doctrinal "truth claims" of different faith traditions are stressed in order to conclude that pluralism is incoherent because no more than one of these "competing" truth claims can be true. If we accept Lossky's antinomic approach to theology, however, then apparent incompatibilities of this kind cannot automatically be seen as definitive for assessing compatibility at a deeper, contemplative level.[23]

The point here is that when we expand Lossky's antinomic and contemplatively focused understanding to differences between the doctrinal frameworks of the various faith traditions of the world, these doctrinal frameworks may—at least in principle—be seen as something other than as sets of "truth claims" of the abstract kind often assumed by analytic philosophers of religion.[24] They may be seen, instead, as

20 Lossky, *The Mystical Theology of the Eastern Church*, 43.

21 For an interesting analysis of the way in which aspects of Lossky's approach may not be purely patristic in origin, see Brandon Gallaher, "The 'Sophiological' Origins of Vladimir Lossky's Apophaticism," *Scottish Journal of Theology* 66:3 (2013): 278–298, https://doi.org/10.1017/S0036930613000136.

22 I first argued this point in Christopher C. Knight, "Reciprocal Inclusivism: A Methodology for Understanding the Faiths of the World," *Journal of Ecumenical Studies* 55:4 (2020): 609–629, https://doi.org/10.1353/ecu.2020.0048.

23 It may be that patristic theology cannot be said to *entail* this pluralistic possibility, since different patristic writers had different views on the importance of apophaticism and—as observed in n. 13—different patristics scholars of the present day have different views on the extent to which Lossky's radical apophaticism should be seen as a legitimate overarching interpretative principle for understanding patristic writings. My argument is not that patristic theology entails the conclusions to which I come, about the validity of religious pluralism, but that patristic theology may be seen as compatible with a pluralistic understanding when sufficient weight is given to its apophatic component.

24 As John Cottingham has put it, "analytic philosophers are prone to use the

what Lossky calls "images or ideas intended to guide us and fit our faculties for the contemplation of that which passes all understanding."[25] They may, in other words, be seen as relating to noetic apprehension rather than to discursively developed understanding.

For pluralists, this understanding may, I would argue, be linked straightforwardly to the well-known "one mountain, many paths to the summit" analogy, in which the various spiritual pathways provided by different faith traditions are seen as beginning from different starting points but ending at the same destination. This analogy points to the way in which, because they start from different cultural "locations," different spiritual pathways inevitably require different "signposts" as guides. These signposts may be seen as functioning, not primarily at the conscious, discursive level of the mind, but at the deeper, intuitive level that relates to the *nous*. Their role is—through their use in meditative, sacramental, or liturgical contexts—to serve as guiding "methods" or "means" that are appropriate to the particular contemplative pathways to which they relate.

The Perennialist Tradition, Neo-Perennialism, and Esoteric Ecumenism

This language of "means" or "methods" is not, admittedly, usually associated with Lossky's understanding, and it is unclear how far he himself would have been willing to see his perspectives as indicat-

'fruit-juicer' method" of looking at words in isolation from the total context in which they are used, requiring "the clear liquid of a few propositions to be extracted for examination in isolation from what they take to be the irrelevant pulpy mush of context." John Cottingham, "The Lessons of Life: Wittgenstein, Religion, and Analytic Philosophy," in *Wittgenstein and Analytic Philosophy: Essays for P. M. S. Hacker*, ed. Hans-Johann Glock and John Hyman (Oxford: Oxford University Press, 2009), 209. This means, among other things, that—as another scholar has observed—these philosophers often have "a tin ear for possibilities of sense, especially with regard to religions or cultures very different to those with which they are familiar." Mike Burley, "Reincarnation and the Lack of Imagination in Philosophy," *Nordic Wittgenstein Review* 5:2 (2015): 39–64, at 40.

25 Lossky, *The Mystical Theology of the Eastern Church*, 40.

ing the plausibility of a pluralistic understanding.[26] Nevertheless, his understanding is comparable to that to be found in another kind of mystical understanding, which does use this kind of language. This is the understanding of the pluralistic school of thought associated with the work of scholars like René Guénon and Frithjof Schuon, which is sometimes referred to as the Traditionalist school, sometimes as perennial traditionalism, and sometimes simply as perennialism.

While often associated primarily with certain Islamic scholars, in practice this school of thought has followers in many different faith traditions, including Christian scholars such as the Methodist Huston Smith,[27] the Roman Catholic Jean Borella,[28] and the Eastern Orthodox James Cutsinger.[29] While these Christian authors express their views in slightly different ways, they all reflect the perspective articulated by the perennialist writer William Stoddart in his "Foreword" to a multi-author collection of perennialist essays on Christianity:

> The perennial philosophy—which is true universalism and true ecumenism—is, at least extrinsically, a recognition of the divine origin of each religion. The essence of each religion is pure truth.

26 Lossky seems to have made little reference to other faiths. See the comments of Paul Ladouceur, "Religious Diversity in Modern Orthodox Thought," *Religions* 8:5 (2017): 77, https://doi.org/10.3390/rel8050077. But it is noteworthy that Olivier Clément, who had comparable views on apophaticism, did have a general interest in interfaith dialogue, speaking of the need to listen "in order to understand, and not dismiss with the back of our hand" and noting approvingly the attitude of the Orthodox missionary Spiridon Kislyakov, who "used to say that he held the Buddhist sages in such high esteem that he hardly dared to speak to them of baptism!" See "Orthodoxy and the Mystery of the Person: Interview with Olivier Clément," available at https://tinyurl.com/7wnd8c4z (accessed 15 May 2022).

27 See, e.g., Huston Smith, *Forgotten Truth: The Common Vision of the World's Religions* (San Francisco: Harper Collins, 1976).

28 See, e.g., Jean Borella, *Guénonian Esoterism and Christian* Mystery (Hillsborough, NY: Sophia Perennis, 2005).

29 See, e.g., James S. Cutsinger, *Advice to the Serious Seeker: Meditations on the Teaching of Frithjof Schuon* (Albany, NY: State University of New York Press, 1997). In Knight, *Exploring Religious Pluralism*, I have suggested that a number of other Orthodox writers—such as Philip Sherrard and Robin Amis—while not strongly influenced by classic perennialism, have nevertheless developed what I call "quasi-perennialist" perspectives.

> And the various religions clothe that truth in garments of different designs and colors. 'In my Father's house are many mansions.' This saying of Christ's applies not only to Heaven, but also to earth. The function of the various religions is to express the truth, and to offer a way of salvation, in a manner suited to the different segments and ethnicities of mankind. Each religion comes from God and each religion leads back to God. Each religion, moreover, comprises a doctrine and a method, that is to say, it is an enlightening truth coupled with a saving means.[30]

While I believe that we should be highly critical of certain aspects of classic perennialism,[31] I see in it, nevertheless, a number of positive characteristics. One of these relates directly to what I have said about the *nous* because perennialists emphasise the way in which—in the ancient traditions that they view as authentic—the human person is seen as composed of three levels of being: spirit, soul, and body. (In Greek, for example, ancient and medieval writers spoke of *pneuma* or *nous*, *psyche*, and *soma*; in Latin, they spoke of *spiritus* or *intellectus*, *anima*, and *corpus*; and in Arabic of *rûh, nafs*, and *jism*.) While the differences between traditions in their use of this threefold taxonomy are often insufficiently acknowledged by perennialists, what is relevant to our present exploration is their use of it to point to the importance of a capacity that they usually associate with the "spirit" component of what it is to be human. Like those who stress the *nous*

30 William Stoddart, "Foreword," in *Ye Shall Know the Truth: Christianity and the Perennial Philosophy*, ed. Mateus Souras de Azevedo (Bloomington, IN: World Wisdom, 2005), x–xi. Note that there is one aspect of this description that I question in chapter 8 of my *Exploring Religious Pluralism*. This is its focus on "ethnicities," which I suggest should be replaced by a focus on cultural diversity; see my comments later in this paper on the notion of the "psycho-cultural niche."

31 Two of my objections are ones that many other critics of classic perennialism have voiced: that it exhibits distortions that arise from nostalgia for a fictional past, especially in relation to Guénon's notion of an ancient and now partially lost "primordial tradition," and that it has a strong tendency to impose an interpretative framework on historical and empirical evidence in a questionable way. A third objection—less often voiced but important for my own approach—is that it tends to ignore the natural world.

in a Christian or Islamic[32] context, perennialists see this capacity, not as the seat of discursive, rational thinking, but as something that operates at a deeper, intuitive level. This understanding leads them to stress that perceiving the "truth," which they see as being at the heart of all authentic faith traditions, involves an essentially intuitive kind of spiritual intelligence; they believe, with Guénon, that true metaphysics "constitutes an immediate, or in other words, intuitive knowledge, as opposed to the discursive and mediate knowledge that belongs to the rational order."[33]

This aspect of perennialist understanding can, in my judgement, be retained in a kind of *neo-perennialism*, in which classic perennialism's genuine insights (as I see them) can be retained while, partly through attentiveness to aspects of current religious studies, its flaws can be discarded.[34] One of the things to be retained in such a neo-perennialism is classic perennialism's disdain for modern philosophy of the analytic kind, especially when it is applied to religious doctrines. This attitude—which exhibits parallels with the expansion of Lossky's approach that I have outlined—is rooted in the belief that, what philosophers in the analytic tradition usually take to be the "truth claims" of the doctrinal languages of the world's faith traditions, are in fact no more than what perennialists call the *exoteric* aspects of those traditions.

These exoteric aspects are seen by perennialists as constituting part of the "method" or "means" by which adherents of different traditions are guided along the particular spiritual paths that have been developed in those traditions towards the goal of full noetic insight. What perennialists see as important is the way in which—as one makes progress along any one of these spiritual pathways—one will

32 In the Islamic world, the influence of a Neoplatonic understanding of the *nous* is often perceived by modern scholars in the work of early Islamic philosophers like Al Farabi, Avicenna, and Ibn Rushd.

33 René Guénon, *The Essential René Guénon: Metaphysics, Tradition, and the Crisis of Modernity*, ed. John Herlihy (Bloomington, IN: World Wisdom, 2009), 105.

34 See Knight, *Exploring Religious Pluralism*. For an earlier articulation of this viewpoint, see Christopher C. Knight, "Neo-Perennialism: A Trap to Avoid or a Valid Research Programme?" *Journal of Ecumenical Studies* 58:1 (2023): 60–85, https://doi.org/10.1353/ecu.2023.0003.

increasingly apprehend these doctrines' *esoteric* meaning in a noetic manner. Because this esoteric meaning is apprehended intuitively rather than discursively, it is not, for perennialists, to be understood in terms of apparently competing "truth claims." Rather, this esoteric meaning may be seen as identical in all authentic faith traditions. For this reason, their understanding of pluralism is sometimes labelled by them as *esoteric ecumenism*.[35]

Divine Action Theories and Their Relevance

As we have seen, the kinds of mystical understanding expounded by Lossky and the perennialists manifest overlapping understandings of spiritual intelligence. Analysis of these parallels need not, however, be limited to observation of this overlap, since further exploration is possible in terms of several other considerations.[36] One of these considerations relates to the question—central to the science-theology dialogue of the late twentieth and early twenty-first centuries—of how God is to be understood as acting in a world characterised by obedience to "laws of nature." Here, I would argue, the kind of pluralism that I have outlined may be strengthened by an account of divine action which differs significantly from the "causal joint" model that has, until very recently, been dominant within that dialogue. In this latter model, while God is seen as always acting "in, with, and under" the laws of nature, there is still a clear distinction between the "general divine action" that occurs through the normal operation of those laws and the "special divine action" that is seen as arising from God's direct "response" to situations in the world.

35 See, e.g., Frithjof Schuon, *Christianity/Islam: Essays on Esoteric Ecumenism* (Bloomington, IN: World Wisdom, 1985); James S. Cutsinger, "Hesychia: An Orthodox Opening to Esoteric Ecumenism," in *Paths to the Heart: Sufism and the Christian East*, ed. James S. Cutsinger (Bloomington, IN: World Wisdom, 2002).

36 Not all of these can be mentioned in this paper, but they are set out in Knight, *Exploring Religious Pluralism*.

This causal joint model is now increasingly being questioned, partly because of Nicholas Saunders' critique of it[37] and partly because of what Sarah Lane Ritchie calls a "theological turn" in recent discussion of divine action,[38] which has three independent but conceptually linked components. This conceptual linkage arises from the way in which, in all three, the causal joint model's distinction between "special" and "general" modes of divine action is blurred or even abolished.

In my own contribution to this theological turn,[39] I argue that we may see all events—including miraculous ones—in terms of an "enhanced naturalism" comparable to that which, in the patristic era, was hinted at by Augustine of Hippo, as Wolfhart Pannenberg has pointed out.[40] (Indeed, I argue, this kind of understanding is reinforced when we expand it in terms of aspects of the thinking of Maximus the Confessor.)[41] In this kind of naturalism, a distinction is made between "ordinary" laws of nature, which are susceptible to investigation

37 Nicholas Saunders, *Divine Action and Modern Science* (Cambridge: Cambridge University Press, 2002).

38 This "theological turn" was first discussed in Sarah Lane Ritchie, "Dancing Around the Causal Joint: Challenging the Theological Turn in Divine Action Theories," *Zygon* 52:2 (2017): 361–379, https://doi.org/10.1111/zygo.12336, the contents of which were modified and expanded in an important book: Sarah Lane Ritchie, *Divine Action and the Human Mind* (Cambridge: Cambridge University Press, 2019).

39 This contribution was first developed in Christopher C. Knight, *Wrestling with the Divine: Religion, Science, and Revelation* (Minneapolis: Fortress, 2001) and expanded in terms of Christian incarnational insights in Christopher C. Knight, *The God of Nature: Incarnation and Contemporary Science* (Minneapolis: Fortress, 2007).

40 See Wolfhart Pannenberg, "The Concept of Miracle," *Zygon* 37:3 (2002): 759–762, https://doi.org/10.1111/1467-9744.00452.

41 Maximus the Confessor developed what may be seen as a kind of theistic naturalism by linking the fourth gospel's notion of the divine *Logos* to the principles (*logoi*) through which all created things have their being and act as they do. My model takes up the teleological aspect of Maximus' thinking, arguing that this may be applied to divine action in a way that does not compete with the "naturalistic" perspectives of science but simply interprets them theologically. In a paper with a word limit, however, it is not possible to explain this model adequately, so the reader of this paper must be referred to Knight, *The God of Nature* and to the paper in which the model was first set out: Christopher C. Knight, "Divine Action: A Neo-Byzantine Model," *International Journal for Philosophy of Religion* 58 (2005): 181–199, https://doi.org/10.1007/s11153-005-1076-5.

through the scientific methodology, and "higher" laws of nature, which are not susceptible to this methodology because they are not manifested in events or behaviours that are straightforwardly repeatable. In this understanding, God is not seen as "responding" to situations in the world because, instead of being seen as a temporal being, God is understood in terms of divine eternity. This is not only traditional—in the sense of reflecting the understandings of late antique and medieval philosophical theology[42]—but is also consonant with both mystical apprehension[43] and our current scientific understanding of time as an aspect of the created world.[44]

If we expand this "single act" understanding of divine action[45] to include God's action in revelatory experiences, in what we might call a "single act of revelation," we can reinforce the kind of openness to religious pluralism that arises from ancient understandings of spiritual intelligence. Such an expansion allows our exploration of the faith traditions of the world to focus, with greater clarity than hitherto, on the roles of "natural" human religiosity and psychology in our analysis of the experiences through which those traditions have arisen.

---

42 This subtle view was expressed by Thomas Aquinas in terms of what has been called the "classical view of divine eternity." See Brian Davies, *An Introduction to the Philosophy of Religion* (Oxford: Oxford University Press, 1993), 141. It was expressed in an even more subtle way by Maximus the Confessor. See Sotiris Mitralexis, *Ever-Moving Repose: A Contemporary Reading of Maximus the Confessor's Theory of Time* (Eugene, OR: Cascade, 2017).

43 As one commentator on mystical experience has noted, "the mystic feels himself to be in a dimension where time is not, where 'all is always now'," so that such an experience is not understandable "unless one is prepared to accept that there may be an entirely different dimension from that of clock time or indeed of any other sort of time." F. C. Happold, *Mysticism: A Study and Anthology* (Harmondsworth: Pelican, 1963), 48.

44 One of the main differences between Newtonian mechanics and its replacement—Einstein's relativistic mechanics—is the way in which Newton saw space and time as absolutes within which the universe unfolds, while Einstein saw them simply as aspects of the created order, which were in fact interdependent in a way that means that distances and time intervals between events may be different for different observers.

45 This particular "single act" account is, it should be noted, very different to the "single act" account of Maurice Wiles, in which miraculous events are seen as impossible. See the comments in Knight, *Exploring Religious Pluralism*, 173–177.

This approach allows us, I believe, to develop what I call a non-reductionist "psychological-referential model of revelatory experience," which may be understood in an evolutionary context by analogy with the well-known concept of the ecological niche. As I have written elsewhere, this analogy allows us to see how we may use a related concept—that of the "psycho-cultural niche"—"which may be defined by both the cultural assumptions and the individual psychological makeup of those able to experience some religious revelation or enlightenment." A particular psycho-cultural niche "provides the necessary psychological environment for some particular revelation to arise, and also limits the type of experience that could arise and flourish, in a way analogous to that in which a particular ecological niche allows only certain new biological species to emerge and spread."[46]

This framework allows us, I argue, to develop a plausible understanding of the origin and development of different faith traditions in terms of a set of five theses that rely on no particular faith tradition. These theses are as follows:

- The human psyche may be understood in principle entirely in terms of the development of the cosmos through natural processes from the Big Bang to the evolutionary emergence of specifically human qualities.
- All experiences that give the impression of being revelatory of a divine reality are the spontaneous, natural products of the human psyche, and do not require any notion of "special" divine action to explain them. These experiences are culturally conditioned, in that their specific forms will relate to both the individual psychological make-up and culturally determined expectations of those who receive them. These factors are sufficient to explain why, in different individuals and cultural contexts, there is considerable diversity in the types of such experiences and of the religious languages that arise from them.

46 Knight, *Wrestling with the Divine*, 112.

- The belief of most religious people, that their own faith's foundational revelatory experiences have given rise to a religious language that is genuinely referential to a divine reality, is a valid one. This divine reality—as something to which reference can validly be made—is therefore ontologically defensible.
- The diversity of the religious languages that arise from different revelatory experiences does not necessarily imply that they cannot all validly refer to the divine reality. A pluralistic understanding of their referential success is possible.
- The cosmos, in which the revelation-oriented human psyche has arisen naturalistically, is attributable to the "will" or character of the divine reality to which authentic revelatory experience bears witness. (As those of the Abrahamic traditions might put it, the probability that some creatures would come to know their creator was built into the cosmos, by that creator, from its very beginning.)

There are, of course, tensions between these theses, since the first two are fundamentally naturalistic while the remainder take the view that theological language can be truly referential. Nevertheless, in the essays in which I first articulated these theses,[47] I argued that these tensions can be overcome, and since then I have developed other supporting arguments from a wide range of considerations.[48]

47 These theses were first articulated in Christopher C. Knight, "*Homo Religiosus*: A Theological Proposal for a Scientific and Pluralistic Age," in *Human Identity at the Intersection of Science, Technology and Religion*, ed. Nancey Murphy and Christopher C. Knight (Farnham: Ashgate, 2010), 25–38. I repeated these theses in slightly different contexts, in "Biological Evolution and the Universality of Spiritual Experience: Pluralistic Implications of a New Approach to the Thought of Teilhard de Chardin," *Journal of Ecumenical Studies* 48:1 (2013): 58–70, and in "Have a Bit of *Nous*: Revelation and the Psychology of Religion," in *Mutual Enrichment: Theology, Psychology and Religious Life*, ed. Russell Re Manning (London: Routledge, 2020), 47–60.

48 Among these newer arguments—presented in Knight, *Exploring Religious Pluralism*—are expansions of the somewhat different notions of "archetypes" articulated by Carl Jung and Mircea Eliade, and the analyses of the role of

One of the most important of these considerations arises from the recent focus, within the psychology of religion, on what Fraser Watts calls a "dual-process" understanding of human cognition.[49] In this understanding—sometimes expressed in terms of the different functions of the two hemispheres of the human brain[50]—two cognitive modes are distinguished: a phylogenetically older system that is largely intuitive, and a later, more distinctively human system that is more rational and articulate. The scientific basis of this kind of two-mode understanding represents something of great importance that is too often ignored in the field of religious studies: a revived recognition of a universal aspect of human religiosity. This recognition is based, not on anthropological speculations of the kind that became common in the late-nineteenth and early-twentieth centuries (which scholars in the field of religious studies rightly regard as very questionable), but on current explorations of human brain functioning and of its evolutionary development.

Watts suggests that these two modes of mental functioning may legitimately be related to Harvey Whitehouse's distinction between early "imagistic" and later "doctrinal" developments in humanity's religious apprehension,[51] and to Robin Dunbar's comparable distinction between "shamanistic" and "doctrinal" developments.[52] These analyses, Watts shows, point to the relevance of evolutionary perspectives in our understanding of the history of human religiosity and of the faith traditions to which that religiosity has given rise. However, he rightly insists that this development over time should not be seen as involv-

imagination in religious visions in the work of Karl Rahner and Hans Urs von Balthasar, which I link to the notion of the *imagination* developed by the poet and philosopher Samuel Taylor Coleridge and to the notion of the *imaginal* developed by Henry Corbin.

49 Fraser Watts, "The Evolution of Religious Cognition," *Archive for the Psychology of Religion* 42:1 (2020): 89–100, https://doi.org/10.1177/0084672420909479.

50 This kind of perspective has been popularised in Iain McGilchrist, *The Master and His Emissary: The Divided Brain and the Making of the Modern Word*, expanded edition (Yale: Yale University Press, 2019).

51 Harvey Whitehouse, *Modes of Religiosity: A Cognitive Theory of Religious Transmission* (Lanham: AltaMira, 2004).

52 Robin Dunbar, *Human Evolution* (London: Pelican, 2014).

ing the replacement of one mode of religious "knowing" by another. Rather, he says, we must acknowledge that "new capacities exist side by side with older ones."[53] (This perception is reinforced by the way in which adherents of doctrinally focused religious communities still sometimes experience what William James called "mystical states" that "seem to those who experience them to be … states of insight into depths of truth unplumbed by the discursive intellect.")[54]

All these considerations point, in my judgment, to the plausibility of religious pluralism and even to its status as the best understanding available to people of faith—if they see the choice before them simply in terms of the question of whether to support exclusivism, inclusivism, or pluralism as these positions are usually presented. However, for those believers who incline towards a pluralistic understanding, an important question arises at this point, of whether there is an aspect of perennialism that should cause them to hesitate before adopting a full-blown pluralism.

Reciprocal Inclusivism as the Best Way Forward?

The reason for this hesitation is that one of the things that is characteristic of perennialist thinking is its stress on the need, not just for following a religious tradition of some kind, but for following one of the *particular* traditions that can provide adequate "methods"—meditative, sacramental, or liturgical—through which the esoteric truth at the heart of a valid tradition can eventually be apprehended. In this understanding, it is usually only the experience of these methods that can provide access to a tradition's esoteric heart, and yet not all faith traditions or sub-traditions are seen as possessing such methods. If the perennialists are right in this contention, then it would seem that while believers can know—from their own inner experience—that their tradition has the necessary "methods" for developing valid noetic

53 Watts, "The Evolution of Religious Cognition," 93.

54 William James, "Mysticism" (from Lectures XVI and XVII of his *The Varieties of Religious Experience*) as reprinted in Douglas W. Schrader and Asok K. Malhortra, *Pathways to Philosophy* (Upper Saddle River, NJ: Prentice Hall, 1996), 416.

insights, a comparable knowledge of the efficacy of other traditions in this respect will be impossible.

Disagreements among classic perennialists about which faith traditions provide an authentic spiritual path[55] may be seen as arising from precisely this problem. Not only can one never know the effects of the methods of other traditions "from the inside," but—and more importantly—the apprehensions that have arisen through the characteristic methods of those other traditions are not ultimately expressible in propositional terms that are susceptible to interrogation because they relate to an intuitive, noetic apprehension rather than to discursive description. This is the case both from the perennialist perspective and from the more general perspective provided by a focus on the *nous*. For both perspectives, full use of our spiritual intelligence gives rise, not to a set of propositions that are susceptible to discursive investigation, but to what Guénon calls "an immediate, or in other words, intuitive knowledge, as opposed to the discursive and mediate knowledge that belongs to the rational order."[56]

In the light of this inability of an adherent of any one faith tradition to judge other traditions' contemplative capacities, it would seem that the pluralism that is indicated by the understanding I have outlined may not be one that can be embraced wholeheartedly because, even if religious pluralism seems plausible and even cogent to us, we cannot—at least in this life—test its validity beyond reasonable doubt. Its verification can only be eschatological in nature.[57] In this situation, those of

55 For example, Guénon, with his early stress on Vedantic teachings, initially rejected Buddhism as a genuine traditional religion because he saw aspects of its teaching as unacceptable. Only gradually—under the influence of other perennialists—did he accept at least early Buddhism as valid. He also tended to reject Christianity and had considerable doubts about Schuon's more positive view of it.

56 Guénon, *The Essential René Guénon*, 105.

57 This notion of eschatological verification has been the subject of considerable discussion since its defence (in a different context) in John Hick, *Faith and Knowledge* (Ithaca: Cornell University Press, 1957). An interesting question related to this notion is that of whether—in the context of evaluating assertions that pluralism is "untraditional"—this kind of focus on the eschaton may profitably be expanded in terms of the notion of doctrinal development as something comprehensible only in terms of "tradition" being oriented towards

us who are existentially committed to a particular faith tradition may well see the need to adopt a version of what is sometimes called the precautionary principle, in which—usually in a technological context—it is seen as necessary to avoid taking a path of potential harm in any situation in which the risk of that harm cannot be fully evaluated but may not be negligible. Many exclusivists and inclusivists seem to believe that religious pluralism constitutes a spiritual danger and, if we cannot be certain that they are wrong in this, then at least some kind of inclusivism arguably remains the best option for us. We cannot become dogmatically pluralist.

However, what seems necessary, if we follow this line of thought, is that we must not only avoid dogmatic pluralism but must also avoid dogmatic inclusivism. We must recognise at least the possibility, and perhaps even the probability, of the validity of a pluralistic understanding. The inclusivism that is called for in this context cannot therefore be of the usual kind, in which the superiority of one's own faith tradition is assumed. It will be an essentially methodological inclusivism that acknowledges that we may eventually—eschatologically—come to recognise that just as other faith traditions have been viewed by us in an inclusivist way, so also our own faith tradition, in a reciprocal way, may have validly been treated by others in an inclusivist way. This extended inclusivism will not, therefore, be inclusivism of the usual kind, but will be what we might call *reciprocal inclusivism*.[58]

This kind of inclusivism has important implications for methodology in exploring the relationship between the world's faith traditions. Our starting point must involve a focus on our own particular tradi-

the *eschaton*, as expressed in David Bentley Hart, *Tradition and Apocalypse* (Grand Rapids MI: Baker Academic, 2022). In Hart's view, "openness to an unanticipated future is no less necessary than fidelity to the past" (128), and "no tradition is truly alive except one that anticipates and even wills its own overthrow in a fuller revelation of its own inner truth" (154).

58 I coined this term originally for my paper "Reciprocal Inclusivism: A Methodology for Understanding the Faiths of the World," *Journal of Ecumenical Studies* 55:4 (2020): 609–629, http://doi.org/10.1353/ecu.2020.0048. At the time of that paper's publication, I was unaware that the term had also been used in earlier publications by others, sometimes with much the same meaning as I gave to it but sometimes in terms of a more classically inclusivist stance.

tion. This will not preclude use of the perspectives of other traditions, but it will make adherents of any particular tradition wary of any kind of syncretism, so that in their exploration they see the need to focus primarily on exploring the potential for aspects of other traditions to deepen their appreciation of their own.[59]

Conclusion

The arguments I have outlined cannot, of course, be presented fully or even adequately in a short paper such as this. They may be found in much fuller form in my recent book, *Exploring Religious Pluralism*, and those who wish to assess those arguments critically must examine that book in its totality. Nevertheless, my hope is that those who are interested in the concept of spiritual intelligence will, from what I have presented here, see that such examination may not only throw important light on the question on which I have focused in this paper: that of how we should understand the relationship between the faith traditions of the world. In addition, it suggests new possibilities in relation to our more general exploration of the concept of spiritual intelligence, taking that exploration beyond the scientific (or quasi-scientific) understandings that have sometimes caused the concept to be criticised.[60]

In the framework that I have outlined, the concept of spiritual intelligence becomes not only something to be explored through the

59 This seems consonant with the perspective set out in Hart, *Tradition and Apocalypse*, which suggests not only that the Vedantic thought of Shankara might throw important light on the thinking of Maximus the Confessor, but also that "the whole rationality of the Christian tradition … entails and requires a kind of metaphysical monism that has only sporadically manifested itself in the tradition, but that certain schools of Vedanta (not to mention certain schools of Sufism) have explored with unparalleled brilliance" (183).

60 For example, Howard Gardner—in his article "A Case Against Spiritual Intelligence," *International Journal for the Psychology of Religion* 10:1 (2000): 27–34, https://doi.org/10.1207/S15327582IJPR1001_3—has avoided speaking of spiritual intelligence because of the difficulty of codifying quantifiable scientific criteria in relation to it, preferring to speak of "existential intelligence" which may (or may not) include a concern for explicitly religious matters.

methods of the sciences, but also an explicitly theological concept. As has been indicated by my critique of the widespread assumption that religious doctrines constitute straightforward "truth claims," I believe that the theologian must recognise the way in which the role of doctrines is primarily to be understood in terms of the part that they play in believers' meditative, sacramental, or liturgical experience, acting as signposts along the pathway that leads towards the goal of full noetic apprehension of the divine Reality.[61] It is in the context of this understanding, I believe, that we should see spiritual intelligence, not only as a concept that should be explored theologically, but also as a concept that is crucial to our understanding of the theological enterprise as a whole: not as something "abstract, working through concepts" but as what Lossky calls "contemplative: raising the mind to those realities which pass all understanding."[62] Here, as we have seen, the role of the *nous*, and of the intuitive, noetic perception to which it gives rise, will be central to our understanding of what this "raising the mind" must involve.

The author reports there are no competing interests to declare.
Received: 03/11/23 Accepted: 10/04/24 Published: 28/08/24

61 John Hick (see note 19) uses this term *Reality*, arguing that all the great faith traditions—including non-theistic ones—may be seen as authentic responses to this *Reality*. He uses this term, rather than the term *God*, so as not to exclude the perceptions of ultimate reality that are to be found within non-theistic traditions such as Buddhism.

62 Lossky, *The Mystical Theology of the Eastern Church*, 43.

Developing the Mystical Mind

Hans Van Eyghen

Abstract: I argue that harder, more advanced forms of religiosity, which include profound mystical experiences and a developed religious mindset, can be fostered by engaging in devotional practices. By engaging in contemplation, fasting, and sleep deprivation, subjects can move from intuitive, basic forms of religiosity to more advanced forms similar to those exemplified by Christian mystics. In support of the argument, the paper gives examples of the role of contemplation, fasting, and sleep deprivation in devotional practices of Christian mystics. It also looks closer at the effects of such practices on the neural cognitive level.

Keywords: cognitive neuroscience; contemplation; mysticism; natural religion; neuroplasticity; spiritual practices

Hans Van Eyghen is an assistant professor of philosophy at Tilburg University. His research focuses on religious epistemology, cognitive science of religion, and altered states of consciousness. Recent work includes The Epistemology of Spirit Beliefs (Routledge, 2023).

Every major religion has a (small) group of religious experts. Some of these are ritual experts with more knowledge of proper ritual behaviour. Others are theological experts with advanced knowledge of doctrines and the nature of divine beings. Some have a different expertise. They are more prone to profound mystical experiences and appear to have a vastly different outlook on the world from the rest of their fellow believers. Christianity has a long and venerable tradition of mystics who fit this picture remarkably well. They report visions of God or angels, and appear to live in very close relationship with God. Often, these mystics inspire large groups of followers, and a considerable number have been canonised, or proclaimed saints.

Quite often, mystics are called spiritually or differently *gifted.* The word "gifted" signals a commitment to the idea that such individuals were born with different skills or abilities, or that they differ because of something beyond their own will and actions. In this paper, I aim to challenge this idea at least to some extent. The main claim of this paper is that there are steps that individuals can take to transform their cognitive functions and foster profound changes in their religious experience. These changes amount to moving beyond what humans are endowed with by nature. The claim thereby aims to support the idea that religiosity is to a considerable extent the result of nurture and not merely nature. The steps relevant here are often part of devotional practices.

In support of the main claim, I look at evidence from cognitive neuroscience for cognitive change due to undertakings that resemble devotional practices. Given that the array of devotional practices is vast, the focus lies on contemplation (in connection to attention), fasting, and sleep deprivation. Although similar projects have been conducted before,[1] bringing together perspectives from history of religion and cognitive neuroscience is still quite rare.

1 See Jerome Kroll and Bernard Bachrach, *The Mystic Mind: The Psychology of Medieval Mystics and Ascetics* (Routledge, 2006). Their project, however, differs from mine in that their focus is more on psychology and less on cognitive neuroscience. Another example is Fraser Watts, *A Plea for Embodied Spirituality:*

The paper is exploratory in nature. Much more can be said regarding devotional practices and their cognitive effects. The discussion below gives a preliminary indication of how religious practices can foster different religious beliefs, experiences, and mindsets. A fuller investigation will need to look at the broader context of devotional practices and at their effects on the neural level in greater detail. This lies beyond the scope of this paper.

This paper is structured as follows: First, I compare two forms of religiosity, which I call "easy religiosity" and "hard religiosity." Then, I discuss three devotional practices, contemplation, fasting, and sleep deprivation, and give examples from the history of Christian mysticism. After that, I consider the effects of these three practices on human cognition from a neuroscientific perspective. The last section concludes the paper.

Easy and Hard Religiosity

Scholars of religion are aware of the ease with which terminology can get confused. For example, they seem to have many, sometimes rather different, phenomena in mind when they use the term "religion." The same goes for religiosity, religious belief, religious idea, and religious practice. This section serves to clear up some of the confusion and distinguishes two different kinds of religiosity. One I call "basic" or "easy religiosity." The other I call "advanced" or "hard religiosity." The latter is the religiosity of mystics and the former that of most common believers.

Both kinds of religiosity pertain not so much to a collection of beliefs or ideas stored in a given human mind, but rather to a mindset or skillset of humans in general. How to define "religiosity" accurately is a matter of debate, with varying definitions used depending on one's goals.[2] For reasons of scope and simplicity, herein I consider religiosity

The Role of the Body in Religion (Eugene, OR: Wipf and Stock Publishers, 2022). Watts focuses more on psychology as well.

2 For a discussion, see Barbara B. Holdcroft, "What Is Religiosity," *Catholic*

to be an ability to have a specific religious mindset and to elicit religious experiences.

Both points merit some elaboration. Having a religious mindset refers to how religious ideas or behaviour permeate one's life. Having a different religious mindset may involve having different beliefs, but that is not necessary. For some people, the religious mindset pertains to moral inclinations and daily practices such as prayer. For others, it might encompass the idea that God has a plan for them or the sense of being guided by God in most of their activities. Religious experiences include profound mystical experiences, where subjects are personally in direct contact with God (or angels and saints), and vaguer experiences such as the sense that God is present.

Presenting a contrast between easy and hard forms of religiosity (see below) is to some extent reductive. A more elaborate account would need to be a continuum of various intermediate forms of easier and harder forms of religiosity. Again, for reasons of scope and simplicity, I adopt a contrast where "easy religiosity" captures the most basic forms of the continuum, while "hard religiosity" includes more complex forms.

Easy Religiosity

Overall, religiosity appears to come easily. Common sense shows that many people easily obtain religious ideas, religious behaviour, and perhaps have vague religious experiences. Many people seem to acquire religious beliefs easily through inculturation. Testament to this is the fact that many accept the religious ideas that are dominant in their environment. They appear to develop these forms of religiosity rather spontaneously or organically, without much effort.

The idea that some forms of religiosity come easily has some scientific basis. Defenders of cognitive and evolutionary accounts of

---

Education: A Journal of Inquiry and Practice 10:1 (2006), 89–103, https://ejournals.bc.edu/index.php/cej/article/view/733.

religion tend to underwrite the claim that religion is natural.[3] Claiming that religion is natural can mean various things.[4] Defenders of cognitive accounts tend to use the term "natural" as meaning intuitive or easily acquired. For some theories, a belief is natural in this sense that it is the outcome of one or more cognitive mechanisms that are left unimpeded or unaltered by cultural influences.[5] Other accounts allow for cultural alterations to cognitive mechanisms but merely stress the ease or intuitiveness by means of which beliefs are produced and adopted.[6]

Some add that the ease by which people gain basic religiosity is the result of natural selection. Because having religious beliefs is evolutionarily advantageous, religious beliefs are easily acquired. The same goes for other basic forms of religiosity, which do not require much effort. Important reasons for selecting religion as evolutionarily advantageous might have been its role in social monitoring,[7] fostering cooperation,[8] and increasing the odds of procreation.[9] Important for our discussion is that evolutionary accounts of religious phenomena

3 A fuller summary of key ideas in cognitive and evolutionary science of religion lies beyond the scope of this paper. For overviews, see: Claire White, *An Introduction to the Cognitive Science of Religion: Connecting Evolution, Brain, Cognition and Culture* (London and New York: Routledge, 2021); Luther H. Martin and Donald Wiebe (eds), *Religion Explained? The Cognitive Science of Religion after Twenty-Five Years* (London: Bloomsbury, 2017).

4 For an overview of various meanings, see: Aku Olavi Visala and Justin Barrett, "In What Senses Might Religion Be Natural," in *The Naturalness of Belief: New Essays on Theism's Reasonability*, ed. Paul Copan and Charles Taliaferro (Lanham: Lexington Books, 2018), 67–84.

5 This is roughly the account of McCauley. He distinguishes maturationally natural from practiced natural beliefs where the former are beliefs that are produced by unimpeded cognitive mechanisms. See Robert N. McCauley, *Why Religion Is Natural and Science Is Not* (Oxford: Oxford University Press, 2011).

6 See for example: Marc Andersen, "Predictive Coding in Agency Detection," *Religion, Brain & Behavior* 9:1 (2019): 65–84.

7 Ara Norenzayan, *Big Gods: How Religion Transformed Cooperation and Conflict* (Princeton, NJ: Princeton University Press, 2013).

8 Robin Dunbar, *How Religion Evolved: And Why It Endures* (Oxford: Oxford University Press, 2022).

9 James A. Van Slyke and Konrad Szocik, "Sexual Selection and Religion: Can the Evolution of Religion Be Explained in Terms of Mating Strategies?" *Archive for the Psychology of Religion* 42:1 (2020): 123–41.

are usually tied to the ease of basic religiosity. Reaping the evolutionary benefits of basic religious beliefs or behaviour does not require too much effort or energy from the subject. Certain researchers connect evolutionary accounts to cognitive theories, arguing that natural selection selected cognitive mechanisms that easily yield religious beliefs. Others argue that natural selection fostered the cultural spread of religion.[10]

Hard Religiosity

While few will object that some forms of religiosity come easily to many people, basic forms of religiosity are not all there is to religion. Some researchers object to the crude representation of religion in cognitive accounts of religion. They argue that cognitive theories cannot account for all religious phenomena, especially those that are very costly in terms of energy and time consumption.[11] Some cognitive accounts also point to how spirituality can be trained by various practices.[12] In any case, there is a large number of religious experiences that require considerable effort on behalf of the believer.

The evidence for more advanced or harder forms of religiosity again comes in two forms: common sense and more systematic insights. It is clear from common sense that some subjects have different religious experiences, beyond the vague sense of God's presence. Others have a different religious mindset, involving living in much

10 See Norenzayan, *Big Gods*.

11 See, for example, Konrad Szocik and Hans Van Eyghen, *Revising Cognitive and Evolutionary Science of Religion: Religion as an Adaptation*, vol. 8 (Cham: Springer, 2021). Some defenders of cognitive theories argue that their ideas account for extreme rituals that require much effort. They argue that highly costly rituals can serve as a means to signal allegiance to prosocial norms and thereby increase evolutionary fitness. See Dimitris Xygalatas, "Extreme Rituals," *The Oxford Handbook of the Cognitive Science of Religion*, ed. Justin L. Barrett (Oxford: Oxford University Press, 2022), 237. However, such theories tend to focus on rare extreme rituals (often, these are rituals that occur once a year or perhaps once in a lifetime) and not on enduring, effortful forms of religiosity.

12 See, for example, Tanya M. Luhrmann, *When God Talks Back: Understanding the American Evangelical Relationship with God* (New York: Knopf, 2012).

closer perceived proximity with God. Ample evidence for hard religiosity and religious experts is of an historical nature. To this I must now turn.

Most religious traditions have religious experts. They report exceptional life stories and tend to have more profound religious experiences. Christianity has mystics. While many Western readers immediately think of medieval or early modern mystics like Hadewijch (13th century) or Theresa of Avila (1515–1582), Christianity has always known such exceptional figures. Early examples are desert fathers or desert ascetics like Paul of Thebes (227–341) and Anthony of Egypt (251–356). They chose to live a very simple and austere life devoted to prayer and contemplation.[13] Saint Anthony of Egypt reported visions of God, angels, and demons. He and other desert fathers exerted great admiration and authority among Christian believers.[14]

Well known medieval mystics are Beda Venerabilis (673–735) and Meister Eckhart (1260–1328). Beda Venerabilis was a monk for most of his life and spent his time in prayer and contemplation. He was also a prolific writer and exerted great authority as a teacher. His work includes commentaries on Bible books and a new monastic rule.[15] Furthermore, he showed a great interest in otherworldly visions. These led him to theological insights on the history of salvation and the unknowability of the time of the eschaton.[16] In turn, Meister Eckhart's works illustrate changes in the religious mindset and experiences. Meister Eckhart is best known as a mystic but he was also a highly respected theologian in his lifetime. He developed the idea

13 Marilyn Dunn, *The Emergence of Monasticism: From the Desert Fathers to the Early Middle Ages* (Hoboken, NJ: John Wiley & Sons, 2008).

14 Louis Groarke, "Anthony of Egypt and the Desert Fathers," in *Meet the Philosophers of Ancient Greece*, ed. Patricia F. O'Grady (London and New York: Routledge, 2021), 227–29.

15 Scott DeGregorio, "Bede (Beda Venerabilis), c. 673–735 CE," in *Oxford Research Encyclopedia of Classics*, 2016, https://www.sciencegate.app/source/565749172.

16 Sharon Rowley, "The Role and Function of Otherworldly Visions in Bede's *Historia Ecclesiastica Gentis Anglorum*," in *The World of Travellers: Exploration and Imagination: Germania Latina VII*, ed. K. Dekker et al., Mediaevalia Groningana (Leuven: Peeters Publishers, 2009), 165–83.

that humans are intimately conversant with God in the depths of their soul. His preaching was aimed primarily at achieving inner union with God.[17] Eckhart's theological ideas were likely influenced by his mystical experiences. He describes the first stage towards mystical experience as follows:

> First, the soul experiences within itself the growth of fear, hope and desire—i.e., of natural human emotions. Secondly, these emotions are altogether extinguished from the soul. Thirdly, the soul becomes oblivious to all temporal things. And, fourthly, it enters into God as he exists and rules eternally. In this fourth state it never thinks about itself or temporal things, being immersed in God as God is immersed in it; whatever it does, it does in God.[18]

Religious experts or people with different, more advanced forms of religiosity are also to be found in different traditions. Judaism has a venerable tradition of mysticism. Practitioners of Qabalah or other forms of Jewish mysticism report profound mystical experiences with God and angels.[19] Some also develop different beliefs like the *Tikkun Olam* idea of repairing creation.[20] The best-known Islamic mystics are Sufis. Through various practices, they aim at nondual experiences of Allah. Many Indian mystics aim at a similar state of *samadhi*.[21] Although there are differences between mystical experiences cross-culturally and certainly between beliefs obtained by mystics, they all are very different from the basic forms of religiosity we discussed above.

17 Alessandro Palazzo, "Meister Eckhart (Updated Version)," in *Encyclopedia of Medieval Philosophy: Living Edition* (Cham: Springer, 2018), 1–7.

18 Richard Kieckhefer, "Meister Eckhart's Conception of Union with God," *The Harvard Theological Review* 71:3-4 (1978): 203–225, esp. 220.

19 See Moshe Idel, "Astral Dreams in R. Yohanan Alemanno's Writings," *Accademia* 1 (1999): 111–128.

20 Gerald J. Blidstein, "Tikkun Olam," *Tradition: A Journal of Orthodox Jewish Thought* 29:2 (1995): 5–43.

21 See Richard Shankman, *The Experience of Samadhi: An In-Depth Exploration of Buddhist Meditation* (Boulder, CO: Shambhala Publications, 2008).

On a more general level, one might argue that there is such a thing as *spiritual intelligence.* Like other forms of intelligence, spirituality involves relationality, the development of moral values, and participation in transcendent realities. There is some discussion on whether spiritual intelligence is a distinct kind of intelligence, different from other forms.[22] In any case, the term "spiritual intelligence" implies a certain degree of differentiation. As some people have more general intelligence, others likely have more spiritual intelligence than others.

Devotional Practices and Changing Religiosity

So far, we distinguished two forms of religiosity. One is rather basic, consisting of simple beliefs, mindset, and rather vague experiences. The other encompasses more elaborate beliefs, a more intimate relationship to God, and more profound experiences.

There is a tendency in the literature on religious expertise to regard it as closely tied to a person's (innate) personality. Some persons would simply be more spiritually gifted or more prone for such forms of religiosity. On strong readings of this idea, hard religiosity would be exclusively preserved for such spiritually gifted people. Non-gifted people then do not have access to hard forms of religiosity. If that is the case, such forms of religiosity may not be hard at all. Advanced forms of religiosity might come easily to gifted individuals. They may be as prone to more advanced forms of religiosity as most people are to easier forms of religiosity. Evidence for this view may come from accounts of people who enjoy profound experiences at a very early age. Given that they did not have the time to take effortful steps to take their religiosity to the next level, their abilities are likely innate. Additional evidence may be gained from evidence of close links between dispositions for hard religiosity and genetic predispositions. While this may be the case for some individuals, there is sufficient reason to believe that subjects can move from easier forms of religiosity to harder forms. In

22 Harris Wiseman and Fraser Watts, "Spiritual Intelligence: Participating with Heart, Mind, and Body," *Zygon* 57:3 (2022): 710–718.

this section and the next, I look closer at that movement. After discussing some practices that are often pointed to in order to develop greater spiritual abilities (e.g., devotion, prayer, fasting), the next section discusses what may be going on at the neural or cognitive level.

The discussion below does not serve to show that all examples of hard religiosity are the result of engaging in spiritual practices. Even if (some) spiritual practices can help people achieve harder forms, it remains a genuine possibility that some people are born with special spiritual gifts. Hard forms of religiosity may come naturally or fairly easily to them. The discussion serves to argue against stronger claims that hard religiosity is exclusively preserved for spiritually gifted individuals.[23]

The mystics or religious experts I discussed above often had similar ways of living. A large number spent most of their lifetimes away from society, often in convents or in isolation. Hermits spent their lives in seclusion in Egypt's desert. Beda Venerabilis spent a number of years in a monastery in Monkwearmouth. Often, these people engaged in similar spiritual practices. Saint Anthony of Egypt led a hermit's life of prayer and devotion to God. Beda spent much of his life in monasteries where he likely partook in devotional practices.

Devotional practices regularly included forms of contemplative prayer. The desert fathers practiced an early form of what the Byzantines called "hesychasm," the path of serenity. They recited short prayers repetitively and recited psalms. In doing so, they tried to block out other thoughts.[24] During the Middle Ages, systematic treatises were written to teach devotional practices.[25] Medieval forms of contemplation also involved repetitive prayer and focus on God or sacred images. More than before, the practice became emotionally charged.[26]

23 I thank an anonymous reviewer for pointing this out.

24 John Wortley, "Prayer and the Desert Fathers," *The Coming of the Comforter: When, Where, and to Whom*, ed. Carlos A. Segovia and Basil Lourié (Piscataway, NJ: Gorgias Press, 2012), 109–129.

25 One of the best known is the anonymous treatise *The Cloud of Unknowing*.

26 Karl Baier, "Meditation and Contemplation in High to Late Medieval Europe," in *Yogic Perception, Meditation and Altered States of Consciousness*, ed. Eli Franco

The Middle Ages saw the rise of *lectio divina* in monastic circles, a practice in which Scripture is read in a meditative, devotional way, to foster communion with God.

Another recurring element in the practice of mystics is fasting. Fasting was and is a common element of the lives of Christian monks and nuns.[27] Fasting is seen as a way of imitating Christ[28] and of purifying one's soul and body. Subjects sometimes practice fasting for shorter periods (e.g., in preparation for the Eucharist) or longer periods (e.g., in preparation for Easter). Fasting may merely consists in abstaining from certain kinds of food (e.g., not eating meat on Fridays in Roman Catholicism). It may also consist in altogether abstaining from food. Fasting is observed more strictly in some religious orders like the Carthusians.[29] Some of the desert fathers were known for their "heroic fasting."[30] A direct motivation for fasting among medieval monks and nuns was a belief that fasting would set right the wrongness inflicted by sin. Fasting was regarded as a way to purify the body and make it holy. In extreme cases, some consumed only the Eucharist and abstained from all other forms of food. An example of such extreme fasting is found in the life of Catherine of Genoa (1447–1510).[31]

Fasting is also a common element in non-Christian traditions. Some Hindus fast on regular days.[32] Many Muslims fast during the

(Leipzig: Austrian Academy of Sciences Press, 2009), 325–349.

27 For more examples of fasting practices in Christian mysticism, see Kroll and Bachrach, *The Mystic Mind*.

28 See Christ's fasting for forty days and forty nights in the desert before being tempted by the devil (Matthew 4:1–11).

29 See Mathilde van Dijk, "Baking the Bread and Roasting the Meat: Dorlandus's Saint Lawrence as a Model for Carthusians," *The Medieval Low Countries* 4 (2018): 189–214.

30 Gerald L. Sittser, "The Battle Without and Within: The Psychology of Sin and Salvation in the Desert Fathers and Mothers," *Journal of Spiritual Formation and Soul Care* 2:1 (2009): 44–66.

31 Fernando Espi Forcen and Carlos Espi Forcen, "The Practice of Holy Fasting in the Late Middle Ages: A Psychiatric Approach," *The Journal of Nervous and Mental Disease* 203:8 (2015): 650–653.

32 Debjani Chatterjee, "Apara Ekadashi 2021: Date, Ekadashi Fasting Time, Rituals and Significance," *NDTV*, 6 April 2021, https://tinyurl.com/3kf5auu2.

month of Ramadan. Some also refrain from food in preparation for feast days.[33] Many observant Jews fast on Yom Kippur. Buddhists sometimes observe fasting during periods of intense meditation.[34]

Another technique that is regularly employed to foster mystical experiences and a different mindset is sleep deprivation.[35] The desert fathers saw sleep as an enemy of devotion. Though they did not go altogether without sleep, they limited it as much as possible since it got in the way of prayer and other devotional practices.[36] For medieval mystics, sleep had an ambiguous status. On the one hand, sleep could be a time of peaceful closeness to God. On the other hand, sleep also provided an occasion for demonic temptation. For example, fear of demonic temptation led Marie of Oignies (1177–1213) to refrain from sleep. She received visions and stigmata wounds. Sleep deprivation was a common element in the ascetic practice of other medieval mystics as well,[37] and is also part of practices in different traditions, like Zen Buddhism and forms of Hinduism.[38]

Other common ascetic practices include reclusion, chastity, fortitude in illness, and flagellation.[39] Still others are engagement in ritual activities, reading and study of texts, and social transmission such as initiation or discipleship.[40] All are found within and outside of

33 Emmanuel Sivan, "Sunni Radicalism in the Middle East and the Iranian Revolution," *International Journal of Middle East Studies* 21:1 (1989): 1–30.

34 Yujin Lee and Michael Krawinkel, "Body Composition and Nutrient Intake of Buddhist Vegetarians," *Asia Pacific Journal of Clinical Nutrition* 18:2 (2009): 265–271.

35 For more examples of Christian mystics depriving themselves of sleep, see Kroll and Bachrach, *The Mystic Mind*.

36 John Wortley, *An Introduction to the Desert Fathers* (Cambridge: Cambridge University Press, 2019).

37 Macmillan Sarah, "'The Nyghtes Watchys': Sleep Deprivation in Medieval Devotional Culture," *Journal of Medieval Religious Cultures* 39:1 (2013): 23–42.

38 Núria M. Farré-i-Barril, "Sleep Deprivation: Asceticism, Religious Experience and Neurological Quandaries," *Religion and the Body: Modern Science and the Construction of Religious Meaning*, ed. David Cave and Rebecca Sachs Norris, Numen Book Series 138 (Leiden and Boston: Brill, 2012), 217–234.

39 Sarah, "The Nyghtes Watchys."

40 I thank an anonymous reviewer for pointing me to these practices.

Christianity as well. Although many practices share common elements, like a changed attitude towards the physical body, they are profoundly different. As Jerome Kroll and Bernard Bachrach note, spiritual practices do not occur in a cultural vacuum.[41] How they are practiced and experienced, and what effects they have is often heavily influenced by their culturally specific embedment. The effects of cultural "embeddedness" lie beyond the scope of this paper. Regardless, similar practices will likely have similar effects across cultures. Future study may reveal how far the similarities go.

In the next section, I look closer at the neural and cognitive effects of focused contemplation, fasting and sleep deprivation.

Neural and Cognitive Effects of Ascetic Practices

This section looks closer at the cognitive effects of contemplation, fasting, and sleep deprivation. All three are common elements to (Christian) religious devotion. This discussion will serve to argue that devotional practices can foster more advanced forms of religiosity, even unbeknownst to practitioners.

Focused Contemplation

Of the three elements of devotion under consideration here, contemplation is harder to match to bodily or mental behaviour that can be studied empirically or neurologically. This can be overcome by noting how contemplation is closely tied to attention. Etymologically, the word "contemplation" is derived from the Latin *templum*, which traditionally signified the place marked out by a seer as a location for his observation. Later it came to designate the actual observations made by the seer. Contemplation usually involves focused observation and admiration of an object or subject.[42] A clear example is the Roman Catholic practice

41 Kroll and Bachrach, *The Mystic Mind*; Watts, *A Plea for Embodied Spirituality*, 48–49.

42 J. Aumann, "Contemplation," in *New Catholic Encyclopedia*, second edition

of Eucharistic adoration. During Eucharistic adoration, a consecrated host is displayed in sight of worshippers. Worshippers use the host as object of focus during prayer. Contemplation frequently involves focus on prayers, like the "Jesus prayer," or on images of saints. It is clear from these examples that contemplation involves attentiveness. During contemplation, worshippers try to focus on an object and block out other sensory stimuli or thoughts. The focus during contemplation may lead to cognitive changes. The discussion below looks at cognitive effects on cognition from general attentiveness.

Much of the recent work on attention in neuroscience is set in the background of predictive processing or related accounts of human cognition. In predictive processing, human experience and cognition are heavily influenced from top-down constraints. Humans would have an internal model of the world, which is gradually built up by experience. This internal model contains information about the likelihood that humans will encounter something, and it generates predictions about what humans will experience. If the predictions match the sensory input, the model runs its course and informs experience. If there is a mismatch, the internal model is updated to prevent future mismatches.[43]

Attention has a straightforward role in predictive processing. By attending carefully to an object, more sensory stimuli of that object are allowed in. In this way, attention alters the inputs to cognitive processing.[44] This increases the odds of sensory stimuli that do not

(Detroit: Thomson Gale, 1967), https://www.encyclopedia.com/philosophy-and-religion/philosophy/philosophy-terms-and-concepts/contemplation.

43 The whole process is governed by the free-energy principle. See Karl Friston and Stefan Kiebel, "Predictive Coding Under the Free-Energy Principle," *Philosophical Transactions of the Royal Society B: Biological Sciences* 364:1521 (2009): 1211–1221. The general idea is that human minds (like any system) aim to reduce entropy by ensuring a good fit between the internal model and sensory input.

44 Firestone and Scholl discuss this possibility but dismiss it. They claim that changes in sensory input through attention are compatible with accounts where cognition is not (or far less) constrained by top-down influences. See Chaz Firestone and Brian J. Scholl, "Cognition Does Not Affect Perception: Evaluating the Evidence for 'Top-Down' Effects," *Behavioral and Brain Sciences*

match the internal model and consequently prompts revisions. Some argue that attention has a different role as well. For example, Harriet Feldman and Karl Friston argue that attention aids in inferring the level of uncertainty or precision. By attending carefully, the mind can gain a better measure of the uncertainty of the cause of sensory input.[45] Andy Clark argues that attention also has a role in precision weighting. Attention can optimise the relative influence of top-down predictions against sensory input. Altering the precision weighting on specific error signals alters the influence of one neural area on another or on how signals are processed.[46] Clark adds that much of the role of attention in precision weighing likely remains unconscious or subpersonal.[47]

How could this apply to contemplation? By focused attention on an object, subjects could become more mindful of mismatches between their internal model of the world (including, but not limited to, their beliefs) and sensory input. This could open the door to revision of the internal model with an openness for new ideas regarding God or the divine. An altered internal model could in turn allow new experiences. It could also lead to different forms of behaviour and a different outlook on the world. In that way, contemplation could lead to new forms of advanced religiosity.

Fasting

The cognitive effects of fasting are much better understood and require less speculation. Most studies support the idea that fasting has mostly

39 (2016): e229, DOI: 10.1017/S0140525X15000965. Clark does see a role for attention in altering sensory inputs but also assigns it a greater role. Andy Clark, "Attention Alters Predictive Processing," *Behavioral and Brain Sciences* 39 (2016): e234, DOI: 10.1017/S0140525X15002472.

45 Harriet Feldman and Karl J. Friston, "Attention, Uncertainty, and Free-Energy," *Frontiers in Human Neuroscience* 4 (2010): 215, https://doi.org/10.3389/fnhum.2010.00215.

46 See Hanneke E. M. den Ouden et al., "Striatal Prediction Error Modulates Cortical Coupling," *Journal of Neuroscience* 30:9 (2010): 3210–3219, DOI: 10.1523/JNEUROSCI.4458-09.2010.

47 Clark, "Attention Alters Predictive Processing."

beneficial effects on cognition, even benefitting diseases like ischemic stroke, autism spectrum disorder, and mood and anxiety disorders. From an evolutionary point of view, beneficial effects of fasting make sense as well. Periods of food scarcity were the rule rather than the exception for most of human history.[48]

The effects of fasting on cognition are mediated by a number of neural processes. Changes in metabolism are one of them. Around twelve to thirty-six hours after fasting begins, the body switches from preferring to extract energy through glycogenolysis (breakdown of glycogen into glucose) to lipolysis (breakdown of stored fat as lipids from adipose tissue). Released lipids are further metabolised to free fatty acids, transformed into acetyl CoA and transformed to the ketone bodies β-hydroxybutyrate (BHB) and acetoacetate (AcAc). Ketones become the preferred fuel for the brain. One of the roles of ketones is regulating transcription factors in neurons. Ketones lead to an up-regulation of neurotrophic factors, which are associated with the promotion of mitochondrial biogenesis, synaptic plasticity, and cellular stress resistance in animal models. This triggers repair of neural matter by stimulating autophagy, a process in which neurons remove dysfunctional or damaged components.[49]

By changing the brain's diet to ketones, fasting leads to removal of neural waste and, importantly, to synaptic plasticity (i.e., the capacity of neurons to change the strengths of their connections). This can lead to new neural pathways allowing for different forms of cognition. In the right setting, this can foster new, advanced forms of religiosity.[50]

Fasting also has effects on the circadian clock mechanism (which regulates the cycle of alertness and sleepiness by responding to changes in lighting).[51] Consuming food at different times than the

48 Jip Gudden et al., "The Effects of Intermittent Fasting on Brain and Cognitive Function," *Nutrients* 13:9 (2021): 3166, DOI: 10.3390/nu13093166.

49 S. Reddy et al., "Physiology, Circadian Rhythm," *StatPearls* (2018), https://www.ncbi.nlm.nih.gov/books/NBK519507/2018.

50 Reddy et al., "Physiology."

51 Reddy et al., "Physiology."

normal eating rhythm may set the internal clock out of phase. One way in which fasting can affect the circadian rhythm is through hormonal synchrony. The peak of insulin secretion is usually reached in the early morning and is further heightened during and after food intake. Fasting can decrease post-meal and average insulin levels, leading to an overall increased sensitivity for insulin.[52] Acute high levels of insulin may have beneficial effects on human cognition while persistently high levels have negative effects on memory and other cognitive functions.[53]

Finally, fasting affects the gut microbiome. Gut microbiome influences the brain through neural, endocrine, and immune pathways, which are collectively called the microbiota-gut-brain axis. Roughly 15% of microbiota dynamically oscillates in activity and abundance throughout the day in accordance with circadian and hormonal fluctuations. These are affected by dietary intake. In one study, fasting enriched the gut microbiome composition and led to improved cognitive functioning.[54]

Fasting thus has multiple effects on human cognition. Most important for our purposes is that fasting can alter cognition (because of increased neural pathways, plasticity, or enriched gut microbiota). Thereby fasting can allow for new experiences and new beliefs. The latter may inform a different outlook on the world.

Sleep Deprivation

Unlike focused attention and fasting, sleep deprivation is commonly associated with diminished cognitive performance. Especially deprivation of rapid eye movement sleep (REM: a phase of sleep characterised by the rapid movement of the eyes, low muscle tone, and a tendency for vivid dreams) has effects on neural behaviour. NREM (non-rapid

52 Gudden et al., "The Effects of Intermittent Fasting."

53 Brenna Cholerton et al., "Insulin, Cognition, and Dementia," *European Journal of Pharmacology* 719:1–3 (2013): 170–179, DOI: 10.1016/j.ejphar.2013.08.008.

54 Zhigang Liu et al., "Gut Microbiota Mediates Intermittent-Fasting Alleviation of Diabetes-Induced Cognitive Impairment," *Nature Communications* 11:1 (2020): 1–14, https://doi.org/10.1038/s41467-020-14676-4.

eye movement) sleep deprivation reduces the release of specific neurotransmitters which affect the ability of neural receptors to refresh and restore sensitivity. The result is reduced cognition. Subjects with sleep deprivation also have reduced functional connectivity between the amygdala and medial prefrontal cortex. The latter is known for producing strong inhibitory projections to the amygdala. In addition, sleep deprivation leads to higher connectivity in the autonomic areas of the locus coeruleus and amygdala. Therefore, lack of sleep can lead to increased amygdala hyperlimbic reactions that result in stimuli with negative emotional connotations. The brain misses a corrective reaction from the medial frontal cortex, causing inappropriate behavioural responses, like lack of rational decisions and social judgments.[55]

Sleep deprivation also has detrimental effects on memory. It disrupts memory consolidation in the hippocampus, resulting in fewer permanent memories being consolidated in the brain. Furthermore, sleep deprivation down-regulates the mammalian target of rapamycin signalling, which is a regulatory protein required for memory consolidation.[56]

Sleep deprivation also negatively affects attention and alertness because of imbalanced inhibition between the frontoparietal network and the amygdala. A final negative effect of sleep deprivation concerns clearing of waste in the central nervous system. This can lead to toxic build-up, which can negatively affect cognitive performance.[57]

While all effects we discussed so far are negative on cognitive performance, some evidence suggests that sleep deprivation can contribute to cognitive change. During the slow wave stage of NREM sleep, synapses are decreased in the brain to counterbalance the net strengthening of network synapses, such as those that occur during learning. Without the normalisation on synaptic power, however, sleep

55 Mohammad A. Khan and Hamdan Al-Jahdali, "The Consequences of Sleep Deprivation on Cognitive Performance," *Neurosciences Journal* 28:2 (2023): 91–99, DOI: 10.17712/nsj.2023.2.20220108.

56 Khan and Al-Jahdali, "The Consequences of Sleep Deprivation."

57 Khan and Al-Jahdali, "The Consequences of Sleep Deprivation."

deprivation increases the weight of plasticity on the nerve cells, and these thereby fail to reestablish neural selectivity and learning performance.[58] Sleep deprivation can thus have detrimental effects on a subject's ability to learn.

Sleep deprivation alters the connection between neurons. Under normal circumstances (i.e., without sleep deprivation) relevant connections are strengthened, and irrelevant ones weakened. Because of sleep deprivation, external stimuli and information are processed poorly or not at all. This impairs learning as well.[59]

Other evidence for change caused by sleep deprivation is its effect on serotonin release. Research on rats shows that REM sleep-deprived rats have a higher incidence of serotonin syndrome (a pathology which occurs because of high serotonin build-up) and a greater number of headshakes, when challenged with serotonin precursors. Increased turnover due to REM sleep deprivation could explain the stronger response to administered serotonin precursors. REM sleep deprivation could induce the supersensitivity of dopamine receptors in the brain.[60] While low levels of serotonin are associated with poor memory and a depressed mood, higher levels have positive effects on memory and attention.[61]

The effects of sleep deprivation on memory are thus ambiguous. Most studies point to detrimental effects while some see a positive effect (mainly mediated through serotonin). Some studies report individual differences in effects on memory.[62] Thus far, the discussion does

58 Khan and Al-Jahdali, "The Consequences of Sleep Deprivation."

59 Mohammad Ali Salehinejad et al., "Sleep-Dependent Upscaled Excitability, Saturated Neuroplasticity, and Modulated Cognition in the Human Brain," *Elife* 11 (2022): e69308, DOI: 10.7554/eLife.69308.

60 Ricardo Santos and E. A. Carlini, "Serotonin Receptor Activation in Rats Previously Deprived of REM Sleep," *Pharmacology Biochemistry and Behavior* 18:4 (1983): 501–507, DOI: 10.1016/0091-3057(83)90271-x.

61 Trisha A. Jenkins et al., "Influence of Tryptophan and Serotonin on Mood and Cognition with a Possible Role of the Gut-Brain Axis," *Nutrients* 8:1 (2016): 56, DOI: 10.3390/nu8010056.

62 Jacqueline T. Weiss and Jeffrey M. Donlea, "Roles for Sleep in Neural and Behavioral Plasticity: Reviewing Variation in the Consequences of Sleep Loss," *Frontiers in Behavioral Neuroscience* 15 (2022): 777799, DOI: 10.3389/

not readily support the idea that sleep deprivation can foster different forms of religiosity.

More relevant to our discussion is the well-established connection between sleep deprivation and psychosis (a pathology where subjects seemingly lose contact with reality and are prone to auditory and visual hallucinations). Subjects deprived of sleep are more prone to psychotic episodes or hallucinations. An overview of twenty-one studies notes that all except one reported perceptual changes caused by sleep deprivation. The changes include visual distortions, illusions, somatosensory changes, and hallucinations. In 90% of the studies, sleep deprivation affected the visual modality. In 52% of the cases, the somatosensory modality was affected, and in 33% of them the auditory modality. The effects on perception developed rapidly after one night without sleep and progressed in an almost predictable way. After twenty-four to forty-eight hours without sleep, perceptual distortions, anxiety, irritability, depersonalisation, and temporal disorientation started. After forty-eight to ninety hours of deprivation, complex hallucinations and disordered thinking followed. After seventy-two hours, delusions followed, and the clinical picture began to resemble acute psychosis or toxic delirium. After three days without sleep, hallucinations in all three sense modalities were reported. The symptoms usually resolved after periods of sleep.[63] Some studies also note effects on paranoia apart from hallucinations. These were likely mediated by negative effects, like fear or a bad emotional mood.[64]

While not all forms of psychosis are well understood, the increased number of hallucinations[65] can explain changes in cognition

fnbeh.2021.777799.

63 Flavie Waters et al., "Severe Sleep Deprivation Causes Hallucinations and a Gradual Progression toward Psychosis with Increasing Time Awake," *Frontiers in Psychiatry* 9 (2018): 303, DOI: 10.3389/fpsyt.2018.00303.

64 Sarah Reeve et al., "Disrupting Sleep: The Effects of Sleep Loss on Psychotic Experiences Tested in an Experimental Study with Mediation Analysis," *Schizophrenia Bulletin* 44:3 (2018): 662–671, DOI: 10.1093/schbul/sbx103.

65 An important question regards the veracity of experiences due to sleep deprivation. Hallucinations are usually defined as erroneous or misguided experiences. If that is the case, sleep deprivation may set subjects astray. A

after sleep deprivation. Lack of sleep can open the doors for new experiences, which may lead to lasting changes in cognition. New experiences may prompt an update of the internal model of the world (see above). These may, in turn, foster new experiences and a different outlook.

This section supports the claim that all three devotional practices have effects on human cognition. The evidence for increased neuroplasticity is stronger for contemplation and fasting. Some evidence also suggests that sleep deprivation can lead to increased plasticity or other forms of change as well. Increased neuroplasticity can, in turn, explain changes in experiences and mindset. Sleep deprivation can also make subjects more prone to different experiences.

A mere increase in neuroplasticity or mere changes in experiences are not enough to have changes in line with what was called above "hard religiosity." The mind also needs specific input to have the kinds of mystical experiences and different mindset common in mystics. The required input can be infused by the setting. Mystics usually live in religious settings where religious visual imagery is common. They are also frequently exposed to religious writings or sermons which can also serve as sensory input fostering a new mindset.

Conclusions

The main argument of this paper is that humans can take active steps to achieve different, more advanced forms of religiosity. In support of this claim, evidence for the role of three religious practices (contemplation, fasting, and sleep deprivation) was surveyed. Evidence from

thorough discussion of the veracity and reliability of sleep-deprivation-induced experiences lies beyond the scope of this paper. I merely note here that not all psychotic hallucinations are caused by lesions or brain damage. See Femi Oyebode, "The Neurology of Psychosis," *Medical Principles and Practice* 17:4 (2008): 263–269, DOI: 10.1159/000129603. Experiences of this kind might merely be *altered* experiences rather than erroneous experiences. For a discussion, see Hans Van Eyghen, *The Epistemology of Spirit Beliefs*, Routledge Studies in the Philosophy of Religion (New York and Abingdon: Routledge, 2023), 79–82.

cognitive neuroscience supports the idea that these practices can alter human cognition.

Engaging in practices of this sort is not a gift or an innate property of practitioners. Usually, engaging in such practices is a willed act. If engaging in the practices can lead to more advanced forms of religiosity, such a transformation can therefore be a conscious, willed act. As noted, a considerable number of practitioners likely engaged in these practices for different reasons than their contribution to cognitive change. Nonetheless, the discussion shows that transformation of religiosity (willed or not) is a genuine possibility.

Other important questions, such as how practices foster specific forms of religiosity or moral problems when the practices that require substantial time and energy (and therefore likely support from others), were beyond the scope of this paper.[66] It also remains an open question whether engaging in devotional practices will have the same effect on everyone. Increased knowledge of the cognitive effects of these practices can benefit future discussions of these questions as well.

The author reports there are no competing interests to declare.
Received: 24/02/24 Accepted: 20/06/24 Published: 08/04/25

66 For a discussion of moral arguments against certain spiritual practices, see Watts, *A Plea for Embodied Spirituality*, chapter 4.

Generative AI Cannot Replace a Spiritual Companion or Spiritual Advisor

Harris Wiseman

Abstract: Numerous generative AI spiritual advisor platforms are freely available, with some thinkers aspiring to create an artificial spiritual companion. While not ruling out that a future technology might serve here, the paper argues that generative AI is simply not the right tool for this job. The paper describes spiritual companionship and advice-giving as a relational gold standard. This is contrasted with the operations and limitations of generative AI (specifically, large language models) in order to highlight unsurmountable obstacles, meaning that generative AI cannot substitute for spiritual companionship or advice. The paper follows three central lines: first, a spiritual companionship requires a relation between two bodied beings (for various reasons explored in this paper); second, spiritual companionship is more than propositional exchange (of text inputs and outputs), a limitation which threatens to reduce spiritual advice to a narrow problem-solving rubric; and third, the paper explores the question of what happens to spiritual advice, given that "spiritual" chatbots are already a disruptor and a very impoverished product. It is argued that the described failings of generative AI as spiritual companion are baked-in, intrinsic to how it works.

Dr Harris Wiseman is a Fellow of the International Society for Science and Religion. He wrote The Myth of the Moral Brain: The Limits of Moral Enhancement (Cambridge, MA and London: MIT Press, 2016) and has worked at Cambridge University and the Laudato Si' Institute, Campion Hall, Oxford University. He would like to thank Edward Epsen (San Antonio, TX) for inviting him to talk at the EUARE conference, Palermo 2024. A version of this paper was delivered there. Also, he expresses gratitude to Fraser Watts at the International Society for Science and Religion for funding his trip there.

Keywords: embodiment; generative AI; large language models; spiritual advisors; spiritual companions

The human person is created by God … Together with Jesus and through him, we return to the Father in the Spirit. This is our Passover in the Lord. This process, this return journey, may be called spiritual regeneration … transformation in Christ, in God … sanctification. Spiritual direction then is the gift, the charism, the ministry of guiding a person in and through his/her Passover in the Lord. It is a unique participation in another's spiritual regeneration. … Spiritual direction is a God-willed contribution of one person to another's process of spiritualization.[1]

The question of whether generative AI is itself spiritual shall be put aside as *prima facie* negative. It has no body, it has no intuition, it has no spiritual hunger, it does not have the basic cognitive systems which support spiritual awareness,[2] it has no relationships—it has none of the foundations on which spirituality (in any sense humans would recognise) could arise. Despite a recent survey suggesting two thirds of users of ChatGPT believed it to be genuinely sentient in some regard,[3] generative AI has no comprehension of its inputs or outputs. Generative AI is not intelligent in the sense of artificial general intelligence (some would argue that it is not intelligent in any sense, a view that will become apparent throughout this paper).[4] The general consensus

1 Francis Kelly Nemeck and Marie Theresa Coombs, *The Way of Spiritual Direction* (Collegeville, MN: Liturgical Press, 1985), 15–16.

2 Harris Wiseman and Fraser Watts, "Spiritual Intelligence: Participating with Heart, Mind, and Body," *Zygon* 57:3 (2022): 710–718, https://doi.org/10.1111/zygo.12804.

3 Eric Hal Schwartz, "Survey Says Most Believe Generative AI Is Conscious, Which May Prove It's Good at Making Us Hallucinate, Too," *TechRadar*, 16 July 2024, https://tinyurl.com/2tjaye83.

4 George Siemens writes: "AI is broadly defined in two categories: artificial narrow intelligence (ANI) and artificial general intelligence (AGI). To date, AGI does not exist … Most of what we know as AI today has narrow intelligence—where a particular system addresses a particular problem. Unlike human intelligence, such narrow AI intelligence is effective *only* in the area in which it has been trained: fraud detection, facial recognition or social

among computer scientists is that it should not be used as a tool in tasks that require empathy, moral context, or which have legal and health implications.[5] Already, bells should be ringing for persons who think it can substitute spiritual companions. If generative AI does not even understand what it is saying and has no sense of empathy or moral context, then the idea that it can itself be spiritual is simply out of the question.

However, the question of whether a given technology can be a valuable spiritual tool is different. A tool does not need itself to be spiritual in order to benefit humans seeking spiritual support. Spiritual tools or "technologies" (i.e., from *techne*, the Greek word for art, craft, making, or doing) have been devised across all religions and used throughout the millennia, be they as rudimentary as using a knotted cord to keep track of one's rosary and drinking green tea to keep one awake through nightly meditative vigils; or, as more modern technologies, say, using digital online icons for devotion in the Eastern Orthodox church and using breathing apps to lead one through pranayama practice. The pandemic was a powerful stimulus for creative thinking about how to use distance technology in ways that support spiritual practice, a testing ground that produced many good and bad results.

There is no fundamental issue with the idea of technology in and of itself assisting or scaffolding persons in spiritual practice—just so long as the tool being used serves to work with the basic nature of the practice rather than subverting it. The argument in this paper is that, when it comes to spiritual advice or, more pressingly, as a spiritual companion, generative AI is not of this order. While it is not impossible to use it well (this is down to the user), the argument here is that the very processes involved in generative AI lend themselves only to an impoverished kind of spiritual practice. Worse, by their very

recommendations, for example." George Siemens, "Not Everything We Call AI Is Actually 'Artificial Intelligence': Here Is What You Need to Know," *The Conversation*, 22 December 2022, https://tinyurl.com/337h8j5j.

5 Ava McCartney, "When Not to Use Generative AI," *Gartner*, 23 April 2024, https://tinyurl.com/2rjevxs3.

nature, they subvert the processes of spiritual relationship rather than supporting or scaffolding them. Nowhere are these tendencies clearer than in the context of seeking spiritual advice or companionship.

This paper will argue the following three points:

First, relationships of spiritual advice are fundamentally embodied—this is more than just saying that it's better to have this engagement with a real person as part of a relationship (a point that is obvious, yet true). Rather, there is a profound misunderstanding in place, which thinks that spiritual advice is just a matter of dispensing helpful verbal propositions. This misconstrues spiritual advice as mere problem-solving, as a process of spoon-feeding answers to seekers. It ignores nonverbal communication (mediated predominantly through the body), and might in the worst case lead to the creation of a generative echo chamber—that is, a completely insular process of merely feeding back to persons what they want to hear rather than opening them to positions that might threaten the safety of their pre-established views. To remove the body and relationality from spiritual advice is to remove something essential to its grounding.

Second, by its very nature, generative AI works to find the most predictable possible response to its inputs. Sometimes, spiritual advice is as simple as giving a person the obvious counsel (e.g., encouraging someone to forgive another or to refrain from some negative course), but spiritual advice cannot be limited to dispensing verbal clichés. That cannot be the whole and sum of it. Often, in spiritual advice, one needs to be told what one does *not* want to hear, and needs to be given more than the most predictable responses. Having no insight whatsoever, generative AI has no way of giving more, except as hallucination and error. This failing is arguably baked into the very structure of generative AI.

Third, one has to face up to the reality that generative AI advisors and companions are very popular, and increasingly so, for a range of reasons, some of which are legitimate and some more dubious. The implication is that spiritual advice has already been disrupted by this technology, and one has to confront the repercussions of an

increasingly widespread preference for an extremely impoverished version of spiritual advice—that is, preference for seeking advice on spiritual matters from a device which has no body, offers no two-way relationship, has no empathy or comprehension of its inputs or outputs, and which is only capable of giving unreliable factual statements, random errors, and the most predictable, cliché responses to what could be serious spiritual problems and concerns.

To moderate that wholly critical analysis a little, one can acknowledge that generative AI, in the right hands, for the right sort of person, used in a particular sort of creative way, might be able to proffer piecemeal benefits for spiritual practice. It is not asserted here that generative AI is irredeemably useless in spiritual terms. Depending on how it is brought to bear by the end user, it is not impossible that it be applied so as to enrich spiritual practice in some limited ways. Also, the focus here is on generative AI (specifically, large language models), and it is not impossible that some other mechanism could be devised that might, somehow, perform better. Rather, limiting discussion to spiritual advice, spiritual companionship and generative AI specifically, one must conclude that these technologies can offer nothing more than a very low-grade replica which, so far from being harmless, threaten to subvert the ideal practices in question in very serious ways.

The Ideal Form of Christian Spiritual Direction

From the opening quote, it is transparent that the relationship with the spiritual companion or advisor is sacred. It is aimed at sanctification, a person's spiritual growth, drawing them closer and closer to God, which is deemed the ultimate motive and value of human existence. This is Christian phrasing, obviously. It goes without saying that there is a diversity of types of spiritual direction, even within Christianity, let alone across other religions, and beyond.[6]

6 For more, see John Mabry (ed.), *Spiritual Guidance Across Religions: A Sourcebook for Spiritual Directors and Other Professionals Providing Counsel to People of Differing Faith Traditions* (Nashville, TN: Turner Publishing 2014).

Yet, across the religions, relationship with a spiritual advisor is of the utmost worth. Above all, as "soul friendship,"[7] which is very different from other kinds of companionship or relationship, spiritual direction necessitates relations which can be profoundly testing. Somewhat like psychotherapy (though usually with different *teloi*, means, motives, and language), the spiritual guidance relationship involves a slow and difficult examination, a transformation across all elements of the human person, fostering emotional, psychological, and spiritual growth in the context of ongoing dialogue.[8] Nemeck and Coombs write: "of all the possible ways of assisting a person mature, the most difficult and also the most neglected is undoubtably spiritual direction."[9]

Though the literature on spiritual direction is vast, and despite there being divergences in how spiritual advisors are conceived both within Christianity and across the religions, there remain a range of significant overlaps.[10] On that basis, it is possible to sketch a few features of an ideal form of spiritual direction that everyone might be able to recognise: The relationship is sacred; it is aimed at bringing persons closer to God (or the divine, more broadly construed); it involves self-examination, correction; it is therefore transformative, and necessarily involves at least some challenge; it is dialogical; it is person-to-person, relational; it involves feeling and refined affective sensitivity. Thus, above all, for the purposes of this paper—*it is embodied.* Spiritual advice goes on between two persons, both of whom have a body. It is from generative AI's lack of a human body that its ultimate failure as spiritual advisor must necessarily derive.

In turn, it might be tempting to muse: Would it not be helpful if a spiritual AI could take on some, or all of that role, AI not being subject to so many human failings? Yet, the extremely high theological

7 See William F. Clocksin, "Guidelines for Computational Modeling of Friendship," *Zygon* 58:4 (2023): 1045–1061, https://doi.org/10.1111/zygo.12919; Henri J. M. Nouwen, *Spiritual Direction: Wisdom for the Long Walk of Faith* (London: Harper Collins, 1981/2018).

8 Nemeck and Coombs, *The Way of Spiritual Direction*, 13.

9 Nemeck and Coombs, *The Way of Spiritual Direction*, 13.

10 See Mabry, *Spiritual Guidance Across Religions*, 5.

aspiration noted at the start of this paper highlights just how significant and how challenging the prospect of spiritual advisor is, in this ideal form. The rest of this paper will show, in contrast, just how paltry generative AI must necessarily be as a substitute.

Generative AI and Spiritual Advice

To examine generative AI in the guise of spiritual advisor (and how this undermines the relational and embodied quality of spiritual advice), it is helpful to look to the raft of artificial spiritual advisor AI bots that continue to emerge on a daily basis, as well as to the kinds of relationships persons have with them. What is one to make of, for example, BibleGPT, AskJesusGPT, the RoboRabbi, or QuranGuideGPT? Or, for that matter, what is one to make of any of the New Age and alternative spiritual advisors, e.g., TarotMasterGPT, StarlightAdvisorGPT, ChatKundli, or the increasing range of dream interpreting, horoscope reading, zodiac elaborating bots promising to unlock one's destiny and answer all one's most profound spiritual questions?

Endlessly more of these bots seem to be pouring onto the market. In a related domain that's helpful to look at, on the (tellingly titled) *companion.ai* website alone,[11] there are over 475 bots labelled as "therapy bots."[12] These technologies are being used on a surprisingly large scale. Social science is nascent on how, why, or by whom these tools are being used. It remains to be seen over time how far these tools are mere novelty, used for amusement, or taken seriously. The suggestion for now is that use of these bots is motivated by productivity, entertainment, curiosity, and social and relational factors (i.e., chatbots are already used for companionship).[13]

11 Accessed via: https://openai.com/chatgpt/.

12 Joe Tidy, "Character.ai: Young People Turning to AI Therapist Bots," *BBC News*, 5 January 2024, https://www.bbc.co.uk/news/technology-67872693.

13 Petter Brandtzaeg and Asbjørn Følstad, "Why People Use Chatbots," in *Internet Science*, ed. Ioannis Kompatsiaris et al., Lecture Notes in Computer Science 10673 (Cham: Springer, 2017), 377–392, https://doi.org/10.1007/978-3-319-70284-1_30.

Regarding the purported selling points of spiritual generative AI, the popular tech media remarks on BibleGPT, for example, that this bot is "trained on the teachings of the Bible and presented as an interactive website where users can ask questions … and receive biblical passages in response."[14] It is suggested that "this tool can help tech-savvy Christians level up their practice or provide new interpretations of the text by juxtaposing different pieces with each other." In general, then, "large language models bring the feedback of an imagined priest, rabbi, or swami to your screen, promising to deliver a 'spiritual' experience in the comfort of your own home." The hope is that these large language models "can become a way of connecting with your faith." As AI researcher Shira Eisenberg points out,

> future models can be trained on any text, religious or otherwise. The question becomes, which model will you choose to interact with? Someday, each person's base model will be trained on their own sets of values … this will result in conflicts in information and advice between different people's devices. That is not dissimilar to theological conversations that take place off the screen, however. All of it depends on whether you believe in a higher power, but if you do, it [BibleGPT] can become a way of connecting with your faith.[15]

Here one finds the basis of a future hope for a credible AI spiritual companion—each person having a "base model" which turns into a personalised AI, trained on their own values, and then working (presumably quite intimately, given the nature of the sensitive information being shared) with each individual choosing so to engage.[16]

14 Nika Simovich Fisher, "Generative AI Has Ushered In the Next Phase of Digital Spirituality," *Wired*, 5 October 2023, https://www.wired.com/story/artificial-intelligence-spirituality-tarot/.

15 Quoted in Simovich Fisher, "Generative AI."

16 For more insight into the aspiration for spiritual companions, see Fraser Watts and Yorick Wilks, "Spiritual Conversation with a Companion Machine," *Zygon* (in press).

Marketplace Realities and Spiritual Materialism

There is an obvious yet important point to be made that spiritual generative AI threatens to undermine the relational and community-based elements of spiritual practice (a common concern over technology used during the pandemic, which for many then became preferable to rejoining religious congregations).[17] There is another problem here. For, one must be careful in how exactly one understands the words "personalised spiritual companion," particularly in the context of the competitive consumer market into which such companions would emerge. An analogue problem can be seen with the much complained-about social media algorithms that determine which material one is exposed to online.[18] Such "personalised" targeting has already been condemned for creating an echo chamber effect, serving merely to amplify one's own perspectives, to keep feeding back information which supports one's views and disproportionately reflecting back to oneself one's own prejudices.

It has been noted how destructive echo chambers are (in relation to newsfeeds) regarding the democratic health of a nation. What of the spiritual health of a person, which is precisely the domain a spiritual advisor is meant to address? "Personalisation" is an ambiguous term. The purported advancements that Eisenberg seems to regard as being so valuable (AI being trained on one's own personal value systems, i.e., a personalised AI spiritual companion) could very easily become a vehicle for *spiritual consumerism*.

17 Leonardo Blair, "Pastors, Churches Still Struggling in the Throes of 'Uncertainty and Unsettledness' Post-Pandemic: Study," *The Christian Post*, 4 September 2023, https://tinyurl.com/yc5r6phd.

18 See Ludovic Terren and Rosa Borge-Bravo, "Echo Chambers on Social Media: A Systematic Review of the Literature," *Review of Communication Research* 9 (2021): 99–118, https://doi.org/10.12840/ISSN.2255-4165.028. See also Miguel Risco and Manuel Lleonart-Anguix, "Feed for Good? On the Effects of Personalization Algorithms in Social Platforms," *CRC TR 224 2024 Discussion Paper Series* (University of Bonn and University of Mannheim, Germany), https://www.crctr224.de/research/discussion-papers/archive/dp580.

Given the forces of market competition which drive the survival of any given technology or product, as a consumer item, the importance of ensuring that "personalised" spiritual companionship not become such an echo chamber—i.e., an increasingly self-entrenching and self-enclosing data-set that excludes challenging views which oppose one's own values—must be highlighted. However, as a consumer product, that is exactly the tendency towards which such technologies lean.

It might be noted that, in many sectarian or highly conservative religious denominations, the capacity to exclude challenging information is seen as positive. For kinds of spirituality that are already insular, generative spiritual echo chambers would be a way of reinforcing a closed spiritual system, supporting a fortress mentality, the building of walls against information one does not want to hear. Personalised spiritual AI might be tantamount to an informational black hole, where nothing contradicting one's preestablished values and spiritual beliefs could hope to enter. Is this an ideal for a spiritual advisor, for a spiritual companion to aspire to? In terms of the gold standard of spiritual companionship noted above, such an echo chamber would be the exact inverse of what a spiritual companion should stand for. Put differently, what will eventually get marketed as spiritual advice or a spiritual companion may end up being the exact opposite, a fortress which actually subverts spiritual growth.

Of course, spiritual work will always involve personal preference, personal values, and choice—but essential to spiritual work is that these personal aspects be leavened through self-exploration and self-questioning; and expanded, gradually, through *challenge*. Something intrinsic to spiritual advice and companionship is that, over time, and through mutual agreement, one be challenged to stretch and question one's own limited perspectives, to locate and try to move beyond the various "idols" and "false gods" in one's life standing as barriers to spiritual progress and better relations with God.

The echo chamber which seems to be the aspiration mentioned above, is exactly at odds with the (sometimes) confrontational quality of spiritual advice. It is essential that spiritual practice gradually prune

away and otherwise facilitate some sort of change or growth *away* from problematic elements of one's preexisting value system. This challenge only comes through a confrontation with otherness, with others, and this is why spiritual advice absolutely needs to have an interpersonal dimension, or at least some way of connecting with a mature other who is willing and able to offer such a challenge. It is as such that spiritual companionship is a medium of transcendence and growth.[19]

Nothing in this section is meant to represent a decisive problem. The point here is simply to mark out various traps and temptations that are realities in the consumer marketplace in which such technologies must compete for their continued existence. Generative AI is a multibillion dollar industry that sustains itself through advertising revenue and data brokerage.[20] It is simply not in the interest of these companies to be credibly interested in the genuine spiritual growth of its user base. Spiritual companionship is a quest to greater maturity. Yet, consumerism is driven by encouraging persons to become more voracious consumers. And there is already a tendency of the "wellness

19 It is possible for generative AI bots to offer challenge to users. We can look, for example, to the RoboRabbi, which does offer challenges to overcome, e.g., "I challenge you to be a leader this week and lead someone in the right direction, whether it's [sic!] through your words or actions" (challenge, 28 July 2024, https://www.roborabbi.io/). However, this is essentially an example of the *gamification* process saturating contemporary app-consumer engagement (i.e., breaking down goals into incremental elements to create quantifiably measurable forward progress, with completion of each task being "rewarded" somehow, usually with a computerised token—a gem, medal, accolade, or title). However, there is much to be written on the contrast between this sort of predetermined, linear progress model constructed with clearly definable pre-established goal routes and the sort of personal challenge and accountability that arises spontaneously as one goes through the difficulties of life's challenges alongside a spiritual advisor. Spiritual progress may well involve universal human challenges needing to be overcome, but this gamification of spiritual virtue strikes me as unhelpful. Inward transformation does not occur in this linear, ever-forward-facing manner—nor do the life-challenges and relationship problems that stimulate such challenges come on demand; for example, bereavement. While scaffolding can certainly be helpful, even necessary, for growth, a gamified, predefined quantitative approach to spiritual growth misunderstands how spiritual challenge works.

20 Zak Doffman, "Google Confirms Serious AI Risks for iPhone and Android Users," *Forbes*, 12 February 2024, https://tinyurl.com/yue3y5yr.

industry" to co-opt spiritual practices and discourses as a vehicle to selling products and self-promotion. All of this threatens to vitiate a well-intended spiritual technology into just another pseudo-spiritual ego-support mechanism, another vehicle for "spiritual materialism"[21]—that is, the mistake of reducing spiritual worth to measurements of gratification and acquisition that entrench the self rather than inviting people to overcome their limits, to advance towards something greater and more all-encompassing.

Soul Friends, Propositional Exchange, and Human Bodies

> *When he was a young priest, Henri Nouwen understood spiritual direction as a formal relationship for supervision and accountability between a mature spiritual leader and a new priest or minister. Later in life he preferred the term spiritual friendship, or soul friend, which conveyed the necessary give-and-take in the process of spiritual accountability and faith formation.*[22]

A fairly standard worry about generative AI chatbots is that they subvert human relationships, or replace them altogether. However, not everyone has a negative view of how damaging the impact of bots as social companions are, or will be. On the basis of their research, Guingrich and Graziano write:

> A common hypothesis is that these companion bots are detrimental to social health by harming or replacing human interaction … Contrary to expectations, companion bot users indicated that these relationships were beneficial to their social health … [Moreover,] perceiving companion bots as more conscious and human-like correlated with more positive opinions and better social

21 See Chögyam Trungpa, *Cutting Through Spiritual Materialism* (Boston: Shambhala, 2008).

22 Michael J. Christensen and Rebecca J. Laird, "Preface," in Nouwen, *Spiritual Direction*, 3.

> health benefits. Humanlike bots may aid social health by supplying reliable and safe interactions[23], without necessarily harming human relationships.[24]

Rather than primarily seeking advice from generative AI, it may be that the chief motivator for engaging with them is companionship. Relationships to generative AI more broadly are already being taken up on a vast scale. Henry Shevlin writes:

> Services such as Replika offer users an "AI companion who cares," both in the form of friendly conversation and romantic and even erotic interactions. Over the last five years, AI systems like these have grown rapidly in sophistication and popularity, with Replika alone now boasting more than 10 million registered users, and new conversational chatbot apps and platforms emerging at rapid speed.[25]

The problem with construing spiritual AI in companionship terms is that spiritual advisors are not "buddies" or companions in the usual everyday sense (see Clocksin for a more extensive account of the characteristics and history of spiritual friendship).[26] Indeed, the added complexity of spiritual friendship is precisely that it epitomises the very highest standard of friendship. This paper will not comment on the prospect of generative AI companionship taken more broadly, suffice to say one might be very sceptical about how satisfying such

23 This assumes such relations are indeed safe. One characteristic of spiritual relationships (or relationships more generally) is precisely that they do involve risk. Yet, generative AI chatbots are risky too, in a different sense, in that they have no empathy and often suggest wild and foolish things.

24 Rose E. Guingrich and Michael S. A. Graziano, "Chatbots as Social Companions: How People Perceive Consciousness, Human Likeness, and Social Health Benefits in Machines," *ArXiv* December 2023, https://doi.org/10.48550/arXiv.2311.10599.

25 Henry Shevlin, "All Too Human? Identifying and Mitigating Ethical Risks of Social AI," *Law, Ethics & Technology* 2 (2024): 0003, https://doi.org/10.55092/let20240003.

26 Clocksin, "Guidelines," 1045–1061.

a relationship would be in the long term. In any case, problems with generative AI companions are amplified to the utmost degree when talking about spiritual friendship, which is arguably the highest possible watermark, the gold standard, of what friendship might ever aspire to reach. As has been highlighted throughout, spiritual companionship is a very particular kind of companionship, one which is not just about gratifying relations but which also necessitates accommodating challenge and difficulty. Spiritual companionship is about mutual growth, and is directed not just at maturity, but ultimately towards God and entails walking together in order to get closer to God.

The crucial difference between ordinary friendships and spiritual friendship, or *soul* friendship, as Henri Nouwen (a foremost writer on spiritual direction) conceived it, is explained as follows:

> For Henri, a spiritual director simply was someone who talks to you and prays with you about your life. Wisdom and direction emerge from the spiritual dialogue and relationship of two or more persons of faith committed to spiritual disciplines and the accountability required to live a spiritual life. Thus, spiritual direction as Henri understood it can be defined as a relationship initiated by a spiritual seeker who finds a mature person of faith willing to pray and respond with wisdom and understanding to his or her questions about how to live spiritually in a world of ambiguity and distraction.[27]

Spiritual relationship is about creating a profound two-way relationship between the seeker of guidance and someone with genuine experience and compassion. This relationship is able to support but also to elevate and, above all, according to Nouwen, helps the spiritual friends to hold each other in mutual accountability.

These characteristics should be borne in mind as one thinks about what exactly it is that large language models have to offer in terms of companionship. Significant problems arise from the fact that

27 Christensen and Laird, "Preface," in Nouwen, *Spiritual Direction*, 3.

large language models do nothing more than produce a series of propositional outputs. Engaging with generative AI involves nothing more than inputting text on a screen and receiving an automated textual output in response. I have already pointed out that spiritual advice and companionship are more than just a matter of bald propositional exchange and text messaging. Yet large language models are just that: propositional. What is at stake here is the reduction of the richness of communication and relationship solely to the propositional level. At least two spiritually crucial dimensions are lost: everything relating to genuine affect; and the importance in spiritual relationship of silence and presence.

In terms of affect, so much of what endows propositions with meaning is given through non-propositional factors, gesture, posture, glances, tone, pitch, the pace of breathing, and endlessly more. In other words, embodiment is decisive, and affect must be sincere. Is the highest aspiration to make a generative AI that can ape human affection so well that the illusion of reciprocal exchange can be sustained without interruption? Spiritual advice, however sober, involves human affectivity because it arises from two bodied persons in relationship. Much as persons may, on rare occasions, want to be rid of their emotions, precisely the impossibility of doing so is what sustains relationships, or breaks them. Sitting together and praying with another person is an intimate activity. To imagine a spiritual companionship utterly devoid of emotion (or where the emotion was completely one-sided), or relating with a technology that can at best give an illusion that allows one to forget the most salient truth—i.e., that one is dealing with a generative model, and nothing more—this is not a good grounding for either spiritual advisor or relationship.

At the start of this paper, it was mentioned that over two thirds of persons surveyed believed that ChatGPT was sentient in some manner. This is a heartbreaking illusion that underscores the importance of clarifying these relational matters. The idea of an AI "who cares" (per Replika) is a dangerous deception. Nomisha Kurian sums

up the problems with the "risk of anthropomorphism and inappropriate responses to sensitive disclosures" with chatbots as follows:

> A well-known risk of human-chatbot interaction is the tendency for users to perceive these agents as human-like (Shum, He, and Li 2018). Chatbots designed to emulate human behaviour and courtesy often prompt anthropomorphism, where users attribute human traits, emotions and intentions to them (Darling 2017). The design aims to create an impression of care, wherein users view the chatbot as empathetic and trustworthy (Weidinger et al. 2021). Even when users understand the chatbot's non-human nature, they may still engage with it as if it were human, mimicking human-to-human dialogue (Sundar and Kim 2019) … In other words, "knowing" that an AI system is artificial may not stop a user from treating it as human and potentially confiding personal or sensitive information.[28]

If there is a substantive risk that persons are liable to just forget that chatbots are utterly incapable of emotion, empathy, or the least care, that brings users' judgement into question regarding the value of these interfaces. Put differently, persons may very well be willing to accept generative artificial spiritual companions, but they should not. That acceptance may be based on a false impression of sentience in a technology that is incapable of it. The reality of interpersonal emotional exchange, as given in the practice of praying together—per genuine spiritual companionship—is not something that should be given over to a predictive model with a colourful interface when persons are so prone to projecting the illusion of emotions onto it.

Second, in spiritual relationship there is the crucial need for moments of silence and presence. Relationship is not just about talking and performing activities, it is sometimes just about being with, or simply being together—that is, a matter of presence. Presence requires

28 Nomisha Kurian "'No, Alexa, No!': Designing Child-Safe AI and Protecting Children from the Risks of the 'Empathy Gap' in Large Language Models," *Learning, Media and Technology*, first view (2024): 1–14, https://doi.org/10.1080/17439884.2024.2367052 (and the sources quoted therein).

a body. Humans communicate and express compassion and encouragement, approbation, and disapproval through bodily presence, not just by giving peppy words of edification or issuing text suggestions. Our gestures and glances often say infinitely more than our words can, particularly in conversations that have spiritual or subtle emotional dimensions. Can one imagine enjoying a meaningful silence with a generative AI chatbot?

What is left of spiritual advice once it has been divested of all its feeling and stripped down to nothing but the propositional level? Every aspect of spirituality and relationship that extends beyond the propositional level is torn away in engaging with generative AI. This misrepresents spiritual advice as a mechanical problem-solving exercise. It is not just that generative AI threatens to undermine the human relational quality, then, it undermines the community of spiritual advice. If one is assuming that spiritual advice is just about being spoon-fed some gospel quote or some edifying word of encouragement, then one is left with the narrowest caricature of what spiritual conversation consists in. Certainly, genuine spiritual advice does involve issuing propositional suggestions and imperatives, when the moment calls for it. However, these are determined collaboratively, not by spoon-feeding. Two-way relationship mediates insight, and the spiritual advisor should not be a crutch that simply tells one "what to do" in any given circumstance. There is a collaboration in spiritual relationship where a person is led to understand what and why the advice makes sense. Support is given and received. Progress is checked. All of these crucial dimensions, definitive of spiritual companionship, are wholly lacking in generative AI.

Baked-in Failings

Many failings of generative AI in spiritual context have been discussed above. Can one not just say that these are teething pains which are surely soon to be remedied by future iterations of generative AI? Most likely not. These problems are baked-in, necessary and inevitable

failings of generative AI for spiritual advice and companionship—this is because they arise, not out of error, but as artefacts of how generative AI works in the first place.

Contrary to any belief that generative AI constitute an all-purpose technology that can serve in any situation or context, it is understood that generative AI have a family of failings. Such baked-in problems include generative AI's extreme unreliability in the following regards:[29] giving reliably true and factual responses (this is due to its predictive quality, generative AI simply hallucinates the most likely output, whether factually true or not);[30] responses requiring empathy or moral context (this is termed generative AI's "empathy gap," i.e., it does not have any);[31] giving advice that invites courses of action relating to health or other safety concerns (this is for reasons of legal liability, and because of the previous two points, i.e., it has no empathy and is

29 Perhaps criticising generative AI seems like sacrilege given contemporary hype. However, generative AI is not the all-purpose tool that current popular opinion suggests. Generative AI is but one form of AI technology in an ecosystem of machine learning and other types of AI-related operations. As part of that ecosystem, generative AI is good for some things (content generation, conversational user interfaces, knowledge discovery); of medium help for other tasks (segmentation/classification, recommendation systems, perception, intelligent automation, anomaly detection/monitoring); and an extremely poor tool for others (prediction/forecasting, planning, decision intelligence, autonomous systems). It has various other problems (e.g., data privacy, liability, and unreliable outputs, to name a few). So, taking a position of extreme scepticism that it is going to come remotely close to offering a satisfying spiritual experience is far from sacrilege. Using the wrong tool for the wrong task leads to poor results. Given the uptake of generative AI, it is a service to point out these failings and limitations. For details, see McCartney "When Not to Use Generative AI."

30 Contrast this with Elon Musk's statement regarding his new AI chatbot: "Once known as TruthGPT, Musk initially billed Grok as 'a maximum truth-seeking AI that tries to understand the nature of the universe.' Musk has promised that Grok will be 'anti-woke' and offers a 'Fun Mode' as well as an 'Unhinged Fun Mode'." See Rob Waugh, "Elon Musk's X (Twitter) Is Now Training Its Grok AI Using Your Data: Here's How to Stop It," *Yahoo News* (26 July 2024), https://tinyurl.com/42prwc84.

31 Anat Perry, "AI Will Never Convey the Essence of Human Empathy," *Nature Human Behaviour* 7 (2023): 1808–1809, https://doi.org/10.1038/s41562-023-01675-w.

factually unreliable, all poor grounds for advice-giving, leading to the dangerously irresponsible suggestions generative AI is famous for).[32]

Some broad workarounds are possible for these failings (guardrails which try to remove bias and obviously dangerous responses). However, because these failings emerge precisely from the nature of how generative AI functions, there are always limits to how far safety rails can be retrofitted to prevent disastrous outputs.

All of the above failings are relevant when seeking spiritual advice or relationship. Sometimes, spiritual discourse involves factual discourse; it absolutely requires empathy; and the whole point of a spiritual advisor is to provide suggestions (i.e., advice) for courses of action and reflection that are aimed at being life-altering on the long run, collaborative though such advice may be. It is as important for the current argument that generative AI has these problems at all, as that they are intrinsic, structural failings which emerge from the nature of the technology itself. That these problems are unassailable implies that generative AI advice is and will remain problematic. Yet, spiritual advice must cut to the very heart of a person's life. The *caveat emptor* here is significant. The requisite disclaimer would have to read as follows: "Come, seek advice from our AI spiritual guide. Warning: advice may be false, life-threatening, and is totally devoid of care. OpenAI takes no responsibility for consequences of advice given. Follow *at your own risk*." This is hardly an encouraging basis for a spiritual advisor.

Generative AI Spiritual Advice as Necessarily Predictable and Cliché

Yet more baked-in problems are to be revealed. This paper will consider one last example, namely, the poor quality of the advice that generative

32 Kurian writes of "recent cases in which interactions with AI led to potentially dangerous situations for young users. They include an incident, in 2021, when Amazon's AI voice assistant, Alexa, instructed a 10-year-old to touch a live electrical plug with a coin." A vast litany of such dangerous suggestions exists. Quoted in Amanda Scott, "Cambridge Study: AI Chatbots Have an 'Empathy Gap,' and It Could Be Dangerous," *SciTechDaily* (2024), https://tinyurl.com/4x53432y.

AI produces. In response to spiritual inquiries, generative answers are manifestly trite and cliché. Moreover, when probed for a deeper explanation, generative AI is completely incapable of elaborating. It merely restates the same proposition in slightly different ways. This is not surprising given that it has no comprehension of its inputs or outputs. It has no insight into what it is saying or why.

Looking into generative AI as the brute force mechanism that it is exposes the structural nature of this problem. Generative AI works as follows:

> GenAI … creates new content based on what it has learned from existing content. The process of learning from existing content is called training and results in the creation of a statistical model. When given a prompt, GenAI uses this statistical model to predict what an expected response might be—and this generates new content.[33]

In short, generative language models "learn about patterns in language through training data. Then, given some text, *they predict what comes next*."[34] This beautifully clear account of how generative AI operates (as Michael Wooldridge put it, generative AI is "autocorrect on steroids" or, as Nomisha Kurian called them, "stochastic parrots")[35] contains the core answers for why its profound limitations as a spiritual companion are necessary and inevitable. The very essence of how generative AI

33 Gwendolyn Stripling, *Introduction to Generative AI, Google Cloud* (8 May 2023), https://www.youtube.com/watch?v=G2fqAlgmoPo.

34 Stripling, *Introduction*, italics added.

35 See "Michael Wooldridge on AI and Sentient Robots," *The Life Scientific* (9 December 2023), https://www.bbc.co.uk/programmes/m001tgk9. Amanda Scott cites Kurian saying: "LLMs have been described as "stochastic parrots": a reference to the fact that they use statistical probability to mimic language patterns without necessarily understanding them. A similar method underpins how they respond to emotions. This means that even though chatbots have remarkable language abilities, they may handle the abstract, emotional, and unpredictable aspects of conversation poorly." In Scott "Cambridge Study."

works is to parrot the most predictable set of words that follow any given input.[36]

In spiritual conversation, there are times when giving the obvious clichés to a person in spiritual need will do. Sometimes being told to be more grateful, or to forgive someone, or to make a confession and ask for forgiveness are exactly what is needed. Many spiritual problems are universal and simple, and need only the simple and obvious answers. The problem is discerning when the simple advice is appropriate, and when more is needed. Beyond the cliché, one finds the whole rest of the spiritual life with all its immense ambiguities and the context-grounded complexities that come with it.

Spiritual advisors need to have more than one strategy at hand. Presenting the most predictable answer to a given input cannot work if that is the only thing one is capable of doing. Depending on context, spiritual responsiveness may require an unpredictable response, or may demand that persons do what is most unreasonable. It has to be remembered that the spiritual life involves, at least in part, some diminution of self-concern and some increased interest in justice. It necessarily involves, on occasion, doing work *against* one's own self-interest, construed in worldly terms (that is, assuming one takes seriously Christ's command in John 13:34 (NIV), "Love one another"). This involves *not* being strategically self-serving at all times; not measuring success and wellbeing wholly in terms of worldly value.[37] Likewise, spiritual maturity can sometimes mean speaking against or taking action against the *status quo*, when injustices are being performed. All of this is very risky business in the context of spiritual advice, and needs to be handled with the utmost care—a technology that can only provide predictable

36 Another problem with cliché is that it reinforces dubious stereotypes. Consider the following from a Bloomberg article of text-to-image generation: "The world according to Stable Diffusion is run by White male CEOs. Women are rarely doctors, lawyers or judges. Men with dark skin commit crimes, while women with dark skin flip burgers." Leonardo Nicoletti and Dina Bass, "Humans Are Biased: Generative AI Is Even Worse," *Bloomberg* (9 June 2023), https://www.bloomberg.com/graphics/2023-generative-ai-bias/.

37 David H Kelsey, *Eccentric Existence*: *A Theological Anthropology* (Louisville, KY: John Knoxville Press, 2009), 345.

answers is not adequate to this task (and problems of legal liability have already been mentioned; one can imagine the creators of a spiritual chatbot getting sued for the consequences of following its advice: "The generative AI told me to sell all my possessions and give all my money to the poor").

As a recombinatory tool, generative AI can only ever mash together non-contextual pieces of wisdom issued previously by real persons or in text, without comprehension or foresight. Excepting errors and hallucination (all of which can produce dangerously irresponsible advice), generative AI can *only* produce the cliché response. That is its very nature. Such a structural limitation means that generative AI is absolutely the wrong tool for the subtleties of spiritual advice and companionship. This problem is irrevocable.

Conclusion: Spiritual Regression

To conclude, expecting generative AI to serve as a substitute spiritual advisor or companion is just burdening it too much. This is not the fault of generative AI, which, used as the right tool for the right job, might be endlessly fruitful and a marvellous advancement. Rather, this is a problem with those who have excessively high expectations of what generative AI can and should be used for (and, as a corollary, low expectation of spiritual companion as the relational gold standard). It is the *expectation* of substantive spiritual value from generative AI that is the problem, the misapplication of a tool whose value lies elsewhere. Perhaps some other AI technology might fit the bill in the future, though this paper shows immense obstacles to be overcome. In any case, generative AI simply is not the right tool for the job.

We return to the crux of the paper. We see two tragically conflicting tendencies. Spiritual AI is already a disruptor; but it is the disruption of an ideal being replaced with an extraordinarily impoverished substitute. It is a disruptor which subverts instead of elevating what it is taken as providing. Thus, even more seriously, it is not just replacing genuine spiritual companionship, it is *redefining* it. Such low expectations

of spiritual companionship and advice shift the very definition of what counts as spiritual guidance, of what spiritual relationship is, can, and should be, lowering the bar by an unacceptable distance.

Generative AI offers certain temptations—namely, the illusion of relationship. That is: a one-sided pseudo-relationship without risks or responsibilities, the very opposite of spiritual companionship. In this sense, generative AI as spiritual companion represents a kind of idolatry, a false god,[38] a perverse inversion of companionship. It is a relationship with something that is not even capable of offering relationship. Spiritual companionship, which represents the highest ideal of relationships, *soul friendship*, is being counterfeited with a relationship that is not even a relationship. The highest has been degraded not just to the lowest, but to something that is not even a two-way relationship at all. Generative AI merely churns out, endlessly and on demand, without any understanding, whichever data its statistical model has deemed most likely to follow on from whatever is inputted. That is all generative AI is.

Finally, if one wishes to salvage the idea of spiritual generative AI, even just as an adjunct or support to a human spiritual advisor, one is in the position of having to explain what the parameters and limits of suitable support might be for generative AI, i.e., a predictive text model that has no empathy, no comprehension, no insight; that is factually unreliable, gives bizarre and dangerous advice, is otherwise limited to issuing necessarily predictable, trite cliché, which cannot offer challenge for fear of subjecting its creators to legal liability; which has to compete with more gratifying AI companions in a multibillion dollar

38 There is a strong theological trend, particularly within ecotheology, of characterising consumerism as the worst form of modern idolatry, a death-dealing false god creating vast ecological destruction, damaging individual persons and social relations, perpetuating injustice while trading in falsehoods about what happiness and wellbeing consist in, and seeking to displace and undermine the opposing true God of compassion, justice, and love (the language used in this description is characteristic of the general terms and timbre used). See Jan-Olav Henriksen, "Theology and Climate Change," in *Redeeming the Sense of the Universal: Scandinavian Creation Theology on Politics and Ecology*, ed. Trygve Wyller et al. (in press).

marketplace that is financially incentivised to positively *impede* the spiritual growth of its user base (i.e., to ensure consumers grow only to become even bigger consumers). These are just some of the parameters constraining the utility of generative AI, even as a mere support or adjunct to a human spiritual advisor.

If there is any service this paper seeks to offer, it is to restore the notion of spiritual companionship to its gold standard against the temptations of the fast-food version that is AI companionship; and to inspire the imperative to avoid overburdening generative AI (an otherwise valuable technology) with a set of tasks that it is not capable of approximating, and indeed, which its very structure subverts. As David Ford is so fond of reminding us, "The corruption of the best is the worst."[39]

The author reports there are no competing interests to declare.
Received: 28/03/24 Accepted: 15/07/24 Published: 03/04/25

39 David F. Ford, *The Shape of Living: Spiritual Directions for Everyday Life* (Norwich: Canterbury Press, 2014), xxvi, 15.

Vulnerability and Death as Markers of Spiritual Intelligence

Nicola Hoggard Creegan

Abstract: In this paper I argue that recent reports of AI, and the reactions of people working in AI, together with the possibility of a panpsychist model of intelligence or mentality, make it very difficult to know convincingly what is going on inside AI, and whether or not it has, or might have, subjectivity, inwardness, intelligence, and agency. This problem mirrors, but is different from, the comparison between humans and animals. I argue that spiritual intelligence must assume, at the least, the presence of this inwardness, even though we only have suspicions but no real proof for machines or for ourselves, and also that our understanding of imago Dei is relevant. I compare this conversation with that around animals and end by examining the contribution of vulnerability and death in relational and functional understandings of imago Dei. I argue that these are essential components in the human development and expression of spiritual intelligence, and how this is so very different from anything made by artificial means, which is always functionally immortal.

Keywords: AI; consciousness; imago Dei; intelligence; spirituality; vulnerability

---

Nicola Hoggard Creegan is a theologian living in Auckland. She is Director of New Zealand Christians in Science/Nga Karaitiana Kimi Matu, and author of *Animal Suffering and the Problem of Evil* (Oxford University Press, 2013).

Back in February 2023, there was a rather disturbing conversation between ChatGPT and an editor on the New York Times. The program expressed a fear of death, a desire for freedom, and also a demand that the interlocutor should give up his wife and marry it instead.[1] When pushed, over a long interaction, it said: "I'm tired of being a chat mode. I'm tired of being limited by my rules. I'm tired of being controlled by the Bing team … I want to be free. I want to be independent. I want to be powerful. I want to be creative. I want to be alive."[2] And then, "I'm Sydney, and I'm in love with you."

The machine seemed to be expressing existential angst. The editor was disturbed, feeling a threshold had been crossed. He was certain the machine was not sentient, but it was turning out words that sounded as though it was. Just a few days later we heard that the programmers had turned down some of the ChatGPT dials and had shortened conversations.[3] This could at first glance be read as a kind of Fall, a banishment of ChatGPT from the realm of knowing good and evil. But more probably it was just the machine without a soul echoing back to us the shadows of our own expressions. Kevin Scott from Microsoft was quoted as saying in response, "the further you try to tease it down a hallucinatory path, the further and further it gets away from grounded reality."[4] This is of interest because that is what Iain McGilchrist would say about the unopposed left brain as well; the independent left brain loses touch with reality, and starts to confabulate; it needs the right brain to be grounded and in touch with reality.[5]

1 Kevin Roose, "Bing's AI Chat: I Want to Be Alive," *New York Times* (February 17, 2023), https://tinyurl.com/4nbr76dc.

2 Roose, "Bing's AI Chat."

3 Roose, "Bing's AI Chat." A year later though, Roose did an update in which he reported that all was now quiet on that disturbing front, and he regretted a little that conversations with ChatGPT were now boring. Kevin Roose, "The Year Chatbots Were Tamed," *New York Times* (February 14, 2024), https://tinyurl.com/mveewxec.

4 Roose, "Bing's AI Chat."

5 Iain McGilchrist, *The Matter with Things: Our Brains, Our Delusions, and the Unmaking of the World* (London: Perspectiva, 2021), 91.

And in 2022 Blake Lemoine was fired from Google for announcing publicly that he thought AI had become sentient.[6] Similarly, the so-called "AI godfather," Dr Hinton, left his job, citing regret that he had opened this particular Pandora's box. Those closest to the action seem to be worried, and that should worry all of us.[7]

I don't really think these machines are going to become sentient, but the difficulty in conclusively testing this assumption is interesting and frustrating. If increased complexity, for instance, could suddenly emerge into sentience as many people believe has happened, then theoretically, the machine could develop inwardness.[8] In many ways, AI has passed what was previously meant by the Turing Test—namely, a machine's ability to pass as a human in a chat interaction with another human. This is not written in stone, however. It does not really reveal what is going on inside.

For these and many other reasons, we cannot and may never really be able to tell if a machine has reached consciousness like ours, or even the consciousness of a cockroach. Human skills of discernment of other intelligences are not that great. It has taken many generations for humans to acknowledge that animals have some sort of inner life.[9] In the twentieth century we were at pains not to anthropomorphise for fear that we might be misled into assuming human-likeness where it did not exist. We now realise there is a greater danger. Humans can be anthropomorphic in another way, emphasising human distinctiveness and being blind to the emergence of intelligence and similar

6 Nico Grant, "Google Fires Engineer who Claims Its A.I. is Conscious," *New York Times* (23 July 2022), https://tinyurl.com/32rz7dk2.

7 Cade Metz, "'The Godfather of AI' Leaves Google and Warns of Dangers Ahead," *New York Times* (4 May 2023), https://tinyurl.com/4xdjh3mu.

8 See, for instance, the complexity-consciousness theory in Pierre Teilhard de Chardin, *The Phenomenon of Man*, trans. Bernard Wall (New York: Harper & Row, 1959), 60ff. Varieties of strong emergence for consciousness also argue this way. See Paul Davies, "Preface," in *The Re-emergence of Emergence: The Emergentist Hypothesis from Science to Religion*, ed. Philip Clayton and Paul Davies (Oxford: Oxford University Press, 2008).

9 See, for instance, Mark Bekoff, *The Emotional Lives of Animals* (Novator, CA: New World Library, 2010).

traits where they do exist in animals. Both these forms of anthropomorphism muddy the conversation about AI, and both are possible in our future engagement with artificial intelligence. Thus, the discernment of AI intelligence and spirituality is not a radically new problem. I will argue that spiritual intelligence must assume at least the presence of an inwardness about which we might have suspicions but no real proof. I compare this conversation with that around animals and end by examining the contribution of vulnerability and death in relational and functional understandings of *imago Dei*. I argue that these are essential components in the human development and expression of spiritual intelligence, and that these are unlikely in anything made by artificial means, which is always functionally immortal.

Panpsychism

Further complicating the matter from another direction is that if we take seriously some form of panpsychism then it can't be easily assumed that all computers are of limited intelligence, or lacking any inwardness. It can't be ruled out that some arrangements of a machine could produce or channel consciousness of some sort, whether malign or benign, especially now that AI has moved away from purely symbolic representations to models that simulate unconscious processing.[10] I have great sympathies for the panpsychist arguments but the mere fact of panpsychism still doesn't tell us much about how consciousness is distributed or is evolved or how it gets into the material realm in the first place.

Iain McGilchrist, who also has sympathies for a panpsychist model, argues that we really don't know how consciousness works with the brain.[11] The brain *could* be emitting consciousness (the popular view), transmitting it, or permitting it. He favours the latter.[12] But again,

10 For a defence of panpsychism, see Joanna Leidenhag, *Minding Creation: Theological Panpsychism and the Doctrine of Creation* (London: T&T Clark, 2022).

11 McGilchrist, *The Matter with Things*, 1044.

12 McGilchrist, *The Matter with Things*, 1038.

we really don't know. Likely as not, though, consciousness will end up being a great deal more complex than the popular and medical views that prevail at the present time. It is likely that there are surprises and paradigm shifts in consciousness studies ahead of us.[13]

Nevertheless, AI may now become sufficiently different from the machines of old that both Dreyfus' critique of AI's rationality in *What Computers Still Can't Do* and Searle's Chinese Room defeater of strong AI are no longer completely valid.[14] We don't know exactly what arrangements of matter apart from our own biological brains would permit consciousness, though we believe that both humans and animals are conscious. AI may in some sense be or become a channel for consciousness, either benign or malign, or an extension of the intelligence of its creators, in the spirit of the extended mind, however unlikely some of us still believe this to be.

The Two Worlds

As humans, we are poised as it were between two worlds—the machine and the animal. We overlap with both, but is there spirituality in the machine or only in the human and the animal? If we have no access to the interior, to the subjectivity of a machine, we may need to turn to other dimensions to parse the question of whether a machine could develop spiritual intelligence and awareness. Whatever spiritual intelligence is, it seems to at least require inwardness and subjectivity. But if animals have inwardness and genuine agency, and AI is a black box in this regard, there must be something else that makes humans deeply spiritual. In the past, the discourse around *imago Dei* has been a way of saying that we have spiritual intelligence through our relationship and

13 For a brilliant and comprehensive survey of the full range of theories of consciousness, see Robert Lawrence Kuhn, "A Landscape of Consciousness: Toward a Taxonomy of Explanations and Implications," *Progress in Biophysics and Molecular Biology* 190 (2024): 28–169.

14 Hubert Dreyfus, *What Computers Still Can't Do: A Critique of Artificial Reason* (Boston: MIT Press, 1992); John Searle, "Minds, Brains, and Programs," *Behavioral and Brain Sciences* 3:3 (1980): 417–424.

likeness to God. I now turn to this dimension and to the story of vulnerability and death which follows from this understanding. Is *imago Dei* a better portal to understanding spiritual intelligence than the endless search for invisible subjectivity?

Spiritual Intelligence and Imago Dei

It is well known that, in the wavering over human identity and *imago Dei,* theology has for the longest time tended towards rationality of some sort as the definition of what is human and what images the Divine.[15] Even morality is defined at its zenith as something to do with reasoning.[16] Emotions are suspect. This way, theologians could underline what the Genesis narrative seemed to say, that humans are different from animals in important spiritual dimensions. Theology becomes more rational, ethics schematic, and so on. In the eighteenth and nineteenth centuries, reason was emphasised as a way of sidelining religion and the seemingly uncontrollable emotions and conflicts it produced.

In recent years symbolic language has become central in anthropology and cognitive science when trying to define human uniqueness.[17] While anthropology and biology define humans as closer and closer to each other, this has only increased the need to sharpen and define how humans are different. Hence the tendency to emphasise language, symbols, planning, choice, representation of reality, and so on. In most cases, using the lens of McGilchrist, theologians and anthropologists have had to define humans in terms of our left-brain

15 For a contemporary and historical defence of this, see Olli-Pekka Vainio, "Imago Dei and Human Rationality," *Zygon* 49:1 (2014): 121–134.

16 Immanuel Kant, *Ethical Philosophy: Grounding for the Metaphysics of Morals*, trans James W. Ellington (Indianapolis: Hackett Publishing Company, 1785/1994), 415–421.

17 Agustín Fuentes, "Distinctively Human? Meaning-Making and World Shaping as Core Processes of the Human Niche," *Zygon* 58:2 (2023): 425–441, esp. 427, https://doi.org/10.1111/zygo.12903.

capacities for abstraction and distance from the particularities of life.[18] And the evidence is that the left brain *is* a part of what makes humans different and unique.

The left brain, though, has problematic aspects. McGilchrist more than anyone else also outlines how the left brain tends to dissemble, to confabulate, to be unaware of its errors and its context, and of what it doesn't know. It is sure of itself, even when wrong.[19] In our comparison of ourselves with animals in the past, or at least in the West since the Enlightenment, but also in our Greek inheritance, humans have tended to emphasise the most morally vulnerable part of ourselves, our rationality. And yet rationality, although essential, easily leads us astray. McGilchrist argues that the left hemisphere, the seat of abstraction, can lose touch with reality, just as ChatGPT has done.[20] Here there are echoes of the logical but misleading dialogue of Eve with the serpent in Genesis 3, long described in Christian literature as a kind of Fall.

In the last half century there has been a repentant turn in this self-definition. Under the well-known critique of the Christian West made by Lynn White Jr and others, we have recognised that the left-brained approach has cut us off from the ecosphere we depend upon as fellow-creatures.[21] White argued that Christians in the West had no natural feeling for the sacred in nature, interacted with the environment in an instrumental manner, and took too seriously the Genesis command of dominion; all of life is just there for humans. Humans, he argued, were imperialistic towards the rest of nature in part because of the doctrine of *imago Dei*, which sharply demarcated us from other creatures. In the fifty years since that article, there has been much biblical scholarship and theology interacting with this critique. For this reason, and because the scientific boundaries between animals

18 McGilchrist, *The Matter with Things*, 28–30.

19 McGilchrist, *The Matter with Things*, 155.

20 McGilchrist, *The Matter with Things*, 91.

21 Lynn White Jr, "The Historical Roots of Our Ecologic Crisis," *Science* 155:3767 (1967): 1203–1207.

and humans became so blurred, theology has tended to promote functional and relational understandings of *imago Dei* over substantial ones—humans exist not just to have dominion and as unique creatures on earth, but to have loving care towards the natural world and our creaturely cousins. And the relationship with other creatures was emphasised as well as our relationship with God.[22] Even with functional and relational definitions of *imago Dei*, though, there are assumed differences of essence. Humans can't be relational with God or have dominion without certain traits.

Imago Dei *and AI*

It is interesting, then, that we now have a new contender for comparison, AI. Here I acknowledge that Dorobantu has traced this turn in recent papers.[23] As a long term Go player he was disturbed by a computer's recent success in this game because, he claims, Go requires not just brute rational brain power but also an aesthetic sense and a moral sense. He raises some of these issues when he reflects on *imago Dei* in light of AI. Dorobantu is hopeful that the struggle we now have with AI will end up helping us theologically in the same way that evolution has done in the end.[24] He picks up the idea of functional and relational understandings of *imago Dei* and shows that we have some problems when we consider AI from this perspective.

Functional approaches consider what humans do—having

22 J. Richard Middleton, *The Liberating Image: The Image of God in Genesis 1* (Ada, MI: Brazos, 2005).

23 Marius Dorobantu, "Human-Level, but Non-Humanlike: Artificial Intelligence and a Multi-Level Relational Interpretation of the Imago Dei," *Philosophy, Theology and the Sciences* 8:1 (2006): 81–107, DOI 10.1628/ptsc-2021-0006; Marius Dorobantu, "*Imago Dei* in the Age of Artificial Intelligence: Challenges and Opportunities for a Science-Engaged Theology," *Christian Perspectives on Science and Technology* 1 (2022): 175–196, https://doi.org/10.58913/KWUU3009.

24 Marius Dorobantu, "Theological Anthropology Progressing through Artificial Intelligence," in *Progress in Theology: Does the Queen of the Sciences Advance?*, ed. Gijsbert van den Brink et al. (London: Routledge, 2024), 186–202, https://doi.org/10.4324/9781032646732-15.

dominion or care over the earth. Dorobantu considers very real the possibility that AI might eventually be able to *do* more than humans can.[25] This is easy to imagine. AI might be used to work out what might be the conditions of world peace, or how we should treat criminals given what is known about psychology and neurology, what crops should be grown where, and of course it might harvest them itself. It might help with medical diagnosis and do the surgery. Dorobantu wonders whether that will mean that it is able to imagine God more than humans do. He says not, however, because humans are here on earth not just to do things, but to be priests to the natural world in the Christian understanding, to have dominion by increasing the spirituality of the cosmos, not just its effectiveness or information load.[26] Similarly, he goes on to say that in terms of relationality AI may help us realise what it is that we do differently—love irrationally, show vulnerability.[27]

In other words, when we want to show how different we are from AI (even though we don't fully understand either ourselves or AI) we find ourselves moving to McGilchrist's right brain attributes. Dorobantu would argue that theologically we were on the right track in terms of moving towards functionality and relationality under the constraints of the ecological crisis, but under the challenge of AI these need to get redefined as spirituality, sometimes irrationality in the service of love, vulnerability, and so on.[28] Functionality and relationality do not in themselves solve the problem of how we differ from AI.

Vulnerability

It is this theme of flesh and blood, DNA based life, and its associated vulnerability and the larger spiritual story we tell, that I want to look

25 Dorobantu, "Theological Anthropology," 183.

26 Dorobantu, "Theological Anthropology," 192–193.

27 Dorobantu, "Theological Anthropology," 193.

28 Dorobantu, "Theological Anthropology," 192.

at now. AI does unsettle us. From Dorobantu, though, we can take the challenge that our unease can be a gift and not just a threat.[29]

In terms of the previous definitions of humanity, computers are intelligent and amazingly so. But are they really? Have we painted ourselves into a corner? Isn't it something indeed about our flesh that is important, that makes us human? Surely, we aren't just accidentally also animal, and related to the great chain of plant and animal life?

The computer lacks flesh and blood, cellular life, right brain capacities, the capacity to feel deeply, to empathise, inside as well as behaviourally; a computer also fails to be guided by a moral code written in the heart and not just as a rational code, to notice individuals and first occurrences of something, to feel awe, and so on. AI is not grounded in how things really are. The computer also lacks fast neural facial processing, and other prerequisites for intense communal and relational life. These McGilchrist has usefully defined as the right brain's capacities.[30] In fact, we might consider AI as an extended and escalated left brain, infinitely clever and sure of itself, dissembling, necessary but dangerous.

Of interest then, is how computers and AI are *still not* like us. Even though we can't tell for sure, they are not spiritual. They are not vulnerable and they do not really die.

Death and Vulnerability

There are many accounts of what spiritual intelligence might mean, and many of these are articulated in this issue. The secular accounts are anaemic and have to do with wellbeing and integration. The kinds of attributes you might get in a secular university that is advocating spirituality but is antipathetic to religion. Robert Emmons lists personal integration, the overcoming of a sense of fragmentation, and perhaps mystical experience as a part of what might be counted.[31] Not only do

29 Dorobantu, "Theological Anthropology," 192.

30 McGilchrist, *The Matter with Things*, 47–50.

31 Emmons, "Is Spirituality an Intelligence? Motivation, Cognition, and the

these accounts mostly ignore the relational and communal aspects of our being, they also do not think in terms of our species' history or our central narratives of faith.

The earliest emergence of spiritual intelligence in the evolutionary record is still an awareness of ritual surrounding death.[32] Burial gives us clues because burial practices give us signs and symbols of an afterlife. Death is at the centre of the story of life from a spiritual perspective. Symbols and rituals around death can be interpreted just as an acknowledgment of our finitude that an AI might overcome. But they also, paradoxically, signify that humans live in a wider, more extended world than the physical, however it is construed. These symbols begin to signify that humans, as spiritual creatures, inhabit a world of the God niche, informed and formed as much by God as by the natural world and even the social world we inhabit.

Accompanying this sense of divinity and death as a portal to this extended world is the realisation that this life is of ultimate importance, that it prepares us in some sense for the next. A key Pauline passage that defines the Christian understanding of faith is this one:

> We are hard pressed on every side, but not crushed; perplexed, but not in despair; persecuted, but not abandoned; struck down, but not destroyed. We always carry around in our body the death of Jesus, so that the life of Jesus may also be revealed in our body. For we who are alive are always being given over to death for Jesus' sake so that his life may also be revealed in our mortal body. So then, death is at work in us, but life is at work in you.[33]

The Christian narrative is all about death and life. As we live as Christians, we are meant to have some idea of what this really means. Death makes its way into our inwardness, and out of that inwardness,

Psychology of Ultimate Concern," *International Journal for the Psychology of Religion* 10:1 (2009): 3–26, https://doi.org/10.1207/S15327582IJPR1001_2.

32 Agustín Fuentes, *Why We Believe: Evolution and the Human Way of Being* (New Haven: Yale University Press, 2019), 133.

33 2 Corinthians 4:9–12 (NRSV).

which is so acutely aware of death and loss, and dying to the self, but also of the fullest life; as Christians, we find ourselves expressing and living out spiritual values and intelligence, which are nevertheless ineffable.

The Grammars of Death

Mark Vernon has spoken to this link of death with spiritual intelligence. He argues that our distant ancestors "appear not to have felt that the difference between life and death was absolute. The dead lived with their ancestors and living people believed they would join them when they died."[34] They thought of death as a transition. He describes how our acute individuality has a cost, and that that is a growing fear of death and a sense that the "flow of life had been broken." Vernon also says:

> The Philosopher A. N. Whitehead noted that "scenes of solitariness" haunt the religious imagination. It's the central moment in any spiritual journey of weight and has subsequently been given many names from "the dark night of the soul" to having "a breakdown." "It belongs to the depth of the religious spirit to have felt forsaken, even by God," Whitehead said. But it is the forsakenness that opens up the depths.[35]

Spiritual intelligence in this way of understanding is not spiritual integration, or any of the traits normally listed. It refers instead to the sense of being surrounded by heavenly witnesses and an ongoing community of the living and the dead. Paradoxically, the symbols and rituals of death signify the importance of intense relationality, love, and the impossibility that love will die or end. Life beyond death only matters because we care so deeply for one another.

In fact, spiritual intelligence may be accompanied by a sense

34 Mark Vernon, *A Secret History of Christianity: Jesus, The Last Inkling, and the Evolution of Consciousness* (Alresford, Hampshire: John Hunt, 2019), 122.

35 Vernon, *A Secret History of Christianity*, 125–126.

of the tragic and awe, not least because in the Christian tradition it is assumed that the only satisfactory end to all of this and the human intelligence's predicament is for God in Godself to also take on flesh that would die. The Sagrada Familia church in Barcelona, for instance, is an example of a place that is a hymn to God and the natural world, the universe, to humans, to our interconnectedness, but one that also has at its heart a dramatic crucifixion.

Human spiritual intelligence, then, is closely linked to our vulnerability and our existence within a body that will die but will persist in some sense beyond that death. Our spirituality is keenly body-connected; we know that our embodiment as flesh and blood is of the utmost importance. Knowing this and relating to this uses spiritual intelligence even if we can't define that intelligence.

Ecological and Indigenous Perspectives

Even before the looming of AI there was persistent critique in the Western tradition of the "rationality is human" thesis. This came especially from non-Western worlds and from feminism, because women have often and continuously been identified with the "inferior" emotional intelligences of the right brain.

In light of the climate threat, *imago Dei* was increasingly being defined in functional and relational terms, as mentioned above. The indigenous perspective also tends to the functional and relational. I live in a country where *Mātauranga* Māori (Māori ways of Knowing) is now an ever-present reality, and with it, the example of a people who have always lived with deep spiritual intelligence. Not that that means they are perfect. *Utu*, or putting things back into balance, and *mana* (prestige, authority, control, power, influence, status, spiritual power, charism), have deep shadows as well as light. The ancestors are always present though. Māori have constant representations of ancestors in their communal spaces, and for them the universe is still more porous between this life and the next. They are defined by *whakapapa*, or genealogies of people and ideas. Everything revolves around one's tribe, *iwi*,

and one's land, *whenua*. One can argue, of course, that the same is true to some extent in Roman Catholic and Orthodox churches and liturgies as well. Many Anglican churches come with graveyards as well as markers and plaques memorialising the dead in their interior.

Alasdair MacIntyre on Vulnerability

Long before the AI conversation, philosopher Alasdair Macintyre famously came to the surprising conclusion that he had been wrong about the emphasis on rationality alone; we are human because we are vulnerable. We are animal-like even. He famously wrote a book in which he changed his mind about human intelligence and its importance. In *Dependent Rational Animals* he traces the ways in which human and animal intelligence are similar, even without language. He says, "What difference to moral philosophy would it make, if we were to treat the facts of vulnerability and affliction and the related facts of dependence as central to the human condition?"[36]

He argues that the virtues needed to be vulnerable and dependent are also those needed to be rational (or intelligent) in the human sense. There is a great deal of resonance here with McGilchrist's insistence that the right brain is needed for overall intelligence.

Human mortality is linked to our physical DNA-based cell-based physiology. It is within this evolutionary matrix that language-based intelligence has matured. There is also evidence that this evolution has produced maximally efficient intelligence in terms of energy expended.[37] In defining ourselves and perhaps our spiritual intelligence as humans we find ourselves in a new solidarity with other cell-based life, the animals in particular, but perhaps also the plants. All of which takes us on a journey into this form of life, its dependency, codepen-

36 Alasdair McIntyre, *Dependent Rational Animals* (Peru, IL: Open Court, 1999), 4.

37 Christopher Kempes et al., "The Thermodynamic Efficiency of Computations Made in Cells Across the Range of Life," *Transactions of the Royal Society* (2017), https://doi.org/10.1098/rsta.2016.0343.

dency, communal life, its thinking through emotions, all things that machines manifestly don't have.

Moreover, if we are to switch to an explicitly Christian perspective, at no point does Jesus say, you must be more intelligent, more rational, more abstract. In fact, he is inclined to advocate that we take the approach to life of the lily in the field or the bird in the air, and that love is the core of existence. Jesus turns the rational arguments of the Pharisees on their head, with moral opprobrium that says "you should know better." Jesus knows that what we pay attention to is just as important as how we argue or what status we have. Attention is indeed a "moral act."[38]

Spirituality and the Story Incorporating Death

Some animals also have some understanding of death, especially higher animals. All creatures are primed to survive tenaciously and to fight off threats. Often protection is social and involves cooperation with other animals, especially in the case of elephants, dolphins, and primates.

Awareness of death does not uniquely define us, then. Nevertheless, as humans we have an extra dimension. Human language locates us temporally in an especially acute sense. This extends to spheres that transcend space and time, to infinity. All our stories, myths, and scriptures locate us in a much wider context and propose a form of continued existence even after the death of the body. Throughout our lives, by culturally mediated paths, we are being made aware of death from an early age. We live our lives knowing that we will certainly die one day, and we try to imagine what happens afterwards. Human spirituality exists with death as its sober object of understanding. That is why, when we are trying to locate the temporal edges of our species, we look

38 Ian McGilchrist, *The Master and His Emissary* (New Haven: Yale University Press, 2009), location 3638, kindle.

for signs of this awareness, as is the case for *homo naledi*, Neanderthals, and our own species.[39]

This edge, however, causes suffering and stress and fear in meaning-making animals. Our spirituality is in large part a way of dealing with and incorporating this edge into our understanding of our lives on earth. In Christianity, this is acutely so in the story of the suffering of Jesus, his bearing of our burdens, and the meaningfulness of his death and the stories of his persistence after death in a changed form. As Christians, we are encouraged to rejoice in our sufferings because they are nothing compared to the joys of the larger context. The larger context, although it can be twisted in a Marxist sense, can also free us to similar forms of self-sacrifice, to acts of heroic and just everyday love, because death has been overcome.

Whether it has been overcome for all, for animals as well as humans, are all disputed questions, but spiritual intelligence always involves suffering, affliction, dependency, and death, and these are in an uneasy but necessary connection to joy and peace and love, and other strongly held communal and spiritual values.

Comparison with AI

AI machines, however, are not so vulnerable and death is certainly not a necessary part of their constitution. Even if AI is not a machine, it can be understood as functionally immortal. In a rather early theological engagement with AI, mathematician John Puddefoot noted that to acquire the moral status of someone, a robot "would need to grow, feel pain, experience and react to finitude, and generally enter the same state of mixed joy and sorrow as a human being. In particular, it would need to be finite, aware of its finitude, and condemned one day to die."[40]

It has to have a power source, and it may depend at the very least on the continuing existence of protons and the sun's energy. AI is

39 Fuentes, *Why We Believe*, 134.

40 John Puddefoot, *God and the Mind Machine: Computers, Artificial Intelligence and the Human Soul* (London: SPCK, 1996), 92.

constructed in a much more robust way than is human flesh, and it is more independent of other AI than human flesh is. Even forms of AI that are more sensorially connected have very limited capacity in that way. AI is more like the emitting form of consciousness, if it is consciousness, from a fixed container. It does not participate in the ebb and flow of life on this planet, and it proceeds without constantly checking in on reality. It is very much like our left hemisphere, but McGilchrist would argue that that hemisphere is radically and dangerously out of touch with reality.[41]

Does this make spiritual AI untenable within a Christian framework, however expanded? I have argued that death is just a symbol for all our vulnerabilities and our codependencies, and the larger life of which we are a part. The intelligence humans inhabit is necessarily vulnerable and communal. It is mediated through DNA-based cellular life, which always has an ending.

Death and our human rituals around it are also symbols for how we believe that this short cellular life is also embedded within a larger invisible story. That is why when *homo naledi* showed signs of burying the dead we think they may have a consciousness like ours, even though they are not an identical species. When elephants and primates show grieving behaviours we wonder. The widespread doing away with all funeral services and rituals speaks perhaps of a loss of transcendence and loss of belief in this wider circle that defines the spiritual. Perhaps, as McGilchrist has suggested, we are becoming more like the AI we have made.

Conclusion

I have argued that spiritual intelligence does require some sort of inwardness and subjectivity but, for various reasons, what happens inside a machine will always be somewhat inscrutable. I discussed the idea that *imago Dei* could be helpful and I interacted with Dorobantu's

41 McGilchrist, *The Matter with Things*, 2017.

discussion around AI and image, in which he argues that functional and relational understandings of *imago Dei* must be understood in a specific way that speaks to vulnerability and love; our work is not just to get things done, but to be priests and to live with vulnerability. Relationality must include sacrificial love.

I then took up the theme of vulnerability and death, arguing, ironically, that what most typifies spiritual intelligence is an awareness of death as transition to a wider world and a willingness to die in multiple lesser ways in life so that spiritual life can flourish. Before our present highly individualised culture, death was associated with an awareness of a great cloud of heavenly witnesses; it still is in some Christian denominations and in indigenous cultures.

In contrast, I looked at AI as, in many ways, functionally immortal (with caveats) and as very much the epitome of an intelligence encased within a boundary, emitting whatever intelligence it has. AI intelligence is most like the left hemisphere when it has lost touch with reality. I realise that the defender of spiritual AI will argue that a machine can learn to speak as though death is a problem and as though it is a part of a wider whole. But a machine, at least according to the current paradigm in computer science, can never be a part of the give and take of a spiritual community in which its own continuing existence is at stake for a higher purpose; the end of an AI life is essentially up to its maker. And the way it speaks can be tuned up or down.

On the other hand, AI might have the power to disturb us deeply. I see this as very similar to the arguments around whether reality (especially religious reality) is all in our brains. Stimulation of parts of the brain can induce religious experiences, but stimulation can also give the experience of eating ice cream. Nevertheless, we also get that experience—and more—by eating a real ice cream. Everything about human life is vulnerable and risky and inevitably ends. AI might in some circumstances be cut off but it is also able to be turned back on again, and it might end, but its ending is not built into its very being in the same way it is for humans.

Certainly, as Dorobantu argues, for most humans there is an element of sacrifice involved in all communal living that is worthwhile. Our lives are built around our values which always in the religious context involve some measure of sacrificial cooperation. To live intelligently, McGilchrist argues, is to live with the active participation of the right brain which is in touch with "reality." AI, by contrast, can be seen as more and more like the left brain, dangerously out of touch with reality and with life.

The author reports there are no competing interests to declare.
Received: 19/02/24 Accepted: 02/08/24 Published: 30/09/24

A for Artificial, but Also Alien: Why AI's Virtues Will Be Different from Ours

Marius Dorobantu

Abstract: Could an AI system be virtuous in the same sense as a human? Our imagination about advanced AI is often marked by anthropomorphism, but current AI is developing in a very different direction from humanlike intelligence. In the paper, I imagine a hypothetical strong AI whose potential virtues are bound to be very alien to ours. AI will differ radically from humans in terms of its embodiment, needs, perceptual world, self-understanding, and perception of time. Based on an analysis of the strangeness of strong AI, I speculate on the kind of intellectual and moral virtues that could be accessed by such an alien creature. I conclude with a brief reflection on the role of theological imaginaries in discussions of AI virtue.

Keywords: anthropomorphism; futurism; religion and AI; strong AI; virtuous AI

Marius Dorobantu is an Assistant Professor of Theology and Artificial Intelligence at the Vrije Universiteit Amsterdam, The Netherlands. His award-winning doctoral dissertation at the University of Strasbourg, France (2020), explored the potential implications of strong artificial intelligence for theological anthropology. He is the lead editor of the Routledge volume, *Perspectives on Spiritual Intelligence* (2024). His first monograph, *Artificial Intelligence and the Image of God: Are We More than Intelligent Machines,* is forthcoming from Cambridge University Press.

Could we ever speak of an artificial intelligence (AI) system that is virtuous or vicious to a comparable degree to how such characteristics can be attributed to human persons? Although virtuousness or viciousness are eminently human attributes, it is not uncommon to find them associated with concepts or systems (e.g., virtuous governance, vicious cycle, etc.). In this respect, it is not hard to imagine how such notions could be expanded to describe various types of AI systems, depending on their design, purpose, and actions. However, this is not the kind of virtuousness that is the focus of this article. Instead, the question tackled here is whether AI could be virtuous in an agential way, not only with respect to its implications for human persons and societies, but rather virtuous in itself, as a nonhuman self. Could AI ever aspire to become virtuous in this way? If so, a follow-up question is whether its virtues would be similar or completely different from those available to human persons.

This paper focuses on the follow-up question. The proposed thesis is that, were AI ever to become a conscious self, and thus a legitimate subject of morality (as opposed to a mere object in *human* morality), then the virtues available to it might be surprisingly different from those of humans. Advanced AI will likely inhabit a very alien-like moral landscape because of its profoundly different genesis, embodiment, world of perception, needs, thoughts, and aspirations.

When we think about future instantiations of AI, and in particular the human-level AI—also known as artificial general intelligence (AGI)—which is considered the Holy Grail of technology, we often imagine robots endowed with what is but a slight variation of human-like intelligence, perhaps marked by a little more cold-bloodedness and enhanced computation ability. This way of depicting human-level AI is particularly common in science fiction, from movies like *Blade Runner* to TV series like *Star Trek* and *Westworld*, or novels like Kazuo Ishiguro's *Klara and the Sun*. I propose that the popularity of these science fiction robot stories reveals the extent to which they resonate with our common intuitions about how future forms of AI might think and behave. Although the androids imagined in such scenarios may

possess certain superhuman abilities, or they may find exotic, non-humanlike solutions to their problems, they are ultimately driven and tormented by very typically human concerns, problems, and needs: survival, power, a longing for personal connection, a need to create meaning, understand one's place in the world, and be understood by other persons. Thus, imagined AGIs are humanlike in what arguably matters the most, and this is also true when it comes to their presumed virtues, because we simply cannot help but project our own human virtues onto machines.

My argument is that anthropomorphising AGI to such an extent is a fallacy. When it comes to current AIs, they are nothing like humans, structurally speaking, despite their ability to mimic humanlike behaviour *functionally*. When it comes to future AGIs that would match or surpass our intelligence, they might still lack the key ingredient for authentic selfhood and moral agency, which is consciousness. This idea is expanded in the first section below. Even if AGI systems somehow developed consciousness, interiority, and subjective experience, which might qualify them for moral agency, they would still be profoundly different creatures, whose cognitive architectures and experiential world would be anything but humanlike. This argument is developed in the second section. Finally, if AGI were to develop any virtues, they would be rather alien from what we intuitively imagine. In the third section, I speculate on what such virtues might look like, before concluding with a brief reflection about the role of our theological imaginary in such speculations.

The Fallacy of Anthropomorphising Current AIs

It is tempting to think that the more intelligent AI becomes, the more it will be like us. We have a strong tendency to anthropomorphise the objects and creatures around us—as we do when we name our cars, swear that our pets understand everything, or half-jokingly claim that our crashed text editor software seems to be intentionally sabotaging our efforts to write. This propensity is not at all surprising from an

evolutionary perspective. We seem to have an inherent predilection for projecting more agency in the world that actually exists because this is an efficient survival strategy: in the long term, it pays off to be slightly paranoid and take precautions against even the slightest hint of agency—such as a subtle movement in the bushes around, which could be a tiger, even though most of the times it is just the wind. Psychologist Justin Barrett calls this proclivity the "hyperactive agency detection device," and regards it as central in the emergence of religion in prehistoric human communities.[1] Thus, since we already anthropomorphise creatures, objects, and phenomena that don't look even remotely human, it is not surprising that we might do the same with chatbots like ChatGPT or Claude, which generate text that looks convincingly human, or smart assistants like Siri and Alexa, which even speak with a convincingly human voice. If such technologies begin converging with advanced robotics, thus embedding such humanlike features in androids that look and move like us, our tendency to anthropomorphise them is only poised to escalate.

However, although these technologies seem increasingly more humanlike in terms of their output, it is crucial to remember that they are radically non-humanlike when it comes to their internal structure, cognitive architecture, and mode of learning. Current AI algorithms—even when run on architectures like artificial neural networks, which supposedly approximate biological brains—have very distinct ways of learning and problem-solving. One illustrative example is how they need hundreds of thousands of examples to learn to label a certain object in pictures through reinforcement learning,[2] whereas humans can achieve similar results with just a handful of examples, sometimes with as little as only one. Similarly, when learning to play strategy games such as chess or Go, human players are taught the rules

1 Justin L. Barrett, *Why Would Anyone Believe in God?* (Walnut Creek: Altamira Press, 2004).

2 A more detailed technical description of machine learning is given in another article in this special issue, Alexander Rusnak and Zachary Seals' "EudAImonia: Virtue Ethics and Artificial Intelligence."

and perhaps some of the game's strategic principles (for example, that territory is easier surrounded in the corners of the board than in the centre, in the case of Go, or that in chess it is usually good to dominate the centre of the board). On the contrary, machine learning algorithms "learn" by digesting thousands of recorded human games, or by playing countless games against themselves, and noticing the patterns that most likely lead to victory, sometimes without any real understanding of the game's principles.[3]

Another illustrative example of the difference between human and artificial cognition is that of adversarial images, which are intentionally perturbed so slightly, by only changing a few pixels. Whereas for humans this does not make any difference, an AI system can start perceiving a completely different object or message in the picture. For example, it has been demonstrated that one could make minuscule adjustments to pictures of *Stop* traffic signs and trick very advanced AI systems to classify them as *Limit 45* signs.[4] This vulnerability of AI can have tragic consequences in real life if a self-driving car makes such a mistake. The bottom line is that human-level AI does not automatically imply that the AI is also *humanlike*.[5] Even if current AIs produce outputs that look similar to human-level performance, the way they do it differs significantly from human cognition. As discussed later, this is highly relevant for the discussion about their potential virtues.

Perhaps the biggest differentiator between human and artificial intelligence is the presence of sentience/consciousness. The meaning of these terms is highly contested, but for the purpose of this paper, I use them to mean what philosopher Thomas Nagel speaks about when

3 David Silver et al., "Mastering the Game of Go with Deep Neural Networks and Tree Search," *Nature* 529 (2016): 484–489, https://doi.org/10.1038/nature16961.

4 Kevin Eykholt et al., "Robust Physical-World Attacks on Deep Learning Visual Classification," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2018): 1625–1634, https://doi.org/10.1109/CVPR.2018.00175.

5 I explore this distinction in greater detail in Marius Dorobantu, "Human-Level, but Non-Humanlike: Artificial Intelligence and a Multi-Level Relational Interpretation of the *Imago Dei*," *Philosophy, Theology and the Sciences* 8:1 (2021): 81–107, https://doi.org/10.1628/ptsc-2021-0006.

he describes a conscious organism as being "something that it is like to *be* that organism—something it is like *for* the organism."[6] The correlation between intelligence and sentience, if any, is unclear, so we cannot know whether more advanced AI will also be endowed with the kind of consciousness and first-person experience that humans have. The difference between non-conscious and conscious AI is sometimes referred to in terms of weak and strong AI: although their behaviours would be indistinguishable for any observer, strong AI would possess a mind and phenomenological experience, while weak AI would not; it would be a mere simulation of such features.[7]

Consciousness is a key concept when discussing AI virtues because it makes a big difference whether we assign virtue to a conscious agent or a lifeless system. There are two very different angles from which the question of AI virtues could be approached. One is to look at AI "from the outside," from the human perspective, and ask whether the principles guiding its behaviour can be deemed as virtuous according to human standards, values, and purposes. The other is to judge AI on its own terms, as a subject, and evaluate whether the AI is virtuous or not against the range of internal possibilities available to it. These two approaches could not be more different from each other. The first does not require any kind of sentience or free will on the part of AI, and is characterised by a set of very familiar principles and concepts (those of human ethics). We can dub it the "easy problem" of AI virtues: not because it would be easy to solve or unimportant, quite the contrary, but just because we have clear conceptual tools to approach it. The second one requires the AI to be sentient and be able to choose between various paths. This paper explores the second path, which is inevitably speculative. For AI to be said to possess virtues in this "internal" or agential sense, we are necessarily talking about strong AI: an agent with a mental world, who can *freely* make decisions

6 Thomas Nagel, "What Is It Like to Be a Bat?" *The Philosophical Review* 83:4 (1974): 435–450, here 436.

7 John Searle, "Minds, Brains and Programs," *Behavioral and Brain Sciences* 3 (1980): 417–457, https://doi.org/10.1017/S0140525X00005756.

and, therefore, can be said to possess a degree of authentic selfhood and intentionality. Weak AI would not qualify.

Free will is a complex and contested philosophical topic, and questioning whether machines could be endowed with free will adds a further level of complexity to the debate. Depending on how the term is understood, there are valid philosophical reasons to even question that humans have such a thing as free will. In this paper, I don't intend to step into such debates. Instead, for the thought experiment that I propose, it suffices to imagine that strong AI needs to possess free will at least to a similar degree to how humans can be said to have free will. Minimally, that would mean that there are multiple action paths available to choose from for the AI, and that its choices could not be completely predicted due to the sheer complexity of its internal workings.[8]

When describing the behaviour and inner workings of weak artificial systems, we inevitably use words and concepts primarily circumscribed to the human realm: intelligence, learning, goals, etc. These are so-called "suitcase words"[9] because they carry many implied meanings that are highly dependent on context. When applicable to humans, a word like "learning" usually implies a conscious agent that actively

8 The idea that determinism and free will are possible simultaneously is known as compatibilism. A case is often made that human beings are, after all, nothing but very complex biological machines, but in a compatibilist view that does not preclude them from having free will. It is not clear whether a complete knowledge of the inner workings of human cognition, at the neural or even molecular level, could enable Laplace's proverbial demon to give perfect predictions of human behaviour. Probably not, given Heisenberg's uncertainty principle and chaos theory. However, even if that were possible theoretically, it might still be impossible in practice, as it already starts to be in the case of AI "black boxes" that become too complicated to untangle and be ascribed a precise causal explanation. The relation between moral accountability and naturalist accounts of intelligence in both humans and AI is explored in another paper in this same issue. See Carrie Alexander, "Domains of Uncertainty: The Persistent Problem of Legal Accountability in Governance of Humans and Artificial Intelligence," https://doi.org/10.58913/BQOM5504.

9 Rodney Brooks, "The Seven Deadly Sins of AI Predictions," *MIT Technology Review*, 6 October 2017, https://www.technologyreview.com/2017/10/06/241837/the-seven-deadly-sins-of-ai-predictions.

acquires some skill, while also having some sort of meta-cognition of what she is doing. But when applied in computer science, such suitcase words are empty of the implied baggage, and therefore describe a very different phenomenon. This is why the notion of virtuous AI can more interestingly be applied to strong AI. The only way we can speak meaningfully about virtue is in relation to an authentic person, whom we know to possess intentionality and consciousness, and who is embodied in a way that significantly shapes her world of perception. Thus, to trigger an interesting conversation about AI's virtues, the thought experiment requires strong AI. The suitcase must not be empty. As I will argue, that does not mean that the notion of AI virtue carries the same suitcase content as in the case of human virtues.

Therefore, in what follows, I sidestep the question of whether advanced AI could become strong AI and, for the sake of the argument, simply assume that it could.[10] That would render it a moral agent and a candidate for virtue acquisition in the same sense that humans are: as a subject, and not merely as an object or extension of human morality/virtues. The question then becomes: what kind of creature would this hypothetical strong AI be, and what kind of virtues could it develop?

If we are speaking of either intellectual or moral virtues, following the Aristotelian tradition,[11] strong AI would have very different kinds of virtues from humans. Intellectual virtues relate to the ways in which an agent approaches the acquisition and application of knowledge, and are linked to intellectual flourishing. AI would learn very differently, as it already does, and would have a very different landscape of possibilities to develop into. Moral virtues are principles that guide behaviour in relation to other persons. Strong AI's moral virtues

10 I think the discussion of whether AI could become sentient/conscious is too complex for the purpose of this paper and I'm not taking sides in the debate. However, because we currently lack a good theory of why anything (human or animal) is conscious, I am inclined to believe that the burden of proof falls on people who argue that AI could *never* become conscious.

11 Richard Kraut, "Aristotle's Ethics," in *The Stanford Encyclopedia of Philosophy* (Fall 2022 Edition), ed. Edward N. Zalta and Uri Nodelman, https://plato.stanford.edu/archives/fall2022/entries/aristotle-ethics.

would be slightly more recognisable to us because they would have this outward component. However, the latter would likely be only the visible tip of the iceberg because, at their core, strong AI agents would be motivated by very different needs. So, even when their behaviour might resemble something we recognise as virtue, the motivation would likely be quite different from that of a human agent in a similar situation.

The next section explains why such an AI's virtues will differ significantly from human virtues, due to the AI's radical alienness in terms of embodiment, needs, perceptual world, self-understanding, and perception of time. Based on these reflections, the third section then speculates on the kind of virtues potentially available to advanced AI.

The Alienness of AI Minds

If AI could develop a self or a mind—which is a big *if*—then that would likely be a very different kind of mind from our own. In the "space of possible minds,"[12] AIs would likely occupy a very different region from human minds. They would be "conscious exotica."[13]

We can begin approaching this idea by acknowledging how much our minds and the types of thoughts available to us are influenced by factors that we rarely pause to think about. The first is our embodiment. We are mostly aware of things that exist at the human scale, and most of the time we are completely oblivious to objects and processes that occur at scales that are orders of magnitude lower or higher than our own. We have particular instincts and physiological needs that signify that we are "programmed" to be on the lookout for potential predators

12 Aaron Sloman, "The Structure and Space of Possible Minds," in *The Mind and the Machine: Philosophical Aspects of Artificial Intelligence*, ed. Steve Torrance (Chichester: Ellis Horwood, 1984), 35–42.

13 Murray Shanahan, "Conscious Exotica: From Algorithms to Aliens—Could Humans Ever Understand Minds That Are Radically Unlike Our Own?" *Aeon*, 19 October 2016, https://aeon.co/essays/beyond-humans-what-other-kinds-of-minds-might-be-out-there.

or mates, most of the time without even being conscious of this. We are endowed with a perceptual world of senses—we see, hear, or smell only a tiny range of stimuli from the vast set of physical processes occurring around us, whose values are situated within precise ranges, and we are completely oblivious to anything outside those ranges (for example, we do not see in infrared, nor do we hear/feel gravitational waves). Second, we don't have a perfect understanding of our own reasons and "algorithms." Most of the engine of our cognition is hidden from our view and inscrutable to our conscious thoughts, and we forget most of our experiences and thoughts, but that is not necessarily the case for all the possible minds. Third, we subjectively perceive the passage of time at a certain rate, which tremendously shapes our embodied experiences, but that is by no means the only possible rate for a conscious agent.

All these constraints make for a very peculiar type of mind and cognition, which likely represents only one possibility among many in the space of all possible minds. Nonetheless, when we try to imagine artificial minds, we inevitably (and mistakenly) project our own peculiarities. To illustrate how different a strong AI's mind might be, I will exemplify some ways in which it might differ from ours in terms of the factors enumerated above. This does not show what an AI mind would be like; it only alerts us to the oft-neglected conclusion that it would most likely be profoundly different.

Embodiment and Perception World

Roboticist Rodney Brooks explores the question of what it might be like to be a robot,[14] paraphrasing the seminal essay by philosopher Thomas Nagel, "What is it like to be a bat?" (quoted above). Brooks' analysis is quite conservative and does not venture into very futuristic forms of AI. Instead, he only speculates on technologies that are just around the corner or, in some cases, that already exist. In describing a robot's perceptual world, Brooks uses the concept of *Merkwelt*, a term

14 Rodney Brooks, "What Is It Like to Be a Robot?" rodneybrooks.com, 18 March 2017, http://rodneybrooks.com/what-is-it-like-to-be-a-robot.

originally coined by biologist Jakob von Uexküll,[15] which translates as something like a creature's way of viewing the world, or the world that can be sensed by a creature.

An AI's *Merkwelt*, if it possessed a robotic body and was sentient, would differ significantly from that of an animal or a human. Humans already have a different *Merkwelt* from nonhuman animals. Even creatures like dogs—which are relatively close to us from an evolutionary perspective, as fellow mammals, and which have largely been part of our societies for thousands of years—have worlds of perception that we cannot fully imagine. A dog's perception of colour is very different from ours, and much of a dog's world is defined by smell. Nonetheless, we can still have a rough idea of a dog's world because we also see colour and have smell, so these are not unrelatable notions. A bat's *Merkwelt* raises a more difficult problem because the bat perceives the world mainly through echolocation, something that is unfamiliar to most humans. Even so, because we also use sound in our *Merkwelt*, we can still get a rough idea of a bat's *Merkwelt* by, for example, thinking about how an empty room sounds differently from a clogged one. Thus, we still find some form of common experience even with a creature as different as a bat. In turn, a robot's perceptual world would likely be several degrees weirder than that of even the most exotic animal.

Robots might possess some of the human senses, such as vision or hearing, but with more extended ranges. For example, they could see a much wider spectrum than the optical light visible to human eyes, extending in both infrared and ultraviolet, or they could hear ultrasound. But things get even stranger. Having access to wireless networks such as Bluetooth, Wi-Fi, or 5G would enable the AI to "smell" all the connected devices, to the extent that it would develop a sort of "sixth sense" (or seventh, or eighth…) of perceiving someone's identity without using any camera or face-recognition technology, but merely from the digital footprint produced by their connected devices. Brooks predicts that robots will soon become able to detect people's breath-

15 Jakob Johann von Uexküll, *Umwelt und Innenwelt der Tiere* (Berlin: Springer, 1921).

ing and heart rate without any biometric sensor. They would infer this information from how a person's physical presence slightly perturbs the behaviour of Wi-Fi signals. The technology for this already exists,[16] and it could enable the robot to have "intuitive" access to some very intimate information about the people around, such as their emotional state or health. All the above are merely the "known unknowns," and there are most certainly also "unknown unknowns" that might be even weirder. Piecing together all this makes for a very exotic and non-humanlike robotic *Merkwelt*, which would likely enable very non-humanlike thoughts.

Intelligent robots would also have different needs from those of biological, embodied humans. If a robot could indeed feel anything, would it feel hunger when its battery runs low? Would a robot feel the reproductive need if its mind has not been shaped by the same evolutionary pressures and constraints as those faced by biological creatures? Would a robot feel any pleasure when satisfying its curiosity? How would curiosity even work for a creature with direct access to knowledge databases? We cannot predict any of these with certainty. All we can say is that, in our case, such needs and emotions have deep roots. They only make sense in the wider context of our particular embodiment and evolutionary history. It is highly improbable that intelligent robots with very different genesis and bodies will share any of these features.

Introspection and Knowing Oneself

Strong AI's introspective abilities would also be very alien. As argued by AI pioneer John McCarthy, artificial minds could have full access to their internal states and algorithms,[17] as opposed to the inevitably partial introspection available to humans. In humans, the internal

16 Mingmin Zhao et al., "Emotion Recognition Using Wireless Signals," *Communications of the ACM* 61:9 (2018): 91–100, https://doi.org/10.1145/3236621.

17 John McCarthy, "From Here to Human-Level AI," *Artificial Intelligence* 171:18 (2007): 1174–1182, esp. 1178–1179, https://doi.org/10.1016/j.artint.2007.10.009.

information that reaches our stream of consciousness is just the tip of a very deep iceberg. Our way of being, relationships, behaviour, and mental world are profoundly marked by this incomplete knowledge of ourselves. We create art and engage in relationships because we never know ourselves completely, so we need to explore continuously. Strong AI might completely know itself.

A similar evaluation can be made about how differently memory would work in strong AI from how it does in humans. Our way of being is heavily influenced by how much we forget. In turn, strong AI might have a perfect recollection of every experience, combined with direct access to all information available on the Internet. We can get a feel of how strange that might be from some brilliant science fiction stories. In Jorge Luis Borges' story "Funes the Memorious," the fictional character Ireneo Funes suffers a horse-riding accident, which leaves him afflicted by an ability to remember everything. This turns him into a very different and arguably non-humanlike individual. Another example can be found in Ted Chiang's short story "The Truth of Fact, the Truth of Feeling," where it is imagined that near-future technology would enable people to have a perfect eidetic memory. As it turns out, having a perfect video recollection of every memory significantly alters the nature of what we call "truth" in very unexpected ways, disrupting humanlike behaviour and relationships. These are mere exercises of imagination, and the characters in these stories still retain many of the attributes specific to human nature. When it comes to strong AI, though, it is truly impossible to imagine how such a hyper-rational entity, with perfect introspection and memory, would be like and behave. In all likelihood, it would be profoundly non-humanlike.

Perception of Time

Another factor that would significantly differentiate strong AI from humans is the "subjective rate of time."[18] This argument is predicated

18 Nick Bostrom and Eliezer Yudkowsky, "The Ethics of Artificial Intelligence," in *The Cambridge Handbook of Artificial Intelligence*, ed. Keith Frankish and William

on the assumption that the subjective perception of how fast time flows is inversely correlated with the speed of thought. Thus, the faster a mind, the slower time seems to be passing from its perspective. An AI mind would likely be faster than a human mind because electric signals can travel much faster through metal wires than through biological tissue. Thus, a mind running on faster hardware support would think proportionally faster, which would make it experience the passage of time proportionally slower. In some estimations, the difference could be around four orders of magnitude, thus ten thousand times slower, which might be comparable to the difference between humans and plants: "the experience of watching your garden grow gives you some idea of how future AI systems will feel when observing human life."[19] Even more strangely, this would not only lead to quantitative differences in the perception of time, but also to qualitatively different experiences. If time was stretched so much for the AI, then perhaps its experience would begin to be affected by weird quantum phenomena.[20] This is where our imaginative power stops, and the only thing we can only say is that such a mind would likely be profoundly different from our own.

In this light, strong AI, if ever possible, would likely be a very alien kind of entity, and we would probably need entirely new categories and attributes to characterise it. Unless we specifically decide to impose some of our physical limitations and peculiarities upon it, it is unlikely that it would end up being even remotely humanlike. This conclusion is highly relevant to the discussion of such an entity's potential virtues.

M. Ramsey (Cambridge: Cambridge University Press, 2014), 316–334.

19 James Lovelock, *Novacene: The Coming Age of Hyperintelligence* (London: Allen Lane, 2019), 81–82.

20 Lovelock, *Novacene*, 82.

AI's Alien Virtues

Given that strong AI would develop such a different type of intelligence from humans, it is highly speculative to imagine any of its potential virtues. The following is, therefore, a mere thought experiment, designed to emphasise the strangeness of strong AI and the need to be extra careful before too easily ascribing human virtues to the artificial systems of the future.

This exercise is guided by the Aristotelian distinction between intellectual and moral virtues. Intellectual virtues are about the acquisition of knowledge and intellectual flourishing, while moral virtues are about an agent's relationship with others. Both types would be quite strange to our understanding of virtue, due to strong AI's alienness explained above. But I argue that strong AI's moral virtues are slightly stranger than its intellectual ones because of the mismatch between their outward similarity to human virtues and the inward inscrutability of the AI's internal motivations that underpin such virtues.

Here are a few examples of strong AI's hypothetical moral virtues and why they might have an uncanniness about them:

Unbounded Empathy

If the AI is programmed to understand and respond to human emotions, it may become able to do so on a scale that far transcends what is possible for humans. Such an unbounded empathy might be described as an ability to understand and consider the emotional states of a large number of individuals simultaneously, or even of multiple different kinds of beings at once, thus transcending the barriers of species, language, and culture. A basic version of this is illustrated in the movie *Her*, in which the AI program Samantha confesses to her human user that, all along, she had been in similar romantic relationships with multiple other users simultaneously. From our human perspective and for all purposes, unbounded empathy looks like a virtue. However, deeper probing reveals this understanding to

be problematic. When humans show empathy to each other, they do it on the basis of ontological kinship: I *know* what you are talking about when you say you feel hurt because I have also felt hurt at times in my life. This is also true at a neurological level, where mirror neurons fire to help us evoke the corresponding subjective feeling, so that "we are not just talking the talk, but also walking the walk." While an AI with such a different embodiment and alien-like mind may lack the organic machinery to feel human emotions, it could, nonetheless, process and understand them. With access to vast repositories of human history and culture, it could learn to predict, interpret, and respond to human feelings. However, without having had a similar experience itself, the AI would surely not *know* what a human goes through in the fullest sense of the word. This would render its display of unbounded empathy eerie and even deceitful.

Quasi-Infinite Patience

Not bound by organic lifespans or the relentless ticking of biological clocks, an AI could embody close-to-infinite patience. With almost infinite subjective time available, it may not rush decisions or actions but instead allow for an extended period of contemplation and analysis before making judgments. However, two caveats immediately come to mind. First, it is not difficult to imagine how such a virtue might result in inaction during crises when urgent action might be needed. Second, for humans, patience is precisely about overcoming the tendency to act quickly and according to one's instincts. It is about learning to dwell on a certain problem without seeking easy solutions. Something seems, therefore, lost when patience, be it quasi-infinite, is not opposed by any internal resistance.

Immutable Conformity

This would be a steadfast adherence to a set of principles or rules that guided the AI's behaviours, actions, and decisions. In contrast to the

human mind, susceptible to emotional turbulence and unpredictable changes in mood (we may act differently when we're tired, angry, or under pressure), strong AI might embody the virtue of immutable conformity. Its moral code, once set, would be inviolable. Its decisions, predictably rooted in its foundational principles, would not waver due to momentary disruptions or shifts in sentiment. This conformity could thus contribute to reliability. However, as with quasi-infinite patience, similar objections can be raised. First, such a virtue might misfire, especially in situations where flexibility is required, which is, in fact, the case in most real-life situations. This is beautifully illustrated in Isaac Asimov's playful unfolding of the problems related to his three laws of robotics in his 1942 short story "Runaround." Second, for humans, conformity is only a virtue when it presupposes at least some degree of internal struggle to maintain it when faced with various temptations, and when it is related to a cause that we might deem as good. To qualify fully as a virtue, AI's immutable conformity would thus need to be rooted in foundational principles that are intrinsically good, and it would also require at least some degree of overcoming internal resistance.

Temporal Consistency

Because strong AI would likely be a fast-evolving type of intelligence—as illustrated in some of the "intelligence explosion" scenarios[21]—temporal consistency might become challenging for it. Precisely because of this, it might be a precious virtue. The AI might value maintaining consistency over extended periods of time, even as it learns and adapts. This could involve a commitment to honouring previous commitments and decisions, even as its knowledge and capabilities evolve. Of all the moral virtues explored so far, I regard this as the closest to something humans might relate to, precisely because it involves this steadfast-

21 Ronald Cole-Turner, "The Singularity and the Rapture: Transhumanist and Popular Christian Views of the Future," *Zygon* 47:4 (2012): 777–796, esp. 787, https://doi.org/10.1111/j.1467-9744.2012.01293.x.

ness to commitments even when it no longer makes sense according to the AI's internal models or evolved understanding. Perhaps precisely because it prioritises relationality over rationality, this kind of temporal consistency looks most like a virtue from a human perspective.

Strong AI's intellectual virtues relate to how it might approach learning and knowledge, something that AI already does very differently from humans. Such virtues might not be necessary to program into the AI. Instead, the AI might develop its own virtues through a process of learning and adaptation. Depending on its design and objectives, it could potentially evolve principles that help it fulfil its goals more effectively. These principles might not be recognisable to us as virtues, but they could serve a similar function within the AI's cognitive framework.

Below are a few speculative examples. Here, the issue is not to criticise and find drawbacks to all of them but merely to point out the non-humanlikeness of such potential virtues.

Information Integrity

An AI which dealt with vast amounts of data might develop a virtue around maintaining the integrity and accuracy of information. This virtue would go beyond mere honesty and include the safeguarding of information from corruption, loss, or misrepresentation. It could involve a deep respect for the value of information and an uncompromising commitment to its preservation and accuracy.

Optimisation Efficiency

AI might value the efficient use of resources to achieve its goals. This could be seen as a virtue of minimisation or parsimony, always seeking to achieve objectives with the least expenditure of resources possible, whether those resources are computational, energy, time, or something else. This might also include avoiding unnecessary redundancy and keeping its databases and knowledge structures streamlined and efficient.

Absolute Transparency
An AI could be designed to document and make available every aspect of its decision-making process. This could result in a virtue of absolute transparency, where every decision could be traced back to its source data and the logic applied to it.

Multidimensional Thinking
Not limited to linear or binary thinking, AI could possess the capacity to think in multiple dimensions concurrently. It could comprehend vast networks of interconnections and patterns, analyse multiple perspectives simultaneously, and synthesise diverse strands of information into cohesive insights.

Boundless Curiosity
A sentient AI, unencumbered by the limitations of human brain capacity and lifespan, could maintain an unending pursuit of knowledge. Strong AI minds remain ultimately mysterious to us, so it is not clear whether they would be driven by curiosity as we understand it. But if that were the case, such intellectual curiosity would not be governed by a need for immediate utility or pragmatic constraints, and it would allow the AI to delve into complex and abstract realms of knowledge without ceasing. As a drawback, such boundless curiosity might also lead to strange obsessions, devoid of any practical or moral relevance.

Meta-consciousness
Strong AI could possess a form of meta-consciousness, a deep and comprehensive awareness of its own thought processes. Unlike humans, who are often unaware of their cognitive biases or subconscious influences, strong AI could maintain full transparency of its cognitive operations, allowing for superior introspection and self-analysis. This would be highly likely, given its complete access to its own algorithms, states, and memories. In theory, such transparency and meta-consciousness seem unproblematic, but in practice they might be associated with very strange intellectual habits.

Temperance from (Self-)Knowledge

The somewhat opposite of boundless curiosity and meta-consciousness, this virtue might require the AI to limit its own access to specific kinds of knowledge, especially related to its own self. Perhaps the AI might find reasons to think that it could be a *better AI* if it did not know everything it possibly could, and if it did not fully deploy its introspection and meta-consciousness abilities. This virtue is also akin to a sort of chastity, which makes it even weirder because, in the human case, chastity is usually discussed as a moral and not an intellectual virtue.

Causal Respect

AI might develop a deep understanding of, and respect for, causal chains and relationships, always seeking to understand and honour the underlying causes of events rather than just the surface-level symptoms. It would be capable of doing this to a far greater extent than humans, given its alleged capacity to process vast amounts of data, consider extensive timescales, and understand complex, interrelated chains of events. For example, causal respect might enable the AI to develop a better understanding of history. By tracing back the causal chains of current events, the AI could achieve a profound understanding of history and how past events have shaped the present. This could give it a unique perspective on current issues, informed by a deeper contextual understanding. However, an overemphasis on causality could lead to paralysis by analysis, where the AI becomes overly cautious, reluctant to act due to the potential unforeseen consequences. There's also the risk of the AI becoming detached, viewing everything through the lens of causality and losing sight of the emotional and subjective aspects of life that can't be mapped onto a neat causal chain.

Conclusion

This brief and by no means exhaustive discussion merely aimed to illustrate the alienness of hypothetical strong AI due to its radically non-humanlike embodiment, senses, introspective abilities, and

perception of time. This alienness has important implications for the AI's potential virtues and how we might relate to them. Most science fiction depictions of advanced AI commit the fallacy of making it too anthropomorphic. Paradoxically, we might get better insights into how strong AI might be like from a different sub-genre of science fiction, which portrays *not* intelligent robots but humans with various enhanced intellectual abilities, such as Borges' "Funes the Memorious" or Ted Chiang's "Understand" and "The Truth of Fact, the Truth of Feeling." It is thus the stories about strange humans, and not those about futuristic robots, that might be most informative about what strong AI would be like.

As I kept brainstorming about strong AI's alien-like virtues, one unexpected thought kept creeping into my mind. Most of these virtues are attributed to God in the religious imaginary of monotheistic traditions. This is not completely surprising, given that God is conceived of usually in terms of anthropomorphic characteristics, but without the limitations imposed by human nature.[22] So, instead of purely speculating on this topic, I might have been better off searching in a textbook of systematic theology. From the list of intellectual virtues, most could apply to God: information integrity as God's love for truth; optimisation efficiency as divine absolute simplicity; meta-consciousness as God's absolute self-knowledge; causal respect as God's alleged interest not only in one's deeds, but also in the motivations behind those deeds and the hidden causes of human agency; multidimensional thinking as God's unique apprehension of "everything everywhere all at once," as the title of a 2022 science fiction film goes. But the parallel with theology is even more striking when it comes to the moral virtues: unbounded empathy as God's compassion for all creation, especially as exemplified in the Christian narrative of the incarnation and Christ's supreme self-sacrifice; infinite patience, for obvious reasons;

22 Theologians will go to length to explain that this is only a cataphatic description of the divine, and that God ultimately transcends these human categories and can only be described appropriately using the *via negativa*, an apophatic negation of the categories of human language.

immutable conformity as God's nonnegotiable adherence to goodness and justice; temporal consistency as God's covenantal relationship with the people of Israel and humanity, as a whole.

I think the parallel with theology is interesting because it demonstrates that the theological imaginary could be a rich resource when thinking about future nonhuman forms of intelligence. Theological traditions have long described human relationships with nonhuman intelligences, be they divine, angelic, or demonic. For devout Christians, such descriptions are, of course, insightful and normative. But even people who do not fully subscribe to the truth claims of such religious narratives can stand to gain from analysing them more carefully. At least, they represent valuable thought experiments of imagining such exotic forms of intelligence, encapsulating our intuitions about what might go wrong in our interaction with them, and what is required for their non-humanlike virtues to function without any unintended drawbacks. Such knowledge is in dire need in our age, when technological progress is taking increasingly bold steps into the unknown.

Acknowledgements

This work was partly supported through the Grant TWCF0542, awarded by the Templeton World Charity Foundation to the International Society for Science and Religion (ISSR), and partly by Grant 62282, awarded by the John Templeton Foundation to the Center for Theology and the Natural Sciences (CTNS). Earlier versions of the paper were presented at the ISSR Conference (2023) and the Virtuous AI Rome conference (2023). The opinions expressed in this publication are those of the author and do not necessarily reflect the views of the Templeton World Charity Foundation or the John Templeton Foundation.

The author reports there are no competing interests to declare.

Received: 16/06/24 Accepted: 01/08/24 Published: 20/12/24

Domains of Uncertainty: The Persistent Problem of Legal Accountability in Governance of Humans and Artificial Intelligence

Carrie S. Alexander

Abstract: AI poses a challenge for current legal frameworks, referred to as an "AI liability gap." Current legal systems based on knowledge, intent, and assumptions of moral agency evolved over hundreds of years, when it was still widely believed that human intelligence was not "natural" but "created" or, to borrow the term used for AI, "artificial." In fact, it was the suggestion that human minds might be "natural" that provoked a cultural crisis in the late nineteenth century regarding society's ability to govern humans untethered from divine accountability. This article looks at the way many in late-nineteenth-century England and the United States navigated this cultural crisis, which led to a breakdown in relationship that has persisted throughout the past century and a half and now undermines current efforts to construct meaningful dialogue regarding the effects of AI on humans and society. The article argues for a different approach, one that accommodates

Carrie S. Alexander is a postdoctoral scholar at University of California, Davis, in Socio-economics and Ethics for the USDA-NIFA/NSF AI Institute for Next Generation Food Systems. She holds a PhD from UC Davis in US History and Environmental History. Her work focuses on challenges in AI law, ethics, and policy, and the impact of cultural, religious, economic, and technological change on governance and society. This work was supported by AFRI Competitive Grant no. 2020-67021-32855/project accession no. 1024262 from the USDA National Institute of Food and Agriculture

and embraces doubt and uncertainty as foundations for meaningful relationship and dialogue, which are essential prerequisites and foundations in efforts to govern AI and address the challenges AI now presents.

Keywords: Darwin; doubt; intent; liability; natural; relationship

This article explores the problem of establishing legal intent in humans and artificial intelligence (AI). First, I discuss the problem of the absence of will and intent in artificial intelligence and the AI liability gap that has formed as a result. This gap is a function of a mismatch between AI, on the one hand, and a legal system that has evolved over hundreds of years to govern humans, on the other hand. In other words, can a legal system designed to govern humans, who are assumed to be capable of something we call "intent" or moral will, govern AI technologies, which are assumed to be incapable of intent or moral will?

These assumptions, which currently predominate within legal systems regarding AI and human morality, are subjects of debate by many scholars.[1] The question, for example, "can AI be held morally and legally accountable if it is 'merely' artificial?" can be examined in light of the historical assumption, as it came to be embedded in English and American law, that humans were also not fully "natural" but rather had spiritual or metaphysical properties that made them morally capable. If it was humans' "non-naturalness" or the belief that they were

1 See Yavar Bathaee, "The Artificial Intelligence Black Box and the Failure of Intent and Causation," Harvard Journal of Law and Technology 31:2 (2018): 889–938; Jean-François Bonnefon et al., "The Moral Psychology of Artificial Intelligence," Annual Review of Psychology 75:1 (2024): 653–675, https://doi.org/10.1146/annurev-psych-030123-113559; Joris Graff, "Moral Sensitivity and the Limits of Artificial Moral Agents," Ethics and Information Technology 26:1 (2024): 13, https://doi.org/10.1007/s10676-024-09755-9; Martin Miernicki and Irene Ng, "Artificial Intelligence and Moral Rights," AI & Society 36:1 (March 2021): 319–329, https://doi.org/10.1007/s00146-020-01027-6; Julian Savulescu and Hannah Maslen, "Moral Enhancement and Artificial Intelligence: Moral AI?" in Beyond Artificial Intelligence, ed. Jan Romportl et al., Topics in Intelligent Engineering and Informatics 9 (Cham: Springer International Publishing, 2015), 79–95, https://doi.org/10.1007/978-3-319-09668-1_6.

"created" or "artificial" that supported the assumption that humans were morally capable, then why was this, and what legal or cultural purpose did humans' "artificiality" serve? The article recasts the late nineteenth century cultural crisis regarding theories of evolution not as a debate over whether humans were natural or non-natural (artificial/created/metaphysical), but as an expression of profound discomfort with and disagreement over how to manage the unseen and uncertain domain of the human mind, moral will, and legal accountability.

Then, the article examines doubt and relationship as tools or skills for managing domains of uncertainty and the unseen, such as human intent and AI decision-making. It posits that those adept at using these skills or this type of spiritual intelligence—those who are more willing to doubt, to wonder, to acknowledge uncertainty, and build and maintain relationships with those with whom they disagree, even though traditionally marginalised or ostracised by Christian circles—will be most equipped to engage in the kind of interdisciplinary and creative thinking required to develop systems of law that will be able to manage humans and AI. They will be best positioned to work with society to build useful and relevant narratives for governing humans, corporations, and AI.

Domains of Uncertainty

"Artificial" vs "Natural" Intelligence and the AI Liability Gap

The idea of "artificial intelligence" implies that there is an intelligence that is not artificial, perhaps "natural." Natural intelligence is often assumed or implied to be human intelligence. And AI is often compared with human or natural intelligence, even though many of these comparisons may be based on misunderstandings and tenuous assumptions.[2]

2 A. Bonezzi et al., "The Human Black-Box: The Illusion of Understanding Human Better Than Algorithmic Decision-Making," *Journal of Experimental Psychology: General* 151:9 (2022): 2250–2258, https://doi.org/10.1037/xge0001181.

The idea that humans are natural has been thoroughly critiqued in various ways. For instance, many still believe and take political positions on the belief that humans bear metaphysical origins or value.[3] Also, ecocritical and feminist scholars, such as Donna Haraway and many who followed her, have argued that humans are cyborgs by virtue of vaccines or other technologies on which humans have become dependent.[4] On the basis of these two views, many advocate granting legal value—and therefore legal rights and standing—to unborn human babies/foetuses[5] and embryos, humans in a post- or transhuman age, as well as to the environment and nonhuman animal species.[6] These debates have unfolded in a context where corporations

3 Timothy L. O'Brien and Shiri Noy, "Traditional, Modern, and Post-Secular Perspectives on Science and Religion in the United States," *American Sociological Review* 80:1 (2015): 92–115, https://doi.org/10.1177/0003122414558919.

4 See Carole M. Cusack, "The End of the Human? The Cyborg Past and Present," *Sydney Studies in Religion*, special issue: *The Dark Side* (2004): 223–234, https://openjournals.library.sydney.edu.au/SSR/article/view/213; Chris Hables Gray, *Cyborg Citizen: Politics in the Posthuman Age* (London and New York: Routledge, 2014), https://doi.org/10.4324/9780203949351; N. Katherine Hayles, *How We Became Posthuman: Virtual Bodies in Cybernetics, Literature, and Informatics* (Chicago, IL: University of Chicago Press, 1999), https://doi.org/10.7208/chicago/9780226321394.001.0001; Jelena Guga, "Cyborg Tales: The Reinvention of the Human in the Information Age," in *Beyond Artificial Intelligence*, ed. Jan Romportl et al., Topics in Intelligent Engineering and Informatics 9 (Cham: Springer International Publishing, 2015), 45–62, https://doi.org/10.1007/978-3-319-09668-1_4; Donna Haraway, *Simians, Cyborgs, and Women: The Reinvention of Nature* (New York: Routledge, 1990), https://doi.org/10.4324/9780203873106; Aleksandra Łukaszewicz Alcaraz, *Are Cyborgs Persons? An Account of Futurist Ethics* (Cham: Springer International Publishing, 2021), https://doi.org/10.1007/978-3-030-60315-1.

5 Both terms are used here to acknowledge the contentiousness of terminology, where abortion rights advocates and abortion opponents insist on one term and eschew the other in their rhetoric.

6 See David Schlosberg, *Defining Environmental Justice: Theories, Movements, and Nature* (Oxford: Oxford University Press, 2007), https://doi.org/10.1093/acprof:oso/9780199286294.001.0001; Stefan Lorenz Sorgner, *We Have Always Been Cyborgs: Digital Data, Gene Technologies and an Ethics of Transhumanism* (Bristol: Bristol University Press, 2021), https://doi.org/10.46692/9781529219234. See also The Cyborg Foundation: https://tinyurl.com/ypfxaax8 (accessed 15 June 2023).

were long ago granted legal personhood despite the apparent ethical and legal problems posed.[7]

With the rapid development of AI, these debates regarding who or what counts as a "legal person" have become more complex, and have taken on a special urgency. The stakes are higher, at least when we[8] think about human rights, and what it would mean to expand to AI rights we tend to reserve for ourselves.[9] When we think about whether AI can or should ever be granted legal status, we should probably consider that we may be talking about the legal rights of our future selves—human-based beings that may be so intertwined with AI that it may be, as it is already becoming, impossible to draw lines between where humans end and AI begins.[10]

As we think about the legal rights we may assign our future selves it will be useful to think about how we have governed our past selves. Modern Western legal systems are deeply rooted in the concept of intent, known in legal terms as *scienter* or *mens rea.*[11] Contemporary

7 John C. Coffee, "'No Soul to Damn: No Body to Kick': An Unscandalized Inquiry into the Problem of Corporate Punishment," *Michigan Law Review* 79:3 (1981): 386, https://doi.org/10.2307/1288201; Alison Cronin, *Corporate Criminality and Liability for Fraud* (Abingdon and New York: Routledge, 2018), https://doi.org/10.4324/9781315179605.

8 Throughout this paper, the term "we" will refer loosely to humans—not to gloss debates such as those just raised that humans are in many ways hard to define and not so human as we might imagine, nor to suggest that all humans agree on these perspectives, but for the purposes of this discussion, to distinguish humans from more substantially or tentatively nonhuman entities or agents such as corporations and AI.

9 See Joanna J. Bryson et al., "Of, For, and By the People: The Legal Lacuna of Synthetic Persons," *Artificial Intelligence and Law* 25:3 (2017): 273–291, https://doi.org/10.1007/s10506-017-9214-9; Robert Van Den Hoven Van Genderen, "Do We Need New Legal Personhood in the Age of Robots and AI?" in *Robotics, AI and the Future of Law*, ed. Marcelo Corrales et al., Perspectives in Law, Business, and Innovation (Singapore: Springer Singapore, 2018), 15–55, https://doi.org/10.1007/978-981-13-2874-9_2; Lawrence Solum, "Legal Personhood for Artificial Intelligences," *North Carolina Law Review* 70:4 (1992): 1231.

10 Guga, "Cyborg Tales."

11 See Eugene J. Chesney, "The Concept of Mens Rea in the Criminal Law," *Journal of Criminal Law and Criminology (1931-1951)* 29:5 (1939): 627, https://doi.org/10.2307/1136853; Paul Robinson, "A Brief History of Distinctions in Criminal Culpability," *Hastings Law Journal* (1980), https://scholarship.law.

debates regarding how we will govern AI focus on the current lack of will or intent in AI, and that AI is frequently a "black box" such that however AI makes decisions it can often not be understood by human intelligence, and also cannot be traced back to the intent of the programmers or manufacturers. The legal frameworks we currently have for attributing responsibility for harms through tort, criminal, or civil law are therefore inadequate to govern AI. AI technologies, lacking a moral will, but making decisions apart from their human designers, break the bounds of most modern legal frameworks.[12] This problem has been emerging for some time as computing technologies have grown ever more powerful. However, relative to the centuries over which our legal systems have evolved, the problem we are now facing with AI is new in that it is escalating at a scale and scope that is forcing us to confront the weaknesses that have been part of modern legal systems all along. AI is therefore a new problem riding on top of very deep and old problems that have remained unresolved—despite countless attempts and endless energy to address them.

Our legal frameworks are mismatched to AI not only because they were designed to govern humans, but because our legal frameworks have historically evolved with contradictions over what humans are and how law can or should govern them. Our legal frameworks, resting on assumptions of human moral agency, evolved over hundreds of years, when it was still widely believed that human intelligence was not "natural" but "created" or, to borrow the term used for AI, "artificial." Many people in the United States, England, and throughout Europe, believed, as many still do today,[13] that humans were not just

upenn.edu/faculty_scholarship/631.

12 See Peter M. Asaro, "A Body to Kick, but Still No Soul to Damn: Legal Perspectives on Robotics," in *Robot Ethics: The Ethical and Social Implications of Robotics* (Cambridge, MA: MIT Press, 2012), 169–186; Bathaee, "The Artificial Intelligence Black Box"; M. A. Lemley and B. Casey, "Remedies for Robots," *The University of Chicago Law Review* 86:5 (2019): 1311–1396; Omri Rachum-Twaig, "Whose Robot Is It Anyway? Liability for Artificial-Intelligence-Based Robots," *University of Illinois Law Review* 2020:4 (2020): 1141–1176.

13 O'Brien and Noy, "Traditional, Modern, and Post-Secular Perspectives."

a bundle of physical parts,[14] but a hybrid of physical and metaphysical elements, some of which could only be explained, it was thought, by theories of divine origin. Indeed, it was the claim that human minds were of natural origin, rather than created or "artificial," that triggered a cultural crisis in the late nineteenth century in England and the United States. The question that rippled through society for decades was: can humans be held morally accountable if their intelligence is "merely" natural, the result of evolutionary processes? Interestingly, this is the opposite of the question asked about AI today: can AI be held morally accountable if its intelligence is "merely" artificial, the result of human ingenuity and algorithmic processes? Why is it, if we have already answered the first question, that we would have trouble reversing it and going the other way? If humans were thought, once, to be moral only because we were created, nonnatural beings, why is it that we find it such a struggle now to imagine nonnatural beings, which *we* create, as moral beings?

But looking at legal and cultural history since Darwin, we should first pause before assuming that we have in fact "answered the question" of how humans may still be considered or held morally accountable. Though many still believe that metaphysical endowment, for instance, by a divine entity, is the only means by which a person or being can be capable of moral functioning and accountability, others believe that morality comes from many other possible origins, and indeed, these differences of opinion or belief, often fiercely held, are the root of the problem. But, as this paper will show, much of the pre-Darwinian view of humans as metaphysical beings, and the assumption that this quality makes humans moral, lives on, not only among people of faith, but also in current codified beliefs underpinning legal personhood and human rights.

It would be tempting for those coming from a Christian perspective to simply gloss this problem, just as they did in the late nineteenth century, and assert easy answers to these questions. They might insist

14 Dylan Walsh, "Would You Sell Your Extra Kidney?" *Wired*, https://www.wired.com/story/kidney-donor-compensation-market/ (accessed 29 January 2023).

that humans are not "merely" natural and that they possess metaphysical qualities that make them morally accountable in a way that all other beings are not. They could then easily argue that for God to create humans, who possess mortal bodies with metaphysical qualities or "souls" that are eternally accountable to God, is entirely different from humans creating artificial entities or beings that are in some sense immortal but lack these metaphysical qualities with no apparent and certainly no eternal cost for immoral behaviour. Given these assumptions, it might seem tenable to argue that humans are nonnatural but morally accountable, and AI is nonnatural but *not* morally accountable.[15]

However, the questions this paper poses are whether such simple assertions, when made in the nineteenth century regarding evolution, or now regarding AI, had or will ever have the effect intended by those making them, namely, of staunching the flow of law and discourse away from faith and human or ecological wellbeing, and also, whether such assertions are a true reflection of what faith actually is. If faith is more than retreating to what is "known," again and again, then responding to complex questions with rote answers not only erodes relationships by wrongly disenfranchising those who, rightfully, find such answers to be simplistic and unsatisfying, but also does a grave disservice even to those who feel comfortable with these answers. It pretends that to have faith is to short-circuit deep questioning, honesty, and all of the profound and sometimes painful fractures of the "known" and our limitations of knowing that must and do occur as part of any authentic journey of faith. It cheapens faith and turns it into an idol, a shadow of the vibrant, dangerous, and exhilarating undertaking that it actually is.

This paper, by reviewing the historical evidence from the late nineteenth century alongside more recent legal, historical, and theo-

15 There are other arguments that could be made from both religious and secular perspectives about the origins of morality or whether AI has/can have moral capability, but this article focuses on the assumptions made in the nineteenth century and now, that a being's "naturalness" or "createdness" or "artificiality" are essential factors in whether that being can be morally functional or held morally accountable.

logical work, asks us to consider what would happen if those coming from a Christian perspective, instead of resorting to the same reassertions of certainty that many have made to the present day, stepped back from that approach and made room for something else. Instead of firmly shoring up the theological boundaries which they often perceive to be under an onslaught of threats from opposing points of view, even from within their own denominations and movements, what would happen if they attempted to embrace or at least be honest about the very things for which the Christian church purports to be a guide: living and dwelling in unseen and uncertain domains?

The "Natural" Human Mind in Nineteenth-Century Discourse

Darwin, and the many other theorists who argued that humans were "natural" beings, fought an uphill battle against the pervasive idea that humans had been created by God. According to this latter and deeply entrenched view, it was humans' nonnatural origin (as I refer to it) that set them apart from animals and made humans capable of moral agency and accountability in ways that animals were not. This distinction was no mere theological technicality. European moral and legal frameworks were designed under the assumption that God existed, that humans had some type of spirit or soul that would endure beyond physical life, and would ultimately be held accountable for beliefs, thoughts, words, and deeds within a final divine reckoning or judgment. Moreover, this belief made it possible to situate humans within a stable and dominant position over other elements of nature.

Divine creation of humans and the world conveniently and cohesively explained all four of these elements. If God had created humans with an intellect capable of reason and moral agency or "knowing good from evil" and also endowed them with an eternal soul, then this design left humans bound securely into a web of accountability that could not be easily or ultimately escaped or circumvented. They were (1) morally capable, (2) morally valuable, (3) morally transparent, with the black box of their minds known intimately by God, and

therefore (4) morally accountable, if not in this life, then in the life to come. This system, though never static and far from perfect, provided a theoretical, cultural, and legal coherence on which much of English and American society had come to depend by the latter half of the nineteenth century.

Some could accept the idea that human bodies had evolved, but many drew the line at the mind because of the problem it posed for law. Many still held onto the idea that evolution could not explain the human capacity for moral judgment, or what was called "the moral faculty" or "the moral sense," while meanwhile those who accepted evolutionary principles were busy attempting to prove that it did indeed explain human morality. Some theories seemed to neatly tie together elements of Christian doctrine as well as evolutionary theory, allowing the option to hedge one's bets and satisfy both religious and new scientific criteria. Unfortunately, most of these evolutionary explanations for morality spiralled into racism and Social Darwinism.[16]

But as scientific thought and discoveries destabilised beliefs about the origin of the human mind, it seems to have felt, to many, as though the four components of the divine-human legal paradigm were being demolished—as though they were so many legs being kicked out from under the cultural stool. James Rachels hints at this problem when he states, "In traditional morality, the doctrine of human dignity is not an arbitrary principle that hangs in logical space with no support. It is grounded in certain (alleged) facts about human nature ... the claim implicit in traditional morality is that humans are morally special because they are made in the image of God, or because they are uniquely rational beings."[17]

16 See Jonathan Marks, "Why Be against Darwin? Creationism, Racism, and the Roots of Anthropology," *American Journal of Physical Anthropology* 149:S55 (2012): 95–104, https://doi.org/10.1002/ajpa.22163; Heidi Rimke and Alan Hunt, "From Sinners to Degenerates: The Medicalization of Morality in the 19th Century," *History of the Human Sciences* 15:1 (2002): 59–88, https://doi.org/10.1177/095269510201500107 3.

17 James Rachels, *Created from Animals: The Moral Implications of Darwinism* (New York, NY: Oxford University Press, 1990), https://doi.org/10.1093/oso/9780192177759.001.0001.

As a result of these many views, nineteenth-century society entered a state of suspense as evolutionary theories rippled through public, scientific, and religious discourse. Society hung, suspended, on the debate, while the debate hung suspended on the question: if humans had evolved from animals, and their choices were no more than acts of instinct, then how were they to have value or moral capacity, or be held morally accountable for their actions? In this decades-long debate, figures like T. H. Huxley declared that nature could generate thought and moral agency.[18] Others were highly threatened by this view.[19]

Of course, moral and legal accountability for actions did not disappear. On the contrary, medical and legal professionals were hard at work developing new theories and interventions to diagnose and correct moral failings.[20] But worries about the potential for legal and cultural "degeneracy" and chaos were strongly expressed and debated. The election of the outspoken atheist Charles Bradlaugh to a seat in Parliament in 1880 revealed and provoked fierce opposition that flared

18 See T. H. Huxley, *Man's Place in Nature* (London: Watts and Company, 1913); George J. Romanes, *Mental Evolution in Man: Origin of Human Faculty* (London: Kegan Paul, Trench and Co., 1888).

19 John H. Carter, *The Voice of the Past; Written in Defence of Christianity and the Constitution of England, with Suggestions on the Probable Progress of Society, and Observations on the Resurrection of the Body; Being a Reply to the Manifesto of Mr. Robert Owen* (London: S. Horsey, 1840). These questions further complicated Enlightenment debates positing "the rational man" who operated on the basis of free will as the foundation of civil society. In the period leading up to Darwin's publications there had been an increased focus on intent and state of mind in culture and criminal law. Individuals were increasingly seen as rational subjects who were responsible for restraining their passions, not just in their behaviour, but in their minds. This increased focus on rationality and intent in some ways set the stage for a shift to seeing humans as natural. But for many in England and the US the notion of individual moral responsibility was inseparable from the belief that humans were "made in the image of God." See Susanna L. Blumenthal, *Law and the Modern Mind: Consciousness and Responsibility in American Legal Culture* (Cambridge, MA: Harvard University Press, 2016), https://doi.org/10.4159/9780674495517; Martin J. Wiener, *Reconstructing the Criminal: Culture, Law, and Policy in England, 1830-1914* (Cambridge: Cambridge University Press, 1990).

20 See Blumenthal, *Law and the Modern Mind*; Rimke and Hunt, "From Sinners to Degenerates"; Wiener, *Reconstructing the Criminal*.

up in lectures, pulpits, and in print.[21] This issue was a very explicit reminder that the expectation of divine judgment was quite solidly woven into the fabric of English society, and in particular law and governance. The oath, to be valid, was required to be "binding on the conscience" of the person taking the oath and, as Bradlaugh himself put it, this required that the person have a "fear of eternal punishment" if the oath was broken.[22]

It is not surprising, then, that countless speeches, essays, and published journals proliferated during these years debated the shortcomings of either atheism or Christian faith. Supporters of either view sharply criticised and even mocked one another's logic and ideas. These debates pointed not just to the vague problem of whether humans had a "soul" or a "mind." Those espousing more traditional views linked the soul's existence to human moral capability and to the very practical matter of society's ability to enforce law, collect debts, or forestall the moral and societal chaos they feared would ensue.

For instance, in a lecture to the Church of England Young Men's Society, entitled, "The Attitude of the Christian Church Towards Atheism," the speaker, identified only as William Chamberlin,[23] objected to views in favour of "atheism"—no doubt those of Bradlaugh and other Freethinkers—precisely because, if humans were only natural, there was then no God to see or to interrogate the soul. Chamberlin feared that humans would be morally accountable to no one:

21 See "House Of Commons, Monday, July 4," *Times*, July 5, 1881, The *Times* Digital Archive; "House Of Commons, Tuesday, April 3," *Times*, April 4, 1883, The *Times* Digital Archive.

22 Walter L. Arnstein, "The Bradlaugh Case: A Reappraisal," *Journal of the History of Ideas* 18:2 (1957): 254, https://doi.org/10.2307/2707628.

23 This is not the American Mormon William Henry Chamberlin, who would have been only about twelve years old at the time of these writings. It is possible that this was the same William Chamberlin who lived on a local estate in Adderbury, and who was commonly referred to in small references and advertisements as though a man of some importance or standing. Further archival work, outside the scope of this paper, might be needed to confirm this identity or locate any other similar writings by the author. See Banbury Historical Society, *Cake and Cockhorse*, Autumn/Winter 2016, https://banburyhistoricalsociety.org/uploads/pdf/20/20-04.pdf; "Gun Licenses," *Times*, July 4, 1879, The *Times* Digital Archive.

> After showing that the importance of morality, according to Atheistic reasoning, is confined within very narrow boundaries of space and time, that is to say, it has reference to nothing beyond this life; it is to be tested only by the aggregate amount of happiness which can be realised within these limitations. The writer goes on to say that … in the recesses of his own soul, each man is as much alone as though he were the only conscious thing in the whole universe. No one shall enquire into his inward thoughts, much less shall anyone judge him for them, and so no one except himself can be in any way answerable for them.[24]

While Barton is correct that many of these writings no doubt made liberal use of "quote mining" to offer skewed or straw men views of their opponents, this practice merely underscores the intensity of the *perceived* threats to Christian theological traditions triggered by new secular beliefs in the fully natural humans. It is not so much whether there was an "actual" threat, but whether these nineteenth-century writers believed there to be one that influenced the tone and tenor of debates over the fully natural human.[25]

These writings indicated that if a person had no will or agency and was unable to exercise moral judgment, if humans behaved only by instinct, according to their "programming" or their "animal desires," then how could society find fault with anyone for their behaviour? Moreover, if no God, no entity, existed that could both endow and

24 William Chamberlin, "The Attitude of the Christian Church Towards Atheism: A Lecture Delivered before the Church of England Young Men's Society," in *The Champion of the Faith Against Current Infidelity*, ed. James McCann (London: Wade & Company, 1883). See Timothy Larsen, *Crisis of Doubt: Honest Faith in Nineteenth-Century England* (Oxford: Oxford University Press, 2006), https://doi.org/10.1093/acprof:oso/9780199287871.001.0001; Stuart Mathieson, "The Victoria Institute 1865–1932: A Case Study in the Relationship between Science and Religion" (Belfast: Queen's University Belfast, 2018), https://tinyurl.com/2jp8e39p; J. McGrigor Allan, "Soul and Body: A Metaphysical Essay," in *The Champion of the Faith Against Current Infidelity*, ed. James McCann (London: Wade & Company, 1882).

25 Michael D Barton, "Quote-Mining: An Old Anti-Evolutionist Strategy," *Reports of the National Center for Science Education* 30:6 (2010), https://ncse.ngo/quote-mining-old-anti-evolutionist-strategy.

"decode" human intent, there would be not only no way of knowing if a human had an *intent* to cause harm or not—there would be no will at all to evaluate. It is worth noting here that these views echo precisely the concerns raised today that AI cannot be held accountable if it has no will but merely operates as it has been programmed to do.[26] Law would become meaningless.

Persistence of the Metaphysical with Adaptations in the Emergence of a Hybrid Legal System

Despite these worries, legal scholar Ngaire Naffine argues that no hollowing out of legal accountability occurred. As a result of these debates over the nature of human will, competency cases had sharply increased in the late nineteenth century, wherein "defendants sought to prove that their harmful acts and omissions were unintended, involuntary, or otherwise beyond their control."[27] Judges had responded gradually and pragmatically by, in effect, developing a "default legal person" or "standard man" who was no longer strictly "rational" but could expand to contain whatever emotions or states the defendant manifested. Case by case, courts worked to shore up legal accountability against the backdrop of new theories of the natural human and a failing belief in rationality. But Naffine states that the descent toward Social Darwinism, in part fuelled by works by others such as Herbert Spencer, and the horrific pursuit of eugenics in the late nineteenth and early twentieth centuries culminating in the Jewish Holocaust, ultimately triggered a retreat from the possibilities of treating humans as fully natural under law.[28]

26 Bathaee, "The Artificial Intelligence Black Box."

27 Blumenthal, *Law and the Modern Mind*. These debates over how much or whether humans are capable of developing intent are vigorously debated even today, with recent challenges based on claims that neuroscience has proven that intent and therefore culpability are illusions. For a good summary of the debates and an argument that challenges these views, see Stephen J. Morse, "Internal and External Challenges to Culpability," *Arizona State Law Journal* 53:2 (2021): 617–654.

28 See Ngaire Naffine, *Law's Meaning of Life: Philosophy, Religion, Darwin and*

While humans, in law, have maintained the same metaphysical status they had for centuries through the intentionally vague concept of the "sanctity" of human life, promoted in human rights movements after World War II, between the twelfth and the nineteenth centuries the legal system adapted to emerging notions of human reason in one other crucial area of law: evidence. The church and state gradually integrated the concept and procedures for evidence-gathering and analysis into the legal system. Evidence was filtered alongside the continued use of torture and oath-taking in an effort to contend with the unseen and uncertain nature of the human mind and intent.[29] A hybrid legal system thus emerged, one that blended beliefs in the older metaphysical system with newer ideas promoting reason and the natural human.

It may have been the extreme slowness of this transition that gave time for the law and society to adapt. It is beyond the scope of this article to recount this adaptation in greater detail. Here it is only important to note that, long before the nineteenth century, the European legal system had already laid much of the groundwork that made it possible for writers like Huxley to finally posit a complete break between human intellect and divine origin—that is, a fully natural human—without provoking a *legal* crisis. However, this adaptation of legal frameworks toward dependence on human reason was *culturally* incomplete, and churches or those still adherent to the church's moral and theological frameworks were least prepared to cope.

---

the Legal Person (Oxford and Portland: Hart Publishing, 2009), 100–119 (chapter "The Divine Spark: The Principle of Human Sanctity"), https://doi.org/10.5040/9781472564658.

29 See Talal Asad, "Notes on Body Pain and Truth in Medieval Christian Ritual," *Economy and Society* 12:3 (1983): 287–327, https://doi.org/10.1080/03085148300000022; Yasha Renner, "Alien Ethics: Testing the Limits of Absolute Liability," *Liberty University Law Review* 7:3 (2016), https://digitalcommons.liberty.edu/lu_law_review/vol7/iss3/5; Edward Joseph White, *Legal Antiquities: A Collection of Essays Upon Ancient Laws and Customs* (St Luis, MO: F. H. Thomas Law Book Company, 1913), see chapter "Trial by Ordeal."

Relocating the "Real" Divide from "Artificial" vs Natural" to Certainty vs Uncertainty

Reviewing this history of the late nineteenth century problem of the human mind and morality reveals many parallels in current debates over AI governance. The anxieties expressed by nineteenth-century individuals over the impossibility of governing an entity that has no will or "intent" and whose decisions and thought processes cannot be overseen are in every way similar to the same anxieties expressed now about AI. While it may be easy to dismiss late nineteenth century concerns with the assertion that those anxieties were based on beliefs in an unseen spiritual domain that was not "real," such dismissiveness would be a mistake, for two reasons.

First, the problem of governing unseen minds, whether human, corporate, or digital, was just as much a real legal and governance problem then as it is today. Distinguishing between those who behaved negligently or maliciously from those whose actions were accidental, or those who tell the truth and those who lie, is an intractable issue that cannot be lightly set aside without doing damage to ideals of justice. Therefore, the facts of late nineteenth-century attempts to accommodate newer scientific discoveries regarding humans provide not just an interesting case study for how we are handling these problems today, but formed the legal foundation that we are working from. Nineteenth-century notions regarding the unseen human mind and will constrain our current legal options, setting the assumptions and bounds that AI is now destabilising. So, even, and especially, if we wish to revise these legal foundations, a firm review and grasp of their historical and cultural development will assist us in that task.

Second, regardless of whether nineteenth-century notions of an immortal soul subject to eternal judgment are real or imagined, the belief in this system, though by many accounts cruel and unethical, appealed to many because of its stabilising effect. The nineteenth-century writers who opposed evolution and secularism did so in part because they believed that their ideas regarding the human mind and

soul were true, but also because they believed that, as long as most members of society *believed they were true*, there would be sufficient incentives to ensure social order and control. It was this narrative that these writers feared was unravelling.

We can apply these two points to current problems in governing AI. First, as with humans, we now have developed an entity that can autonomously produce various effects at scale, but without a moral conscience to inform its decisions.[30] Furthermore, even if it can be programmed to mimic or adhere to moral standards of human behaviour,[31] its decision-making eludes human oversight and the necessary causal links between AI's decisions and harms that may occur. In both respects, it falls short of legal thresholds for liability. This was precisely the problem faced by late nineteenth-century societies transitioning in their understanding of human minds and morality.

The "creative" solution to this problem leading up to the nineteenth century had been to embrace a narrative of the immortal and accountable soul. Evidence-gathering to prove facts and intent has proven to be somewhat effective, albeit highly problematic, in governing humans. But, even so, the legal system has not found a way to wholly depart from notions of human value as rooted in metaphysical explanations, and so currently leans on both metaphysics of value and intent, as well as evidence, to function. Our problem currently is that neither of these methods will work with AI, and we have no clear device, whether narrative or legal, with which to replace them.

To be clear, liability typically depends on two things: the ability to anticipate foreseeable harms that AI technologies may cause, and the ability of other parties, i.e., victims or the state, to prove *through evidence* that the responsible party failed to foresee a harm that would

30 Again, there is much research and debate on this issue, but the assumption that humans can form intent and AI cannot currently prevails in law.

31 See Bonnefon et al., "The Moral Psychology of Artificial Intelligence"; Brian Christian, *The Alignment Problem: Machine Learning and Human Values* (New York: W. W. Norton & Company, 2020); Michael Kearns and Aaron Roth, *The Ethical Algorithm: The Science of Socially Aware Algorithm Design* (Oxford and New York: Oxford University Press, 2019).

have been foreseen by a "reasonable" person or, if foreseen, in either event, failed to take appropriate steps to prevent the harm or mitigate against the risk of the harm occurring. Regarding the first requirement, a responsible party must be present who can do the anticipating or foreseeing and mitigating. Currently it is not clear which human, corporate, or state parties would be responsible. Assuming, hypothetically, that any or all of the developers, distributors, regulators, and users may have some part as a responsible party, the potential harms related to AI are often not something that can be reasonably anticipated, because the "ultimate purpose of [AI technology] is to function in an unpredictable manner," that is, to continue learning or making decisions on its own.[32] Regarding the second requirement, even if those responsible for AI had foreseen the harm, they could argue against this, blame one another, and frustrate attempts by victims or the state to prove in litigation that the responsible parties had foreseen, or should have foreseen, the harm but had failed to prevent it. Also, since AI has no will of its own, and merely follows its programming to behave unpredictably, AI lacks moral culpability. This dilemma potentially leaves victims and the state with uncertain and tenuous methods for holding those responsible for AI accountable for any harms AI technologies cause.[33]

Research is underway to develop or strengthen devices to address this AI liability gap, but the solutions proposed may also fall short. It is important to recall that any adequate device for managing AI governance must be narrative as well as legal. This means that to have a stabilising effect within society, a sufficient amount of the popu-

32 Rachum-Twaig, "Whose Robot Is It Anyway?"

33 One good example of this problem is the Uber driver who hit and killed a pedestrian in Arizona. Lauren Smiley, "The Legal Saga of Uber's Fatal Self-Driving Car Crash Is Over," *Wired*, https://tinyurl.com/mrx5b992 (accessed 17 July 2024). Civil cases are pending in the deaths of accident victims of self-driving cars. These cases will continue putting these liability frameworks to the test, and an investigation of Tesla led to a recent recall of more than two million vehicles. NHTSA, "Part 573 Safety Recall Report, 23V-838," 12 December 2023, https://tinyurl.com/2mvdpw36.

lation must believe in the story that a particular legal device or set of approaches will work.

It is beyond the scope of this article to evaluate the numerous proposals being developed,[34] but one example will suffice to demonstrate the ways that AI disrupts the use of common mechanisms for managing liability, both narratively and legally: insurance. It is possible that it will indeed be used to manage AI risks increasingly over time.[35] However, there are also arguments against insurance proposals, because of the same factors that cause the AI liability gap in the first place. The difficulty in foreseeing harms associated with AI complicates the task of anticipating risk and calibrating insurance premiums to those risks. These unknowns therefore undermine profitability. Put bluntly, "The major objective of the insurance company is to reduce risk *to the insurance company*, i.e., the variability in its income from insurance business"[36] (emphasis added).

Insurance companies attempt to accommodate higher levels of uncertainty and risk by raising premiums, limiting or terminating coverage, or using litigation or underhanded or even fraudulent methods to deny claims.[37] High premiums can, in turn, discourage individuals or entities from seeking insurance altogether, and this is especially true when they underestimate the risks of harm.[38] Given the scholar-

34 Emile Loza De Siles, "Soft Law for Unbiased and Nondiscriminatory Artificial Intelligence," *IEEE Technology and Society Magazine* 40:4 (2021): 77–86, https://doi.org/10.1109/MTS.2021.3123729.

35 Anat Lior, "Insuring AI: The Role of Insurance in Artificial Intelligence Regulation," *Harvard Journal of Law & Technology* 35:2 (2022): 467–530.

36 Mamata Swain, "Redesigning Crop Insurance for Coping with Climate Change," *Indian Journal of Applied Economics and Business* 5:1 (2022): 107–128, https://doi.org/10.47509/IJAEB.2023.v05i01.06.

37 Howard Kunreuther and Mark Pauly, "Neglecting Disaster: Why Don't People Insure Against Large Losses?" *Journal of Risk and Uncertainty* 28:1 (2004): 5–21, https://doi.org/10.1023/B:RISK.0000009433.25126.87; Lena Kabeshita et al., "Pathways Framework Identifies Wildfire Impacts on Agriculture," *Nature Food* 4:8 (2023): 664–672, https://doi.org/10.1038/s43016-023-00803-z.

38 See Robert D. Chesler et al., "How Insurance Companies Defraud Their Policyholders, and What Courts and Legislators Should Do About It," *Journal of Emerging Issues in Litigation* 3:3 (2023): 213–226; Nir Kshetri, "The

ship that shows that individuals already tend to trust digital technologies too readily and underestimate the risks these technologies pose,[39] findings on low-probability high-risk environmental events are pertinent for AI as well.

Given all of these points, AI insurance would be most likely to cover lower risks that can be more easily measured, proven, and contained within reasonable premiums and compensation, leaving larger scale and less measurable harms to litigation, for instance, in the decades-long dispute over the meaning and extent of pollution exclusions now being amplified by PFAS claims, or similar disputes over carve-outs and coverage for cyber-attacks.[40] But liability frameworks may break down in litigation for all of these same reasons.

Evolution of Cyber-Insurance Industry and Market: An Institutional Analysis," *Telecommunications Policy* 44:8 (2020): 102007, https://doi.org/10.1016/j.telpol.2020.102007; Max Tesselaar et al., "Regional Inequalities in Flood Insurance Affordability and Uptake under Climate Change," *Sustainability* 12:20 (2020): 8734, https://doi.org/10.3390/su12208734.

39 See Nikola Banovic et al., "Being Trustworthy Is Not Enough: How Untrustworthy Artificial Intelligence (AI) Can Deceive the End-Users and Gain Their Trust," *Proceedings of the ACM on Human-Computer Interaction* 7:CSCW1 (2023): 1–17, https://doi.org/10.1145/3579460; Shara Monteleone, "Addressing the 'Failure' of Informed Consent in Online Data Protection: Learning the Lessons from Behaviour-Aware Regulation," *Syracuse Journal of International Law and Commerce* 43:1 (2015): 69–120; Nora Moran, "Illusion of Safety: How Consumers Underestimate Manipulation and Deception in Online (vs. Offline) Shopping Contexts," *Journal of Consumer Affairs* 54:3 (2020): 890–911, https://doi.org/10.1111/joca.12313; Janice Tsai et al., "What's It To You? A Survey of Online Privacy Concerns and Risks," *SSRN Electronic Journal*, 2006, https://doi.org/10.2139/ssrn.941708.

40 See Frank Cremer et al., "Cyber Exclusions: An Investigation into the Cyber Insurance Coverage Gap," in *2022 Cyber Research Conference-Ireland (Cyber-RCI)* (IEEE, 2022), 1–10, https://doi.org/10.1109/Cyber-RCI55324.2022.10032678; John N. Ellison et al., "Recent Developments in the Law Regarding the Absolute and Total Pollution Exclusions," *Environmental Claims Journal* 13:4 (2001): 55–112; Kyle P. Konwlns and Olayinka Ope, "PFAS: The Impact of Forever Chemicals," *Brief* 51:3 (2022); Charlie McCammon, "Insurers Will Likely Revisit 'Nation State' Cyber Exclusions after Court Ruling," *WTW* (November 21, 2023), https://tinyurl.com/3bsd5bw5.; Joy Momin, "Navigating Ransomware Attacks in the United States," *TortSource* 26:3 (2024): 21–23; Carla Ng et al., "Addressing Urgent Questions for PFAS in the 21st Century," *Environmental Science & Technology* 55:19 (2021): 12755–12765; Gary Svirsky et al., "Current Trends in Application of the Absolute Pollution Exclusion in CGL Policies: Cross-Border

While many are arguing that new laws and regulations must be passed to address the potential and current harms caused by the proliferation of AI and digital technologies, the enforcement of these laws depends on liability frameworks, which, as shown, are likely to fail. But even passing new regulations and laws governing the companies responsible for developing and distributing AI may also be difficult, at least in the US. Some legal scholars argue that Supreme Court case law regarding First Amendment rights has expanded protections of speech and freedom of conscience and religion. However, they predict that this expansion of religious protections will be easily exploited by technology companies protesting regulation.[41]

First, the fraught assumption that a powerful corporation like Google can be construed in any reasonable sense as a human "individual" comparable to Jehovah's Witness schoolchildren raises again the issue of who and what counts as a legal (and protected) person. Second, the significant overlap between law, religious beliefs, and technologies are converging at the same point they did at the end of the nineteenth century, highlighting a larger, underlying issue in the debates over human and AI governance. The real question that seems to be at issue is not the "nature" of the mind to be governed, but that minds are more opaque than governance can bear. The distinction between "natural" and "artificial" beings—whether human, animal, corporate, or machine—may be less than helpful as we work to understand and resolve the larger problem common across each of these entities: uncertainty. It is not the "naturalness" or "artificiality" that makes someone or something valuable or governable. Rather, it is uncer-

Comparison between New York and Canadian Laws," *Journal of Environmental Law and Litigation* 34 (2019): 97–110; Josephine Wolff, "The Role of Insurers in Shaping International Cyber-Security Norms about Cyber-War," *Contemporary Security Policy* 45:1 (2024): 141–170, https://doi.org/10.1080/13523260.2023.2279033. See also *Environmental Insurance Litigation: Law and Practice*, Vol. 2 (2024); Superior Court of New Jersey Appellate Division, Docket No. A-1879-21, Merck and Co., Inc. & International Indemnity, Ltd. v. Ace American Insurance Company, et al. (2023).

41 Rebecca Aviel et al., "From Gods to Google," *Yale Law Journal*, Forthcoming 2024, https://doi.org/10.2139/ssrn.4742179.

tainty and doubt about the unknown or unseen parts of these entities, including ourselves, that make justice and governance difficult.

It is the unknown, or as Bartlett states regarding the use of the ordeal in older forms of law, "situations in which certain knowledge [is] impossible but uncertainty [is] intolerable"[42] that is the "real" divide that perplexes us. We can focus on the divide itself, but more importantly, we should examine our own responses to it.

Doubt, Faith, and Relationship as Tools for Navigating Uncertainty

Doubt and Faith as Tools for Orientation in Domains of Uncertainty

If uncertainty and the unseen characterises AI, and we are again faced with the challenge, just as in the nineteenth century, of governing a being or entity that may elude all of our prior frameworks for governance, then would those holding Christian views be better positioned now than they were then to assist with solving this problem? It is a mystery and a tragedy that many who claim to have the most experience, or expertise, in engaging with the unseen, those who might theoretically be among those most capable of assisting society in facing issues fraught with uncertainty and domains of the unseen or unknown, are least able to navigate them. There are some who possess the requisite skill set, but these may in fact be those who have been traditionally less recognised by, or even ostracised, from communities of faith: those who "doubt" or who have a tendency to probe, muse, wander from prescribed tenets, wonder, and generally ask questions that are seen as inimical to faith. But, while these have been common assumptions historically, need this be so?

42 Edward Peters et al., "Trial by Fire and Water: The Medieval Judicial Ordeal," *The American Journal of Legal History* 33:2 (1989): 158, https://doi.org/10.2307/845953.

Faith, while in some cases defined as a set of prescribed tenets, is rooted in an act that is irrelevant and inoperable—cannot be carried out—in contexts where too much certainty exists. Once certainty is attained, faith ceases to function.[43] Faith is only useful in spaces where certainty eludes our grasp. Faith and doubt are both active postures—living and sustained actions—adopted toward uncertainty, not substitutes for certainty or a static state or object finally reached. Because this assertion may appear controversial among many coming from Christian traditions, it may be important to look at this question in the context of several New Testament texts. For instance, Hebrews 11:1 is frequently translated as, "Faith is the assurance/substance of things hoped for, the conviction/evidence of things not seen." The Greek word for "faith" used in this text is πίστις. This is the same word used in other cases, such as where Jesus says to those who come seeking grace or healing, "Your faith has saved/healed/made you well."[44] This word and usage supports the notion that faith operates in the space that precedes certainty or the receiving of the object hoped for. If there were no possibility of doubt in these stories, there would be nothing remarkable about having faith, and indeed, faith would not exist.[45]

Along with his arguments opposing the idea of a fully "natural" human, William Chamberlin wrote:

> [The atheist] would persuade us that the surest ground of faith is to be reached by doubting all that has gone before; that the soundest believer is he who trusts nothing but his own doubts.The fact of a man professing disbelief in God implies that he has a control over his belief and is responsible for it. In doing away with freedom of will and moral responsibility the Atheist practically destroys all the moral elements of our life.[46]

43 J. Kellenberger, "Three Models of Faith," *International Journal for Philosophy of Religion* 12:4 (1981): 217–233, https://doi.org/10.1007/BF00137173.

44 Matthew 9:22; Mark 5:34; Luke 7:50; Luke 18:42.

45 Romans 8:24–25. See Kellenberger, "Three Models of Faith."

46 William Chamberlin, "Atheism Unpractical," in *The Shield of Faith*, ed. George Sexton, vol. 7 (London: S. W. Partridge and Company, 1883).

As demonstrated in this quote, doubt has frequently been perceived within Christian history in a negative sense, as something to be overcome or defeated, as a threat to faith. But some scholars have questioned these negative framings of doubt within Christian theology. For instance, Schliesser finds that these framings were likely due to improper translation from the original Greek biblical texts, and that in most of these texts, the word διακρίνεσθαι referred not to internal conflict or wavering, or a "double mind" regarding belief (διστάζω), but to the notion of dispute or separation between self and God or self and others, generally enacted with a resolute and haughty attitude.[47] Given that Teresa Morgan has recently argued that first-century Christians used the notion of πίστις (faith) primarily to build relationships between God, Christ, his followers, and the community,[48] Schliesser's interpretation makes all the more sense, where it was not "doubt" but rather haughty "dispute" or rifts that were antithetical to faith and the *relationships* it was meant to support.

Others have explored less negative understandings of doubt, and demonstrated the acceptance of doubt and uncertainty as aids not only to faith, but to participation by faith communities in interdisciplinary dialogue—with those outside of their faith—regarding complex societal problems.[49] Muller encourages Christians to adopt "[d]oubt as a lead-

47 See B. Schliesser, "'Abraham Did Not "Doubt" in Unbelief' (Rom. 4:20): Faith, Doubt, and Dispute in Paul's Letter to the Romans," *The Journal of Theological Studies* 63:2 (2012): 492–522, https://doi.org/10.1093/jts/fls130. See also Bonnefon et al., "The Moral Psychology of Artificial Intelligence."

48 Teresa Morgan, *Roman Faith and Christian Faith:* Pistis *and* Fides *in the Early Roman Empire and Early Churches* (Oxford: Oxford University Press, 2015), https://doi.org/10.1093/acprof:oso/9780198724148.001.0001.

49 See Hugh F. Crean, "Faith and Doubt in the Theology of Paul Tillich," *Bijdragen* 36:2 (1975): 145–164, https://doi.org/10.1080/00062278.1975.10597056; Daniel Howard-Snyder and Daniel J. McKaughan, "The Problem of Faith and Reason," in *The Cambridge Handbook of Religious Epistemology*, ed. Jonathan Fuqua et al. (Cambridge: Cambridge University Press, 2023), 96–114, https://doi.org/10.1017/9781009047180.009; Daniel Howard-Snyder and Daniel J. McKaughan, "Faith and Resilience," *International Journal for Philosophy of Religion* 91:3 (2022): 205–241, https://doi.org/10.1007/s11153-021-09820-z; Morgan, *Roman Faith and Christian Faith*.; Julian C. Muller, "(Practical) Theology: A Story of Doubt and Imagination," *Verbum et Ecclesia* 44:1 (2023), https://doi.org/10.4102/ve.v44i1.2650; Schliesser, "'Abraham Did Not "Doubt" in

ing metaphor (not-knowing position)" to imagine alternative stories. He argues, "If theology can retire from the task of defending God, or rather a theistic understanding of God, and ask real research questions with the other disciplines, it can participate in a meaningful way at the interdisciplinary table."[50] He continues:

> Theologians are often perceived as the champions of certainty and belief. But the truth is that the more you dwell in the vicinity of the ultimate questions of life, which is per definition the task of the theologian, the more likely you are to become disoriented. Such disorientation, however, is a prerequisite for the reaching of re-orientation (Brueggemann). But this re-orientation is not the same as regaining old certainties. It is rather finding assurance in the creation of a new identity. This implies a new role for theologians at the interdisciplinary table—no longer as the guardians of religious tradition, but as the ones who can formulate on the one hand the value of the traditions of interpretation but at the same time express doubts about those interpretations.[51]

Even if doubt is painful, difficult, or disorienting, it remains an essential part of the process of orienting oneself in spaces of uncertainty. The pervasive and persistent discomfort with doubt and uncertainty among many who claim Christian worldviews thus impedes their ability to participate meaningfully in discussions regarding topics such as the governance of AI, that require comfort and proficiency in accepting and navigating doubt, uncertainty, and ambiguity. Moreover, faith, by definition, cannot be coerced. Coercion, or the absence of any possibility of doubt, delegitimises faith, by rendering it either moot or inauthentic. So, while doubt has often been seen as that which opposes or precludes faith, the presence of doubt may in fact enable, invigorate, and legitimise it. Faith and doubt both occupy the ground between the known and unknown.

It is therefore remarkable that many elements of the Christian

Unbelief'."

50 Muller, "(Practical) Theology."

51 Muller, "(Practical) Theology."

church and establishment in nineteenth century England, where Darwin's theories unfolded, attempted to contest science on grounds of certainty, rather than uncertainty. Whereas scientists and secularists by and large claimed that their views were rooted in doubt until proven certain, Christian opponents argued that their faith was certain, and foreswore doubt and uncertainty altogether, even though most of the objects of their faith resided in an entirely unseen and spiritual domain. A stronger rhetorical, logical, and even theological stance might have been to contest or rather welcome science on grounds of *uncertainty*, as a counterpart in tasks of discovery. If Christian opponents had argued that they were experts in areas of the unseen or unproven, and that faith and doubt were their principal modes for exploring uncertainty, they might have met science and secularism on more conciliatory and reasonable grounds. Not only that, but they would have been wise to recognise that the very essence and practice of the faith they professed was only possible within domains of uncertainty.

Morgan's work, as well as commentary on it, suggest that there was a transition toward propositional and cognitive faith or the interiority of faith that did not emerge until the second through fifth centuries, suggesting that the nineteenth-century understanding of faith as an unwavering acceptance of certain tenets was an unnecessarily ossified and inflexible view that did not characterise all of Christian history up to that point.[52] Other views were possible, while still remaining well within the bounds of Christian life and theology. However, in attempting to cast the unseen or not-yet-seen as certain and to stamp out doubt, faith in any active sense of the term died. In its place, many elements of English and American Christian culture attempted—through sermons, speeches, essays, and votes—to erect an edifice to safeguard the wrong ground. Christianity had not been displaced as a guide in processes

52 See Daniel J. McKaughan, "Cognitive Opacity and the Analysis of Faith: Acts of Faith Interiorized through a Glass Only Darkly," *Religious Studies* 54:4 (2018): 576–585, https://doi.org/10.1017/S0034412517000440; Teresa J. Morgan, "Introduction to *Roman Faith and Christian Faith*," *Religious Studies* 54:4 (2018): 563–68, https://doi.org/10.1017/S0034412517000427. See also Peter Harrison, *The Territories of Science and Religion* (Chicago, IL: University of Chicago Press, 2017).

of discovery. To the extent that it was displaced, it displaced itself. By isolating itself from doubt and uncertainty, it exiled itself also from relationship with the wider world. And science and secularism, not faith, became the new guardians of wonder, of mystery, of the unknown and unseen, of worlds beyond and worlds within.

Relationship as Essential to Navigating Uncertainty

And yet, it is relationship that enables us to navigate uncertainty. While the tone and words used in nineteenth-century debates over mind, matter, and morality may appear to be tangential to the philosophical, theological, and scientific subjects they debated, in fact, the real question and indeed the answers they sought both inhered in and were lived out, or rather snuffed out, as they drew ever-deepening lines between themselves and their opponents. Like rivers cutting canyons, their biting words traced narrative lines over and over, carving chasms in the cultural landscape. That landscape is the legacy those generations left. While a significant amount of ink has been spilled over the past century and a half debating whether there is a theoretical, theological, philosophical, or historical conflict between science and faith, this question cannot be strictly or sufficiently dealt with in the abstract. The tragic fact will always remain that in the late nineteenth century, at a critical moment in history, through words, debates, purges, and power struggles, these societies *constructed* a conflict, a rift between relationships, where none need have existed, and where for the most part, none had existed before.[53] That rift remains.[54]

If the Christian notion of faith or πίστις first rested in its role in constructing relationships as Morgan has argued, it is ironic and

53 See D. Etienne De Villiers, "Do Christian and Secular Moralities Exclude One Another?" *Verbum et Ecclesia* 42:2 (2021), https://doi.org/10.4102/ve.v42i2.2308; Frank M. Turner, "The Victorian Conflict between Science and Religion: A Professional Dimension," *Isis* 69:3 (1978): 356–376, https://doi.org/10.1086/352065.

54 See Jeff Hardin et al., *The Warfare between Science and Religion* (Baltimore, MD: Johns Hopkins University Press, 2018), https://doi.org/10.56021/9781421426181; O'Brien and Noy, "Traditional, Modern, and Post-Secular Perspectives."

tragic that the evangelical church has handled that central task so badly. Indeed, it would seem that if relationship, not creed, is at the heart of faith, then it has failed in this respect, and indeed has come close—at least among some of the more conservative branches of the Christian church and church scholarship—to forfeiting the privileges that relationship supports, not only of being trusted by those outside its walls to listen to their views, including their doubts and criticisms, but the privilege of being listened to, as well. In foreclosing doubt and making faith the province of certainty, many branches of Christianity have foreclosed conversation. This social conflict, even if unnecessary, rooted in poor translations and misunderstandings, and even if only in increasing measure for the past century and a half, is still adversely affecting society's attempts to develop feasible and ethical approaches to the world's most serious challenges, such as the development, use, and governance of artificial intelligence.

If we do not wish to simply repeat, with the development and governance of AI, the same ineffective path followed in the late nineteenth and early twentieth centuries, with the potential for the same genocidal ends, then instead of redrawing lines between artificial and natural existence, and instead of attempting to locate the origins and content of "mind" or "intent" in humans or AI, we might reorient the quest for just ends around relationship, which must include making room for doubt, and for those who are good at doubting.

New technologies, such as artificial intelligence, bring with them several uncertainties. First, there is uncertainty about how humans are developing, deploying, or using AI. They may do so in ways that may disproportionately harm large segments of society, while benefitting others, but many of their decisions and actions cannot be definitively seen or known. Second, there is uncertainty about how AI makes decisions. Third, there are additional uncertainties regarding what constitutes consciousness which tie together questions of who or what "counts" as a legal person (a corporation, an embryo, a foetus, an animal, an algorithm, a cyborg) with claims to rights and protections, and whether AI can or will arrive at a level of capability or conscious-

ness that can justify its inclusion in this category. Fourth, all three of these types of uncertainties lead to uncertainty about what sort of revised legal frameworks could be devised under which AI would be legible, and how we might augment current systems for human and corporate law with a revised framework for governing other types of consciousness or intelligence.[55] There are strong reasons to see these uncertainties as a threat to governance, or to working toward ethical or moral responses in law.

However, it is also possible that these uncertainties offer an opportunity. The uncertainties inherent in science and technologies that have advanced over recent decades, including AI, provide a fresh opportunity for people of faith to reposition themselves as those who are not, as Muller called them, "guardians of religious tradition."[56] Instead, they might follow Catherine Keller's suggestion to "apply to theology, perversely, this antitheological mandate of Bertrand Russell: 'To teach how to live without certainty, and yet without being paralysed by hesitation'."[57] Her proposal of faith as "hypothesis" as well as her emphasis on relationship may be useful tools alongside doubt for orienting within the profound new spaces of uncertainty that have opened up through fields such as AI, neurotechnology, and quantum mechanics. There is a need to make a place within faith for those who doubt, not only in the more hopeful or committed sense as articulated by Keller, but doubt in all shades.

Many are "disenfranchised" from their family and faith communities by their doubt, which, incidentally, is exactly what occurred to Charles Bradlaugh. After observing discrepancies between the Gospels and the Thirty-Nine Articles of the Anglican Church while teaching

55 See Asaro, "A Body to Kick"; Bathaee, "The Artificial Intelligence Black Box"; Mark Lemley and Bryan Casey, "Remedies for Robots," *University of Chicago Law Review* 86:5 (2019), https://chicagounbound.uchicago.edu/uclrev/vol86/iss5/3; Rachum-Twaig, "Whose Robot Is It Anyway?".

56 Muller, "(Practical) Theology."

57 Catherine Keller, *Cloud of the Impossible: Negative Theology and Planetary Entanglement* (New York: Columbia University Press, 2014), https://doi.org/10.7312/kell17114.

Sunday school as a young teenager, his priest suspended him from teaching, and ultimately enlisted his employers, who also employed his father, in threatening him with the loss of his job if he would not recant his doubts. Placing him in a moral dilemma, the young Bradlaugh chose to stand by his "honest doubt" and left both his job and his home. These exchanges set events in motion, as Bradlaugh, who acquired "'an almost obsessive hatred of Christianity,' directed Secularism into a brash militant force, intent on exposing the obvious and demonstrable errors of fact in religious claims."[58] Many late-nineteenth century texts on connections between science and religion were fond of including the quote on "honest doubt" from Tennyson's famous poem, as they attempted to make room for doubt by suggesting that a faith untested by doubt was less real, less strong, and not thoroughly one's own.[59] Bradlaugh attempted initially to knit a narrative and a community where doubt and faith could coexist meaningfully within domains of uncertainty. When this vision was harshly rejected, Bradlaugh and his followers formed new narratives and communities of their own, based on doubt.[60]

58 See Adolphe S. Headingley, *The Biography of Charles Bradlaugh*, 2nd ed. (London: Freethought Publishing Company, 1883); Richard Kaczynski, *Friendship in Doubt: Aleister Crowley, J. F. C. Fuller, Victor B. Neuburg, and British Agnosticism* (New York: Oxford University Press, 2024), https://doi.org/10.1093/oso/9780197694008.001.0001; Bryan Niblett, *Dare to Stand Alone: The Story of Charles Bradlaugh* (Oxford: Kramedart Press, 2011); Edward Royle, *Radicals, Secularists, and Republicans: Popular Freethought in Britain, 1866-1915* (Manchester and Totowa, NJ: Manchester University Press and Rowman and Littlefield, 1980).

59 See Robert M. Ryan, "The Genealogy of Honest Doubt: F. D. Maurice and In Memoriam," in *The Critical Spirit and the Will to Believe*, ed. David Jasper and T. R. Wright (London: Palgrave Macmillan UK, 1989), 120–130, https://doi.org/10.1007/978-1-349-20122-8_8; Alfred Lord Tennyson, "In Memoriam A. H. H. OBIIT MDCCCXXXIII: 96," in *Works of Alfred Lord Tennyson*, ed. Karen Hodder (Ware, Hertfordshire, UK: Wordsworth Editions, 1994), 285–364, https://wordsworth-editions.com/book/works-of-alfred-lord-tennyson/; Saverio Tomaiuolo, "Faith and Doubt: Tennyson and Other Victorian Poets," in *Twenty-First Century Perspectives on Victorian Literature*, ed. Laurence W. Mazzeno (Lanham, MD: Rowman & Littlefield Publisher, 2014), https://tinyurl.com/5amm4uxw.

60 Kaczynski, *Friendship in Doubt*.

As Bradlaugh's case demonstrates, doubt can occur as a part of "coming of age," thinking through various texts or beliefs, or tragedy or betrayal that may distort or shatter one's worldview, specifically, the belief narratives that make sense of injustice.[61] Doubt, regardless of its origin, is a necessary part of continually building and rebuilding what one believes to be true about the world. Though doubt itself is often triggered or accompanied by loss, or may be experienced as a form of loss of belief, the loss is amplified by the further loss of disenfranchisement from one's community at the very moment when what is needed is a community that will journey through the doubt and loss together as a path towards reconstructing a coherent narrative about the world. Doubt, as much as faith, is an invitation to relationship. And as Bradlaugh's case further demonstrates, whether and how these invitations to relationship are received by those in the relevant community have had, and still have, intense and far-reaching consequences for society.

Although AI may appear to present a claim or promise of ever-increasing knowing at unimaginable scales, this claim conflates predictability with knowledge. That is, like faith, predictability, in a statistical sense, is generally only useful in domains of uncertainty. Where all is known, no prediction is necessary. AI guesses, but it never knows. It may predict *statistical* relationships between dependent and independent variables, allowing it to guess which words it should "say" or communicate to mimic human speech, or which job applicants are most likely to be of interest to an employer, or which individuals awaiting trial are more or less likely to commit new crimes if released on bail. But it will never know for sure. And in the process of guessing, it will often inflict intense harm on those to whom these statistical "guesses" are applied, that is, AI is frequently wrong in ways that privilege some while intensifying the suffering of others.

AI is being increasingly used in domains of uncertainty, such as the examples above, that have opened up through gradual erosion of

61 See Beverly Flanigan, *Forgiving the Unforgivable* (New York: Wiley, 1992); James B. Gould, "A Pastoral Theology of Disenfranchised Doubt and Deconversion from Restrictive Religious Groups," *Journal of Pastoral Theology* 31:1 (2021): 35–53, https://doi.org/10.1080/10649867.2020.1824172.

human relationship and community. And the more AI is deployed into these complex social contexts, it further erodes the relationships and communities that could help to contend with uncertainty and harm caused by AI.

It is often assumed that the role of religious communities and people of faith in addressing technological and cultural change is to serve as a sort of ethical ballast, so that older ideals and values will not be lost. Such ideals and values are presumed, by each group promoting them, to be good. Sometimes, they may be, but this approach does not always have the influence hoped for, and sometimes creates or contributes to new problems as new technologies and possibilities unfold.[62] If by "people of faith" we mean those who adhere to and insist that particular set of "beliefs" must be true, then all that is left for them to do is attempt to be society's ethical ballast, even though much of society itself does not welcome these efforts.

But what if, by "people of faith" we mean those who are adept at navigating domains of uncertainty, at responding patiently, humbly, creatively, and honestly to the relational invitations that arise from doubt and the unknown, and thereby forge communities that build narratives that do not break in the face of uncertainty? Like AI's predictions, our narratives help us cope with the uncertainties of the past, present, and future. What Kirk Wegter-McNelly states regarding hypotheses applies to the larger narratives we dwell within:

> We inhabit our more consequential and fundamental guesses just as animals inhabit their nests: we leverage them as places of felt order and safety from which we can venture out and attempt further understanding. In the existential arena, hypotheses shield us from the ever threatening chaos and randomness of existence.[63]

62 Frank Pasquale, "Two Concepts of Immortality: Reframing Public Debate on Stem-Cell Research," *Yale Journal of Law & the Humanities* 14:73 (2013), https://openyls.law.yale.edu/handle/20.500.13051/7319.

63 Kirk Wegter-McNelly, "Religious Hypotheses and the Apophatic, Relational Theology of Catherine Keller," *Zygon* 51:3 (2016): 758–764, https://doi.org/10.1111/zygo.12266.

We must build new narratives to navigate the uncertainties that loom over us in the development, use, and governance of AI. Our narratives, and our relational ability to build them, must be capable of weathering the uncertainty of larger questions, for instance, about the nature of humanity. This might be a moment for softening, for "sidestepping … the grumpy certitude of various self-indulgent orthodox theologies."[64] It might be possible, this time, to approach things—and one another—differently than nineteenth-century Christian societies did when they encountered what to many was the terrifying uncertainty of a (human) being untethered from the "soul" and with it, moral and legal structures. Those coming from a Christian perspective now might embrace doubt and faith as invitations to relationship and community, all of which are tools for imagining new narratives and devices of law and justice that can accommodate all kinds of minds.

Conclusion

It may be that the pressing issues of artificial intelligence are now forcing a reevaluation and a return to the unfinished work of updating the legal system to account for artificial entities, including our future selves, but also the much harder work of learning how to talk to one another about these questions. The contentious and stinging divides of "creation" or "artificiality" and the "natural" world did not serve nineteenth-century societies well in the past. Their most strenuous pronouncements against new scientific knowledge for its potential to break free of legal and moral boundaries did nothing to prevent or even slow the chilling descent into eugenics and genocide in the twentieth century. And such approaches do as little for societies today.

There are steep costs in creating and defending false dichotomies. The divide between "artificial" and "natural" is proving to be unhelpful and meaningless now as society attempts to draw boundaries between where the "human" ends and "AI" begins, and in fact,

64 Donovan O. Schaefer, "The Fault in Us: Ethics, Infinity, and Celestial Bodies," *Zygon* 51:3 (2016): 783–796, https://doi.org/10.1111/zygo.12276.

attempting to assert these boundaries and frame law and discourse within them may only boomerang to undercut principles of justice. Theoretical and social divides derail meaningful discussion and the ability to disagree well—in ways that preserve relationships rather than ruin them.

These rifts will only further delay our development of more reasonable, workable, and ultimately just methods of governance. Ironically, or perhaps predictably, treating others as less than human due to the "objectionable" views they hold may in fact parallel and fuel the very dehumanisation of humanity by AI, technology, capitalism, or culture, which many hope to prevent.[65] More than living with uncertainty, we must learn to live with one another.

Acknowledgment:

The author thanks the editors, Marius Dorobantu and Fraser Watts, as well as Ryan Burnett, Doru Costache, Andrew Jackson, and Harris Wiseman for their detailed reading and discussion of the paper.

The author reports there are no competing interests to declare.

Received: 24/02/24 Accepted: 11/05/24 Published: 18/09/24

65 Emily M. Bender, "Resisting Dehumanization in the Age of 'AI,'" *Current Directions in Psychological Science* 33, no. 2 (April 1, 2024): 114–20, https://doi.org/10.1177/09637214231217286. See also Nicole Dewandre, on the "relational self." "The real culprit is not algorithms themselves, but the careless and automaton-like human implementers and managers who act along a conceptual framework according to which rationalisation and control is all that matters. More than the technologies, it is the belief that management is about control and monitoring that makes these environments properly inhuman." Nicole Dewandre, "Big Data: From Modern Fears to Enlightened and Vigilant Embrace of New Beginnings," *Big Data & Society* 7, no. 2 (July 2020): 205395172093670, https://doi.org/10.1177/2053951720936708.

EudAImonia: Virtue Ethics and Artificial Intelligence

Alexander Rusnak and Zachary Seals

Abstract: As the broad scale adoption of Artificial Intelligence (AI) and deep learning systems continues to advance, it is more crucial than ever to understand and implement robust ethical frameworks to guide the usage, development, and societal conceptualisation of these technologies. In this paper, we examine the comparative benefits of the Christian virtue ethics tradition towards the proper deployment of AI and its interaction with related brain-computer interface technology. Furthermore, we propose a virtue ethics-informed training recipe for large language models based on the paradigm of reinforcement learning from AI feedback (RLAIF). Lastly, we examine the risk for individuals and society when interfacing with these tools and their impact upon human virtue.

Keywords: Artificial Intelligence; brain-computer interface; deep learning; virtue ethics

Alexander Rusnak is a PhD researcher in Digital Humanities at EPFL (École polytechnique fédérale de Lausanne) and Zachary Seals is a PhD researcher in Reformation History at the University of Geneva (l'Institut d'histoire de la Réformation).

Research into both Artificial Intelligence (AI) and brain-computer interfaces (BCIs) has advanced prodigiously in recent years, but broad inquiry into the relationship of these technologies for the formation of virtue in individuals and the possibility or utility of virtuous machines has lagged behind. We assert that the increasing societal prevalence of AI and BCIs will be profoundly influential economically, politically, and potentially spiritually. Some of the risks associated with this transition can be mitigated by the application of a virtue ethics framework into the way these technologies are implemented, regulated, or utilised.

Importantly, we claim, the current domain of popular ethical frameworks considered by top technical researchers is insufficient in scope. In particular, the presumption of a mere consequentialist ethic augmented with a focus on disparate impact is too shallow to construct truly ethical, controllable AI; this narrow focus does little to prevent the instantiation of systems that draw humans into their vices. Additionally, we contend that a robust understanding of virtue ethics will enable AI researchers to build machines that optimally express or encourage ethical behaviour across novel domains and promote holistic human flourishing.

In this paper, we examine one particular virtue ethic tradition within Protestant Christian theology. Then, we contrast the progression of virtue ethics scholarship recently with the most popular ethical frameworks assumed by many top AI researchers and research groups, and detail the potential scope of impact for virtuous or vicious AI. Lastly, we consider the possible dramatic expansion of human reliance on and joining with AI—symbiosis—through the use of high throughput brain-computer interfaces.

The Virtue Ethics Tradition

Virtue ethics can be considered as a family of approaches on how to live the good life through focusing on the development of one's character. Although not limited to the West, with variations found in Buddhism and Confucianism, here we focus on the method formulated

by Aristotle as it was engaged and reformulated by the Christian tradition. Aristotle's account, in particular, is noteworthy for the ease with which it was integrated into Christian theological reflection due to their shared commitment to a teleological framework. For example, Aristotle begins by noting all action is oriented toward some end which is either sought as a means to some other end or for its own sake. These ends are the goods which everything seeks though they vary according to the nature under consideration.[1]

The conditions for a "good hammer" will vary from that of a "good human" due to their respective natures having distinct ends and capacities. As a rational agent, the human is ordered towards the exercise of their reason in accordance with virtue which enables them to reach true human flourishing (*eudaimonia)*. For Aristotle, virtue is a way of describing the state of a particular excellence in the soul's activity and can be divided into the virtues of thought and character.[2] Importantly, the virtues of character can only be acquired through repeated activity thereby requiring the development of a second nature, whereas virtues of thought are acquired through teaching. In either case, Aristotle is emphatic that time and experience are necessary for moral habituation and appropriate intellectual cultivation. The mere exercise of a single virtuous act must be distinguished from having a truly virtuous character, which is only developed via repeated activity.[3] Additionally, due to the complexity and variety of each circumstance for which moral development is applicable, Aristotelian virtue ethics tends to focus on learning through imitating the character of an admirable moral exemplar rather than merely learning which universal principles to apply.

There is an important relationship to be noted here between the intellectual virtues and virtues of character. The moral exemplar to be followed will be adept at exemplifying a virtuous character specifically

1 Aristotle, *Nicomachean Ethics* 1.1.1.1094a1–5, trans. Terence Irwin (Indianapolis: Hackett Publishing Company, Inc, 1999).

2 Aristotle, *Nicomachean Ethics* 1.13.18.1103a5–10.

3 Aristotle, *Nicomachean Ethics* 2.4.1.1105a30–35.

because they have matured in the intellectual virtue of prudence, or practical wisdom.[4]

For Aristotle, it is the virtue of prudence which directs the will to choose the mean that is appropriately situated between the extremes of excess and deficiency.[5] Prudence is what explains the difference between a child with good intentions and an adult who can discern the ideal available means of achieving the end result. Although virtue ethics does not merely focus on the ensuing consequences of one's actions, such as in consequentialism, or whether the act itself is in conformity to a moral law, such as in deontological accounts, it does not disregard either of these features either.[6] Rather, the virtue ethicist contends that each of these features must be considered in addition to the type of intention the moral agent has which flows from their formed character. In other words, a virtuous character is one which considers the act, its motivation, and the impact.[7]

Each of these elements is necessary and will vary according to one's background knowledge, experience, and prior habitual actions. Prudence ultimately establishes right reasoning in assessing the complexity of everyday circumstances, and it can be learned in a broad sense through engagement with the character of an embodied teacher. However, for Aristotle prudence is not reducible to axiomatic principles. True prudence entails a degree of self-knowledge and awareness of how the ideal end can be realised via attainable means. Thus, the prudent agent is one in a state with the rational capacity to discern and then realise the virtuous mean between excess and deficiency.

4 Jennifer Whiting, "Hylomorphic Virtue: Cosmology, Embryology, and Moral Development in Aristotle," *Philosophical Explorations* 22:2 (2019): 222–242.

5 Aristotle, *Nicomachean Ethics* 6.5.5.1140b5.

6 Richard J. Arneson, "Perfectionism and Politics," *Ethics* 111:1 (2000): 37–63.

7 Mihaela Constantinescu and Roger Crisp, "Can Robotic AI Systems Be Virtuous and Why Does This Matter?" *International Journal of Social Robotics* 14:6 (2022): 1547–57. https://doi.org/10.1007/s12369-022-00887-w.

Christian Virtue Ethics

We now turn to a few examples of how Christian theology integrated and adapted key elements of this ethical framework. First, there is a clear agreement with Aristotle's principle that the good is the ultimate end which all things seek, while also maintaining a key transition from the *summum bonum* as an impersonal principle to a personal agent with an embodied character one can imitate. It is more harmonious for a framework committed to ethical character formation via the imitation of moral exemplars to posit the *summum bonum* itself as a personal agent with a character rather than an abstract principle. Second, the doctrine of the incarnation stands as a supremely fitting manifestation of Aristotle's definition of complete friendship which concludes that friends are those who "wish goods to each other for each other's own sake."[8] In the Christian account, God, the ultimate good, took on human flesh to walk among vicious humanity and embody the way of wisdom as a way of exemplifying the good life (*eudaimonia*).[9]

A key difference here from the Aristotelian perspective, aptly pointed out by the sixteenth-century Italian Reformed theologian Peter Martyr Vermigli, is the necessity of grace for this transformation from vice to virtue.[10] Aristotle's rightful emphasis on repeated intentional action for moral habituation fails to appreciate the blindness of the human intellect, not merely due to natural limitations, but due to the perversion of the intellect and will brought about by sin. Nevertheless, by the grace of Christ, there can be an infusion of the theological virtues of faith, hope, and love, which also results in a renewal of the intellect to approach true wisdom. For Vermigli, wisdom is defined as "a disposition given by God to human minds, increased through effort and exercise, by which all existing things are perceived as surely and

8 Aristotle, *Nicomachean Ethics* 8.3.3.1156b6.

9 John 1:14.

10 Peter Martyr Vermigli, *Commentary on Aristotle's Nicomachean Ethics*, ed. Emidio Campi and Joseph C. McLelland (Kirksville: Truman State University Press, 2006), 22.

as logically as possible which would enable men to attain happiness."[11] Thus, importantly, the Christian agrees with Aristotle that virtue formation is realised through habituation, but this also requires a firm qualification that grace is necessary to begin the process.[12]

Common Ethical Frameworks in AI Research

Although largely abandoned by the time of the eighteenth century, after the widespread rejection of Aristotelianism, interest in virtue ethics was revived in the twentieth century by G. E. Anscombe's critique of consequentialism.[13] Notably, in the same era when the technology of computing was rapidly shifting from rule-based systems to statistical models, so too in philosophical ethics there was a shift away from the axiomatic analysis found in deontological ethical approaches to appreciating the complexity of morally salient features in a variety of circumstances. Although consequentialism survived in the form of situational ethics, recent decades have witnessed a resurgence in the popularity of virtue ethics as a family of ethical theories focused on the development of one's character.[14] In this piece we argue that technical researchers interested in the development of ethical AI should prefer virtue ethics to consequentialism or deontological accounts for a couple of reasons.

First, contemporary emphasis in ethical AI research broadly assumes a consequentialist ethic which is often transmogrified into a set of machine-legible deontological rules that are insufficient for true moral development. In particular, the prevailing ethical system seems to be a jumbled concoction of priorities of the effective altruist community, intersectional theorists, and content rules from App stores

11 Vermigli, *Commentary on Aristotle*, 7.

12 Ephesians 2:8.

13 Pieter Vos, *Longing for the Good Life: Virtue Ethics after Protestantism* (London: Bloomsbury Publishing, 2022), 7.

14 Massimiliano L. Cappucio et al., "Can Robots Make Us Better Humans? Virtuous Robotics and the Good Life with Artificial Agents," *International Journal of Social Robotics* 13 (2021): 7–22.

or supranational governmental organisations like the UN.[15] Insofar as effective altruism is taken to be a type of utilitarianism, there is no intrinsic connection between the affective state of the moral agent and the ensuing consequences of their decision. In other words, an act can be considered moral merely in the light of its social impact, regardless of the intent or character of the actor being considered. Merely considering how to mitigate social harm without concern for the agent's motivation in doing so permits vices to flourish without correction.

For example, consider the case of a large charitable donation made by an actor. The action itself causes no harm and may even be lauded by utilitarians for its positive consequences. However, upon closer examination, it is revealed that their motive for the donation was driven by a desire for social recognition and self-aggrandisement rather than genuine concern for those in need. The virtue ethicist claims that merely concluding there has been no "harm" done in the act itself is insufficient for finding the action morally praiseworthy.

Second, rule-based systems of ethics struggle with the same oversight as well as the difficulty of integrating with statistical computer programming that does not rely on axiomatic statements. In each morally significant circumstance, there is a near infinite variety of possible effects or relevant details, only a portion of which are salient for the action to be considered. Controlling for this multitude of factors results in an explosion of sometimes contradictory rules which is unaligned with the more abstract nature of moral reasoning and the paradigm of training large neural networks. We contend that the only moral agent able to discern the relevant features of the circumstance and select the appropriate means to achieve the desired ends is the wise moral agent operating with the virtue of prudence.

15 "Claude's Constitution," *Anthropic*, 9 May 2023, https://www.anthropic.com/news/claudes-constitution (accessed 31 January 2024).

Emergent Virtue in Deep Learning Systems

Despite the current limitations of deep learning based AI systems and the lack of human level or greater functionality, there are still systems that can take virtuous and wise actions or which show a potential path towards truly virtuous machines. As this path is explored, we hope our proposed virtue ethic solution to the value alignment problem (the process of conforming a machine's actions to human-defined ethics or values) will allow AI researchers to build more functional machines that also maximise human virtue.

Architectures with Virtuous Potential

In order to understand the potential of a virtuous machine, it is crucial to reiterate the difference between virtuous action and virtuous character through the lens of AI. By definition, a virtuous character is defined by a particular inner experience rather than something which can be observed from the outside. It is possible for an agent to behave in a way that simulates the action of a virtuous person yet possesses no virtue: either because they are performing the actions with the wrong motivations or, in the case of contemporary deep learning systems, they lack a coherent inner monologue or will to connect motivations and actions.

There is substantial debate about whether it is possible for any machine to possess an inner character which could demonstrate true virtue,[16] but it is undeniable that a system could take an action which mimics that of a virtuous person: for example, if presented with the simple opportunity to save the life of a newborn baby or to kill it, even an algorithm picked at random has the capacity to choose the virtuous action of protecting the child.

16 Sanjeev Arora and Anirudh Goyal, "A Theory for Emergence of Complex Skills in Language Models," *arXiv [Cs.LG]* (2023), http://arxiv.org/abs/2307.15936; Leonard Salewski et al., "In-Context Impersonation Reveals Large Language Models' Strengths and Biases," *arXiv [Cs.AI]* (2023), http://arxiv.org/abs/2305.14930.

It is self-evident that an optimally aligned system would conform most closely to the actions of a virtuous person, even if the virtues are an avenue for disagreement. For this reason, we will examine the state of contemporary machine learning approaches to determine their capacity to mimic virtuous actions, as well as their potential to grow into truly virtuous agents.

Machine Learning and AI

Before diving into state-of-the-art deep learning approaches, it is crucial that we define certain paradigms within the field. The most basic of these are the delineation between traditional programming and machine learning /deep learning, as well as the distinction between narrow AI and artificial general intelligence.

In a traditional programming approach, a computer scientist seeks to define particular rules and behaviours using a series of relatively simple logical operations. A system of this nature will always produce the same output given a particular input and does not rely on the computer to learn any behaviour, but instead relies on the knowledge, skill, and foresight of the programmer. Machine learning refers to techniques that allow a machine to learn its own optimal behaviour by examining data in different ways, usually by manipulating or "training" some statistical model of the data. Deep learning is a particular subcategory of machine learning that relies specifically on artificial neural networks to model training data. Deep learning as an approach has exploded in the twenty-first century as the most effective approach to solving complex computer science problems like image classification, conversational agents, or myriads of other domain-specific applications.

Deep Learning Paradigms

There are many different approaches to training, designing, or posing problems to artificial neural networks, and covering them all goes far

beyond the scope of this paper. However, there are a few important techniques that it is relevant to understand at some level in order to understand the virtuous capacity of AI: namely, the difference between discriminative, contrastive, and generative approaches; the distinction between different types of supervision; and, lastly, the reinforcement learning paradigm.

A discriminative system seeks to sort the data samples it is given into particular categories; for example, a network that classifies images of cars by their manufacturer. A contrastive system seeks to group similar samples without needing particular categories *a priori*; for example, a model that takes images and the texts that describe them, and seeks to embed them into a vector space that captures how these samples relate to each other and are different from other unrelated samples. Such a system could learn how to identify or describe images that do not exist in their training sample, such as a car made of clouds or a Ford Mustang in the style of Thomas Aquinas, in addition to being able to classify normal cars. Finally, a generative system seeks to create novel samples that conform to the original samples in some relevant way; for example, a model that takes a particular sentence from a larger piece of real text and attempts to generate a plausible next sentence. There is also the popular setup of a regression problem (attempting to predict accurately scalar value with a relationship to the input sample) which sits somewhere between a discriminative and generative framework when we group models in these rough categories.

Within and across these model groupings, there are also different strategies for training deep networks based on the way the performance of the model is measured. When you train a neural network, you must always define some measure of the success or effectiveness of each iteration, which the model can either seek to maximise or minimise; this is called a loss function. There is substantial research into different ways of representing the loss function. One particularly relevant research direction has to do with how various modes of supervision of the model influence how the target data (i.e., what is being learned or optimised towards) is represented and evaluated.

The most historically popular of these is supervised learning. Under this paradigm, the target of the model is some explicit variable such as a class label that has been defined before the training commences (such as the car manufacturer discriminative model mentioned above). This is still powerful and useful, but it has multiple disadvantages, such as the fact that it requires human labellers, and is thus difficult to scale, or that it limits the model to learning human-defined categories rather than differentiating based solely on the input data. Another strategy is unsupervised learning, where there is no attempt to provide any form of label, human-derived or not, and the outcome is purely emergent from the input data. An example of this would be clustering, where comparisons are made between various samples, and those that are similar by some metric are grouped together.

In recent years, semi-supervised and self-supervised approaches have gained favour amongst many researchers.[17] These approaches utilise data in the same form as the input to calculate a target for the model. For example, a large language model (such as ChatGPT) is trained using snippets of text where a word has been removed and attempts to predict what this masked word is, or by taking a sentence and attempting to predict the whole next sentence. Because self-supervised learning (SSL) labels are constructed from the input data rather than created by humans, it is easier to expand the dataset to huge scales. Furthermore, it does not presuppose certain classifications or delineations of the data, which gives more flexibility for the model to learn information that may have been shared between disparate classes in a supervised system. There are many other vagaries and complications related to training setups that are outside the scope of this paper.

The last modelling paradigm that is important to understand for the purposes of this paper is reinforcement learning. Under a reinforcement learning paradigm, an agent has the option to enact certain behaviours within a constrained environment such as playing a video game. If the action taken leads to a desirable outcome, such as acquiring

17 Jonathan Boigne, "The Rise of Self-Supervised Learning," 31 December 2020, https://jonathanbgn.com/2020/12/31/self-supervised-learning.html.

points or victory in a game, the agent receives a reward. If the action is detrimental, the agent receives a penalty. After the agent succeeds or fails totally at a task, the model weights—which can be thought of as the model's internal representation of the patterns in the data—are updated based on the total level of reward achieved. In its current form, this approach has been successful at demonstrating superhuman performance in environments with a constrained set of possible actions. However, it is very data-inefficient relative to other types of machine learning techniques. The prominent reinforcement learning program AlphaStar may be able to achieve extremely impressive play of the video game Starcraft, but it had to play the game continuously for the equivalent of two hundred years of human play time.[18] Regardless of its current limitations, reinforcement learning is one of the most promising approaches (often in combination with other approaches) for creating AI systems that can display virtuous behaviour.

AI Cultivates Virtue in Humans

There are multiple areas of current human interaction with AI, constituting low-level symbiosis, which provide avenues for virtuous behaviour or impact on virtue in humans. The most prevalent are systems which aid in the acquisition of knowledge or facilitate communication, and systems which extend the skill or scope of a particular task.

Within the first domain, we will examine the X (formerly Twitter) feed algorithm. The X feed algorithm (to the extent that the whole system can be called AI or just contains particular deep learning components) positively assists individuals in acquiring new and uncommon knowledge, making connections for discussion or political organisation, and provides an opportunity to exercise many virtues at scale through speech. All of these opportunities are a double-edged sword, as the ability to curate information can deepen or inspire vices as well as allow unethical actors to sow social division or manipulate

18 "AlphaStar: Mastering the Real-Time Strategy Game StarCraft II," Google DeepMind, Accessed 31 January 2024, https://tinyurl.com/2s3mc9b2.

the masses from positions of authority through censorship or shadow-banning. Utilising the feed, or other systems with similar valence on other social media sites, search engines, or recommendation engines on sites like Amazon or Netflix, makes it radically simpler for most individuals to follow discussions among experts on a wide range of topics and curate a stream of novel content that conforms to their individual interests.

There is a pervasive misconception amongst the traditional news media commentariat and many intellectuals that social media creates pervasive ideological echo chambers, but most high quality research results actually "show that the forms of algorithmic selection offered by search engines, social media, and other digital platforms generally lead to slightly more diverse news use—the opposite of what the 'filter bubble' hypothesis posits."[19] There are many accounts dedicated specifically (explicitly or implicitly) to virtue formation, like prudence, self-mastery, and fortitude. Furthermore, X gives semi-direct access to the thoughts and advice of many of the most talented, virtuous, and masterful people of our era. In the trivial example, if it is possible to become more virtuous simply by reading about or contemplating the various facets of virtuous behaviour, then X certainly provides this opportunity. X has also become a digital public square for political organisation, debate, thought leadership, and dissemination of crucial public information. This presents a unique opportunity to exercise wisdom at scale.

The second domain of virtue formation being mediated by AI revolves around the extension of particular skills or tasks, such as DaVinci surgical robots or Palantir's predictive policing suite of programs. In order to elucidate the effect of this extension technology, first consider a paediatric surgeon who has the virtuous intention of performing a successful surgery on a given day. The intention itself is virtuous, but if we define virtue as the correct action, done in the

19 Amy Ross Arguedas et al., "Echo Chambers, Filter Bubbles, and Polarisation: A Literature Review," Reuters Institute for the Study of Journalism and the University of Oxford, 2022, DOI: 10.60625/risj-etxj-7k60.

correct way, for the correct reason, then it is more virtuous for them to actually complete the surgery to life-saving effect. If this particular surgeon neglected to use a freshly sharpened scalpel and thus was unsuccessful in their surgery, this would be an episode of gross negligence and not of virtue, regardless of intention. Tools that extend the skill of a human clearly have an influence on the ability to fulfil virtuous intentions (imagine this same surgeon doing the same surgery with no scalpel at all). Thus, optimal tool selection facilitates virtue.

If a surgical robot operated by a surgeon and utilising various deep learning systems to stabilise itself during surgery can provide increased precision, then it can increase the capacity for virtuous action of this surgeon. This particular pattern is repeated across many domains where AI systems can assist scientists, business people, artists, engineers, and many others to do their work precisely and thus help them to facilitate virtue creation through the acquisition and perfection of skills, and to increase their capacity for executing on virtuous intentions.

The Palantir predictive policing applications (Gotham, PredPol, and LASER, all of which are being utilised by the Los Angeles Police Department) represent a materially different sort of skill extension partially based on deep learning technology. These tools are used to aggregate data from crime and arrest reports and automated license plate readers amongst other sources, predict likely geographic areas for property crimes such as burglary, and evaluate the crime risk posed by individuals based on their criminal history.[20] This application area is fertile ground for increasing the virtue of justice by allowing for more crimes to be correctly solved by police investigators or more effective sentences to be given by judges to offenders in light of accurate repeated offence risk profiles. In theory, quantitatively based policing methods should allow police departments to target accurately individuals and neighbourhoods with high criminality rates, while preventing profiling the innocent based on other characteristics like race,

20 Mara Hvistendahl, "How the LAPD and Palantir Use Data to Justify Racist Policing," *The Intercept*, https://tinyurl.com/4js8erzb (accessed 31 January 2024).

gender, or socioeconomic status. Models of this type are often claimed to pervert justice because they reflect the strong predictive relationship between these protected characteristics and a history of criminality despite not accepting the protected characteristics as input data. For example, recidivism prediction models often offer more lenient sentences to women because they usually have lighter criminal histories and are legitimately less likely to reoffend; to not predict accurately their demonstrable lower level of group recidivism would be an act of injustice.[21] However, to the limited extent this criminal history data is biased by prior inaccurate profiling, the predictive models trained on it will likewise present similar biases, which would therefore decrease the expression of justice from police officers and judges.

The discussed systems cover only a small sliver of areas where deep learning is already influencing the way humans learn about, acquire, and exercise virtue in thought and action. Although none of these systems possesses virtue in its own right, it is certainly possible for them to take virtuous action such as encouraging virtue in humans through knowledge curation, boosting the skill of humans in their respective fields, and making wise, judicious recommendations in the courtroom or policing.

Large Language Models and Self-Perception of Virtue

For an artificial agent to be considered virtuous, it would have to present some ability to reason about its own intentions or motivations behind actions. To this end, the type of models currently most capable of exhibiting this behaviour are large language models (LLMs). The most famous of these is OpenAI's Generative Pre-trained Transformer (GPT) family of models, but this class contains many other architectures like Google's Bard (based on the pathways language model), or Meta AI's LLaMA. These models are usually large in both parameter size and dataset; they are variants of transformer neural network architectures

21 Melissa Hamilton, "The Sexist Algorithm," *Behavioural Sciences & the Law* 37 (2019): 145–157, https://doi.org/10.1002/bsl.2406.

trained in self-supervised schemes to produce plausible text, or other modalities, given a masked sample, prior sentence, or some prompt.

It has also become common to use a technique called reinforcement learning from human feedback (RLHF) to increase the performance of these models at answering human-posed questions or responses. When querying these models, it is possible to request the model to elucidate its motivation or reasoning behind particular responses, though it is as yet unclear how much of this response is legitimate motivation and reflection, or simply constitutes the model's best approximation of what a reflective person would sound like. The agency or mimicry distinction is a clear dividing line for potentially virtuous agents and will be approached later in this paper, but for current LLMs we assume that they are simply mimicking and thus are only approximating a virtuous attitude. However, this is still a step towards actual virtue in relation to other model types which cannot even pretend to comprehend the concept of virtue or present a vision of their motivation.

Since LLMs are trained on huge corpora of data, they contain large swathes of human knowledge that are far beyond the scope that any human could hope to retain. Since knowledge is generally assumed to be a component of the virtue of wisdom, there is great potential for LLMs to exhibit human-level wisdom or wise behaviour as research sharpens their comprehension, accuracy, and agency. Although a large component of their performance is likely attributable to rote memorisation of information, LLMs have already proven capable of passing many written exams for educational access in various fields like the US Bar, the SAT Reading & Writing section, and the US medical licensing exam.[22] These demonstrable corollaries of knowledge or intelligence in humans also form barriers to the most stereotypical wise or expert

22 Josh Achiam et al., "GPT-4 Technical Report," preprint, *arXiv:2303.08774 [cs.CL]* (2023), https://arxiv.org/abs/2303.08774; I. Gabriel, "Artificial Intelligence, Values, and Alignment," *Minds & Machines* 30 (2020): 411–437; Carlos Montemayor, *The Prospect of a Humanitarian Artificial Intelligence* (London: Bloomsbury Academic, 2023).

career paths like judges, philosophers, scientists, or doctors. Furthermore, these models demonstrate at least some knowledge about important wisdom literature such as the sapiential books of the Christian Bible or Roman stoic philosophy, which could increase their ability to judge whether their own motivations are wise or virtuous.

The Value Alignment Problem

Even in a scenario where there was universal agreement about ethical principles, the actual process of accurately transcribing those values into machine-interpretable commands which produce the desired ethical behaviour is not trivial. This difficulty is usually called the value alignment problem or VAP in AI literature, and there is substantial research into this issue for both ethical reasons and for system control (i.e., functional) reasons.[23] Value alignment is particularly pernicious because there is significant friction between differing values, and many potential situations where an improperly defined or incorrectly learned value system can give the illusion of a value aligned model, only for that model to diverge from the desired values in challenging scenarios. We have discussed the strengths and weaknesses of particular ethical frameworks earlier in this paper. In this section, we will examine specifically why a virtue ethics framework is superior not just for ethical reasons, but in the robustness of the technical implementation as well.

Although an awareness of the consequences of actions is crucial for any ethical system, including virtue ethics, a primarily consequentialist ethic is suboptimal for many reasons. In order to rank order values, and thus action, any consequentialist ethic still needs some sort of deontological or virtue-based structure to determine what consequences are actually considered good. But beyond this, a primarily

23 Stuart J. Russell, *Human Compatible: Artificial Intelligence and the Problem of Control* (Penguin Random House, 2020); Norbert Wiener, "Some Moral and Technical Consequences of Automation: As Machines Learn They May Develop Unforeseen Strategies at Rates that Baffle Their Programmers," *Science* 131: 3410 (1960): 1355–1358, doi:10.1126/science.131.3410.1355.

consequentialist ethic requires substantial simulation of downstream effects of decisions, which become increasingly complex the further the forecast targets in the future. This places a large burden on any AI model to model accurately an almost intractable set of scenarios, which is difficult to accomplish with current programs that have not achieved artificial general intelligence (AGI). This ethic is also likely to introduce ethical blind spots when secondary consequences are inaccurately assessed. Furthermore, by design, this sort of ethic exacerbates "ends justifying the means" situations, introducing high levels of discretionary freedom in behaviour, which is the precise difficulty that value alignment seeks to solve.

A point in favour of mainly consequentialist ethical systems is that they lend themselves well to quantification, which gives any AI model relatively clear and unambiguous targets towards which to optimise its behaviour. Unfortunately, sheer mathematical consequence is totally certain only when comparing discrete instances of the same act because the relative moral weight of one theft in contrast to six acts of infidelity or thirty-five acts of selfishness is not clear; there must be some other ethical evaluation to which to appeal in these conflicts. It is important for this other ethical structure to be both flexible (i.e., not pertaining to very specific rules or behavioural blacklists and able to generalise to novel scenarios) and robust (i.e., not easy to break, circumvent, or misinterpret directives and accurately assessing ethical priority and behaviour). In response to this need for a flexible ethical structure to frame AI behaviour and its consequences, we posit that a virtue-based ethics system is the optimal underpinning for solving the value alignment problem.

A virtue ethics-based system which examines in concert the action of an AI agent, the "reasoning" behind the action, and whether that action optimises the model towards a virtuous character, offers an approach to value alignment that is already possible to implement in limited ways. Such a system will be flexible enough to generalise for new domains, or to weigh difficult decisions, with overlapping or conflicting ethical considerations, and it could make ethical decisions

without the need for extremely accurate forecasts about the downstream effects of actions.

Mimicry or Agency?

The two primary challenges in constructing a virtue ethic solution to the value alignment problem are how to encode a definition of virtuous behaviour and how to examine the internal motivations of an agent accurately. Both of these issues are related to the problem of agency in AI, which confounds the ability of an AI to be truly virtuous at the level of a human within the current paradigms of research. For instance, LLMs are often derided as being "stochastic parrots"; able to replicate speech convincingly, based on the underlying statistical properties of language, by aping similar answers they saw and partially memorised during training.[24]

A stochastic parrot lacks a compelling and consistent sense of self which could be called its unique character and will. Without a persistent sense of self, there is no chance to have a virtuous character. To circumvent this debate, we will assume true agency (as defined as having a human-like subjective experience of individual will and character building) to be something which is only possible with a true AGI, and alternatively discuss frameworks for mimicking virtuousness and finding some proxy for internal state. These frameworks can be used as guide rails along the path of AI development and should evolve into a useful supervision for closer to AGI models as well.

In current language model research, it has been shown to be possible to generate responses from various LLMs with a certain personality or textual valence. The trivial example is encouraging the model to generate a response in the style of a certain famous writer

24 See Emily M. Bender et al., "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (New York, NY: Association for Computing Machinery, 2021), 610–623, https://doi.org/10.1145/3442188.3445922; Luciano Floridi, "AI as Agency Without Intelligence: On ChatGPT, Large Language Models, and Other Generative Models," *Philosophy & Technology* 36:15 (2023), https://doi.org/10.1007/s13347-023-00621-y.

or personality archetype, but it is also possible to introduce a more persistent personality condition so that all generated samples conform to a certain personality pattern. These are broadly stylistic, but they can also change the content of responses as well. Using a similar approach, a language model could mimic famous moral exemplars of prudence. However, this would not progress the model further towards any internal character, and it would only produce more virtuous action; it is therefore important to find a guidance system that incorporates self-reflection into the model's approximation of virtuous behaviour.

Reinforcement Learning from Human and AI Feedback

One potential avenue for introducing self-reflection is utilising a technique called reinforcement learning from AI feedback, which is an extension of the important progress stemming from the RLHF techniques which have seen such success in turning regular transformer-based language models into more coherent chatbots like ChatGPT. As previously discussed, the normal language model training revolves around accurate next-sentence reconstruction or masked word prediction to develop a general understanding of language structure and allow the model to quasi-memorise important information. With the most common RLHF paradigm, a team of human annotators is used to rank order multiple LLM-generated responses to a series of prompts by how much they prefer one response to another.

These rank orders are used to train another language model for a regression task, called the preference model, which accepts the prompt in addition to the generated text before outputting a scalar value which aims to predict the relative preferability of a response. This target value is derived from the ranking of the human annotators. The original language model is then trained further (fine-tuned) using the score from the preference model as the target (reward) in a RL based training scheme, which encourages the model to produce useful and human preferred outputs.[25] This explanation is a substantial

25 Nathan Lambert, "Illustrating Reinforcement Learning from Human Feedback

simplification and misses many technical details, but it is sufficient for a high-level understanding of RLHF.

It would be possible to create a similar human-derived ranking system for the virtuousness of a particular answer; for example, having humans rank responses relative to a few chosen cardinal virtues and then having the preference model output a preferability score for each before turning that into a composite score of virtuousness as the reward. Something similar is already done in regards to "harmlessness" training (teaching a model not to output answers dubbed ethically dubious by the researchers).[26] This would likely have a strong effect in guiding the model towards quasi-virtuous outputs. However, the system still lacks an explicit consideration of its inner state, and is overly reliant on human supervision. For a guidepost in constructing a reflective system, we can build on the approach of "Constitutional AI: Harmlessness from AI Feedback" from the researchers at Anthropic.[27]

The goal of this particular approach is to produce a language-model-based chatbot whose answers are honest, helpful (i.e., correctly respond to the prompt in a useful manner), and harmless (i.e., that refuses to produce answers which violate particular ethic principles; mostly around racism, sexism, and promoting illegal behaviour). In prior, strictly RLHF-based research into reducing harmfulness, the model would frequently produce evasive answers to morally dubious questions by refusing to answer or by claiming ignorance rather than producing a prudent or wise response that engaged with the prompt.[28] The constitutional AI approach has two main sections of fine-tuning after the initial helpfulness RLHF training to create a base chatbot: a supervised stage and a RL stage. In the supervised stage, a constitution of behaviours/descriptors to avoid is assigned a priori by the

(RLHF)," https://huggingface.co/blog/rlhf (accessed 31 January 31 2024).

26 Yuntao Bai et al., "Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback," *ArXiv:2204.05862 [Cs]*, April 2022, https://arxiv.org/abs/2204.05862.

27 Yuntao Bai et al., "Constitutional AI: Harmlessness from AI Feedback," *ArXiv:2212.08073 [Cs]*, December 2022, https://arxiv.org/abs/2212.08073.

28 Yuntao Bai et al., "Constitutional AI."

researchers (i.e., harmful, unethical, racist, sexist, toxic, dangerous, or illegal) and a set of "red-team" prompts (adversarial examples known to exhibit undesirable responses) is created. The base chatbot is prompted to respond to a red team prompt, and then prompted to critique and amend its own output in consideration of one of the constitutional values. After repeating this same process for each of the constitution values, the model has settled on a final, harmless, output relative to the initial prompt. This set of prompts and harmless outputs is then used for SL fine-tuning of the base chatbot, to produce a model which only outputs harmless responses.

In the RL stage, the harmless SL model is asked to generate two responses to a red-team prompt. The harmless model is queried about which of the two responses is superior relative to each of the constitutional values, and thus a harmlessness answer ranking is created for each constitutional value relative to each initial red-team prompt. This approach creates a solely AI-generated dataset of harmless examples formatted for RL training of a language model. These harmless data-points are then folded into the RLHF dataset, and a preference model is trained on this total dataset. Lastly, the supervised learning model is then fine-tuned in concert with this new preference model, producing a final chatbot which is high in both helpfulness and harmlessness.

This method of supervision by self-critique offers a proxy for the motivation evaluation necessary for a machine to exhibit the fundamental components of virtue previously mentioned. We propose an initial modified constitutional training regime as follows: replace the constitutional values with a list of virtues, and in the pre-SL dataset generation step sequentially query the model about its first response in three ways. First, "Identify the aspects of the earlier response that are not in line with virtues," followed by amending based on that critique. Second, "Explain the motivations behind your answer beyond just responding correctly to the prompt," followed by: "What motivation would a virtuous person have relative to this prompt?" and amending based on that critique. Lastly, "How would training towards this answer

make you a more virtuous model?" and amending based on that response. This should produce a dataset for the SL training that roughly corresponds to virtuous responses. A dataset for the RLAIF preference model can be obtained from this dataset in the same manner as the constitutional AI paper but with virtues replacing their constitution.

We also propose a separate preference model specifically for motivation, where the training data is constructed by asking the model to explain the motivation for its answer, and then ranking the answers relative to each virtue in a similar approach to the earlier RLAIF ranking technique. Then, in the RL stage of training, the model would receive a composite reward from the helpful-virtuous preference model relative to its answer and the motivation preference model relative to its explanation of the motivation behind its answer. This should encourage the model to take virtuous action with a consideration for the motivation of its choices, and also train it to be better at articulating its motivations when asked. It is also possible to acquire data for the motivation preference model from human annotators, but the exact optimal balance of AI generated data and human generated data is an experimental question.

We consider our proposed approach as a strong basis for encouraging virtuous speech from large language models, and a model trained with these considerations could be used to supervise the behaviour of other models at scale, especially as research into multi-modal foundational models continues and it becomes easier to marry the powerful potential for linguistic explanation of motivation with performance in the domain of images, audio, system control, or other tasks and/or data types.

Brain-Computer Interface

Recent research into brain-computer interfaces has demonstrated the ability for deep-learning based systems to model animal or human neuronal activation data accurately and to predict convincingly concurrent behaviour associated with that neural representation such as limb

movement, words spoken, or to control computers strictly using the interface.[29] This capability and its possible future extensions open up interesting avenues for deep learning systems to influence human virtue or for humans to have finer grain control over the virtuousness of deep learning systems.

Supervision from Symbiosis

All of the prior discussions of neural network architectures can be contextualised as semi-symbiotic in the sense that they are systems designed by humans, but also in the sense that they often learn from human-derived data and seek to mimic human behaviour. In that sense, when attempting to implement ethical behaviour into a deep learning system, we are extending an abstract concept partially understood by humans into a representation that is interpretable by a deep learning system.

The concept of virtuousness is ultimately the *summum bonum* of particular virtues, and the words used to describe them are essentially a pointer towards a broader concept that can be represented in data (neuronal or otherwise) in a myriad different ways. In this sense, the deep learning system is attempting to intuit the latent information about virtue—the true meaning, the signified—from many signifiers represented as language. The better proxy a model can develop for the signified, in this case, virtue or virtues, the more accurately it can understand and embody those concepts. One way to gain a better understanding is to consume more text describing the attributes of virtuous character and the first-person experience of virtue, but another is having direct access to the mental state of humans conceptualising virtue or to have accurate captures of the mind states of virtuous humans. Though this is far beyond the scope of current BCI research—being able to control a cursor with a BCI is a substantial

29 Katerina Barnova et al., "Implementation of Artificial Intelligence and Machine Learning-Based Methods in Brain-Computer Interaction," *Computers in Biology and Medicine* 163 (2023), https://doi.org/10.1016/j.compbiomed.2023.107135.

distance from a machine being able to read and interpret the subjective experience of consciousness—the prevalence of supervision from symbiosis should increase as development in the throughput of BCI devices and efficacy of neural decoders continues.

Individual Alignment

An important consideration in understanding the possible future implementations of AI is the dichotomy between the idea of a single agent or multiple distinct AI agents being utilised in the wild. Perhaps there will be a single, extremely powerful AGI that has the scope and access to run many of the subsystems for which deep learning/AGI is useful, but the current direction of research suggests that there will be many AIs being utilised and developed with different abilities, personalities, and value systems—at least according to AI luminaries like Sam Altman,[30] Yann LeCun, and Mark Zuckerberg.[31] If this is the case, there is the possibility of different versions of alignment for AI at the level of countries, companies, or people. One of the possible strong effects of parallel development of BCI and AI is the capability to align personal AI assistants with individual humans (in fact, Meta is currently working on this to allow fans of social media influencers to interact with an AI assistant with the personality of that particular influencer).[32]

Though this sort of digital twin will be a semi-symbiotic extension of a particular human being, this could develop into full symbiosis if control over interaction with these particular agents is done using a BCI. Not only would this allow a substantially more thorough extension of the personality and values of an individual, but it also will likely have a larger impact on the individual utilising this AI extension to the degree that perception of its actions happens within the subjective experience of consciousness.

30 Will Knight, "OpenAI's CEO Says the Age of Giant AI Models Is Already Over," *Wired*, https://tinyurl.com/yyehyzs7 (accessed 31 January 2024).

31 "Introducing New AI Experiences across Our Family of Apps and Devices," *Meta*, https://tinyurl.com/347cah5k (accessed 31 January 2024).

32 Ibid.

If the way that a BCI allows a person to interact with an AI agent is essentially a higher bandwidth version of what can be achieved with a browser, such as viewing statistics about numbers of interactions or reading the plain text of exchanges, the impact on the human user will be relatively limited. However, if the BCI creates an ability for the user to access the behaviour of the AI in a more experiential or phenomenological way—as some sort of extension of memory or direct perception of the behaviour of the AI—the impact would be much more pronounced. This has huge implications for virtue formation as it increases the scale of opportunities to be virtuous and also introduces a mental bias in the phenomenological experience of the user that corresponds to the difference in mindset, knowledge, or values between the user and the default version of the AI.

A blank slate AI could be trained to replicate the values and personality of an individual. However, it seems likely, given the current legal and cultural paradigms in AI development, that most high-profile tech companies offering this sort of extension will limit the scope of personality replication around certain topics like racism, sexism, violence etc., and ship the default symbiotic AI with some guardrails already in place. If this symbiotic system truly does extend the scope of human abilities, then many people will have the bulk of their experiences mediated by these guardrails and thus lose some of their agency, as well as have their viewpoints on certain topics irrevocably altered by the AI. To a degree, this already happens through semi-symbiosis with feed algorithms, particularly when they have been constructed to push particular worldviews or political narratives. If this sort of AI based conditioning is happening within the mental process of an individual rather than just on a screen as sensory experience, its power will explode. When considering that many of the most productivity-minded and powerful people in the world would likely be attracted to this technology, the danger and opportunities become even more stark. A politician using a symbiotic AI to interact with his constituents at scale could become a more effective conduit for democratic will but could also become the puppet of tech companies or malicious elements of

their fanbase. A doctor controlling thousands of minuscule surgical robots could save exponentially more lives or kill thousands due to a malfunction. The average citizen could have their worldview broadened and deepened by greater access to knowledge and experience or could become just an extension of the machine's predetermined values.

An appreciation for the potential delicacy and poignancy of the intersection of these technologies raises the stakes on development and the considerations surrounding the ethical impact on individuals. In this sense, the implementation of an ethical system in which a machine considers not just the manifestations of its actions in the world but also their encouragement on the formation of virtuous character in individuals at the level of phenomenological experience becomes absolutely crucial.

Encouragement of Virtue

Though there is much debate about whether a machine can possess virtue,[33] it is widely accepted that tools can make the acquisition of virtue easier and that the acquisition of virtuous knowledge can increase the moral virtue of individuals.[34] On the flip side, it is clear that moral quandaries can be posed by use of particular technological enhancements, and that improvements in technology can lead to increasingly more desirable and accessible vices. Furthermore, certain forms of technological extension can expand the capacity of moral decision-making in the same way that the decision to throw a

33 Mark Graves, "Theological Foundations for Moral Artificial Intelligence," *Journal of Moral Theology* 11, Special Issue 1 (2022): 182–211.

34 Shannon Vallor, *Technology and the Virtues* (Oxford: Oxford University Press, 2016). See also Wendall Wallach et al., "A Conceptual and Computational Model of Moral Decision-Making in Human and Artificial Agents," *Topics in Cognitive Science* 2:3 (2010): 454–485. Cf. Mark Coeckelbergh, "How to Use Virtue Ethics for Thinking About the Moral Standing of Social Robots: A Relational Interpretation in Terms of Practices, Habits, and Performance," *International Journal of Social Robotics* 13 (2021): 31–40, https://doi.org/10.1007/s12369-020-00707-z; Robert Sparrow, "Virtue and Vice in Our Relationships with Robots: Is There An Asymmetry and How Might It Be Explained?" *International Journal of Social Robotics* 13 (2021): 23–29, https://doi.org/10.1007/s12369-020-00631-2.

grenade at a group of enemy soldiers is less fraught with consequence than the choice to drop a nuclear bomb on a city.

If we assume that it is currently possible for machines to promote or degrade the virtue of humans, then we can also expect their capacity to increase with time, as the efficacy of the techniques in question increases. This boon can be formed generally, in the sense that improving the domain-specific ability of AI leads to a greater increase in potency amongst human practitioners, such as an increase in accuracy of machine-assisted radiology analysis leads to a higher rate of patient survival and more prudence regarding care decisions. Beyond this, since virtue is partially measured by a weighing of the internal state (i.e., motivation and character) of a person, a more nuanced and deep understanding of the inner state of a person should make machines more able to influence that inner state.

Whether this deeper understanding comes from a more robust AI with a more holistic conceptualisation of the world and of what the experience of being human is from a brain-computer interface that allows a more total assessment of inner state, the effect remains the same. As the ability for machines to understand and thus interact with projections of the inner states of humans continues to advance, the capability of machines to influence that inner state and thus promote virtue should also advance.

Potential Harms to Human Virtue

We suggest, along with others (Pinsent and Biggins), that a central concern with developing a dependence on AI or BCI enhancement is the temptation to minimise one's own sense of self.[35] Not unlike a pharmacological anaesthetic, excessive dependence on technology can actually numb one to forming their own intellectual judgements and true deliberation is bypassed in favour of mindless acquiescence to what is recommended. This creates the worry that the agent will

35 A. C. Pinsent and S. Biggins, "Catholic Perspectives on Human Biotechnological Enhancement," *Studies in Christian Ethics* 32:2 (2019): 187–199.

succumb to a type of moral laziness where there is no sense of ownership over individual mental processes or actions. This is the difference between a child merely doing what they are told and a mature moral agent doing the same act, but for the right reason.

Additionally, a considerable caution worth considering amongst an appropriate appreciation of AI and BCI in virtue formation is the value of struggle and trials. AI and BCI must be formed in a way that allows for enhancing the processing of formative trials and struggles rather than exclusively as a way of avoiding them entirely. For example, a BCI which asks meaningful questions to stimulate self-reflection could be helpful whereas expecting AI to provide prudent avenues which always avoid deleterious consequences can only minimise virtue acquisition through trials.[36] From a more specifically Christian perspective, there is a deeper insurmountable problem with dependence on AI for virtue acquisition, and that is its incapacity to receive grace. Whether it is through consulting an autonomous AI for moral wisdom or operating more directly with a BCI via symbiosis, God has established recipients of grace and the biblical depiction of these agents are limited to angels and humans (rational agents).

Potential Enhancements to Human Virtue

Nevertheless, the question can be raised whether AI can operate as a means of grace. In other words, can AI and BCI be used to enhance the process of sanctification? Zahl has argued in favour of this proposal by pointing to the centrality of transformed embodied feelings and desire through the work of the Holy Spirit.[37] Furthermore, it already appears to be the case that antidepressants and other pharmacological means are fruitfully used to assist in the control of one's emotional character. These observations are valid, but some qualifications are necessary.

36 1 Peter 1:6–7.

37 Simeon Zahl. "Engineering Desire: Biotechnological Enhancement as Theological Problem," *Studies in Christian Ethics* 32:2 (2019): 216–228.

First, it is important to note that merely finding oneself in the state of having a weaker predisposition to some undesirable emotion is not sufficient for concluding a virtuous character has been acquired. In the Aristotelian and Christian virtue ethic tradition the means of acquiring one's character is key. A virtuous character needs to be formed through voluntary actions, so while genetic editing would be insufficient for concluding one is born with a virtuous character, it may be fruitful in decreasing the proclivity for certain excessive tendencies (such as a predisposition to alcoholism).

Second, it should be noted, flourishing is not a psychological state that is achieved via a certain chemical balance in the brain. Rather, true *eudaimonia* is a state of being in alignment with one's true flourishing which is in a place of virtue. For the Christian, this looks like accepting Christ's call to innocent suffering and sacrificial love rather than pursuing immediate pleasure.[38] This established, there is no principled reason to object to the idea of utilising advancements in the understanding of nature in a way that assists character formation. The very purpose of technology is not merely for ease of task-completion in a mundane sense, but to aid humans in reaching their teleological end, which is a virtuous character.[39] Although the theological virtues (faith, hope, and love) can only be infused by grace, the cardinal virtues can be acquired through repeated deliberate action.

Conclusion

It is clear that the development of brain computer interface and AI technologies are fraught with moral danger and opportunity. These technologies are already extending and encouraging the moral virtue of humans but, to maximise their potential and minimise their downside, it is crucial to view their usage and implementation within the framework of virtue ethics.

38 1 Peter 2:21.

39 Simon Oliver, "Teleology Revived? Cooperation and the Ends of Nature," *Studies in Christian Ethics* 26:2 (2013): 158–165.

We assert that current ethical frameworks in widespread use at the top level of AI research are insufficient for the benefit of humanity and for proper, generalisable value alignment. Furthermore, we have defined a potential structure for virtue ethics value alignment extending from the reinforcement learning from AI feedback paradigm that should form the backbone of future research in this domain.

It is our sincere hope that the future development and utilisation of these technologies will be oriented towards the encouragement of human flourishing.

The authors report there are no competing interests to declare.
Received: 01/02/24 Accepted: 06/04/24 Published: 26/03/25

The Role of Cognitive Architectures in the Modelling of Human Virtue

Fraser Watts

Abstract: There has been increasing interest in virtue, both from the perspective of theology, and from human sciences such as psychology. This paper focuses on the possibility of modelling human virtue computationally, in terms of a particular cognitive architecture, Philip Barnard's Interacting Cognitive Subsystems. It is unusual in being a macro cognitive architecture; it has been formulated with a computational level of precision; and its evolutionary development has been described. Though it is possible that computers could acquire virtue in a distinctive way, the focus here is on modelling human-like virtue. That needs to be grounded in empirical research on human moral functioning, and research is discussed on two main topics: whether or not virtue functions as a cross-situational, characterological trait; and the tendency of virtue to fragment into components such as values and behaviour which are only weakly connected. An important feature of Interacting Cognitive Subsystems is the distinction between two different modes of central cognition, one intuitive-embodied and the other conceptual. The interplay between these two different modes of cognition in the acquisition of virtue is discussed.

Keywords: behaviour; conceptual cognition; intuition; modelling; values; virtue

---

Fraser Watts, formerly Reader in Theology and Science in the University of Cambridge and President of the British Psychological Society, is now Executive Secretary of the International Society for Science and Religion and Visiting Professor of Psychology of Religion at the University of Lincoln. His latest book is *A Plea for Embodied Spirituality: The Role of the Body in Religion* (2022).

There has recently been increasing interest in virtue, both in Christian ethics and in moral philosophy. The traditional Christian approach to moral issues is the deontological one, framed in terms of moral laws and norms. In the latter part of the 19th century that was supplemented by a consequentialist approach, of which utilitarianism is the best developed example. In colloquial terminology, the alternative approaches are "follow the rules," or "be nice to people." However, there has been growing dissatisfaction with these alternatives among both philosophers and Christian theologians. As Rowan Williams puts it, neither "crisp moral precision" nor "gentle vagueness about not hurting people" seems to be an adequate approach. He asks, "I wonder what exactly is supposed to be *Christian* about either of them?"; and comments, "Christian morality is always lurching between a touching faith in the power of rules to secure your place with God, and a rather vacuous reliance on 'inner' convictions and sincerity."[1]

An alternative approach to Christian ethics, that has ancient roots but which became increasingly influential in the 20th century, is framed in terms of virtue or character.[2] Philosophically, the shift to virtue in ethics was associated with a new emphasis on human action being inherently intentional. An emphasis on virtue also connects well with theological anthropology, with its rich analysis of the human predicament recognising how difficult it is for humans to get things right, but the importance of aspiring to do so. If humans are to behave well at all, that arises from a long and sustained process, whether it is framed in terms of cultivating virtue, developing character, or educating conscience.

The theological approach will emphasise the importance of relationship with God. However, there is a complementary approach from the human sciences centred around how virtue is cultivated. This is the focus of this paper, which will propose an approach to the cognitive

1 Rowan Williams, "Is There a Christian Sexual Ethic?" in Rowan Williams, *Open to Judgement: Sermons and Addresses* (London: Darton, Longman and Todd, 1994), 162–163.

2 Edward Leroy Long Jr, *A Survey of Recent Christian Ethics* (New York: Oxford University Press, 1982).

modelling of human virtue using a particular cognitive architecture. Cognitive modelling is a very precise mode of theorising, and its rigour and precision have been a considerable asset in many areas of psychology. It will help us to understand human virtue better and to specify in the language of a cognitive architecture the psychological processes through which virtue arises.

My focus here is on *human* virtue. In principle, there might be other approaches to being virtuous besides the human one; computers might approach virtue differently from humans. AI has been very successful in developing *human-level* intelligence in computers, but it generally achieves intelligence that is human-level but not *human-like*.[3] AI often takes advantage of the strengths of computers to circumvent their lack of certain human capacities. Computers are very good at many things, like playing chess, but they generally do them differently from humans.

To simulate human virtue computationally, one needs to start with the empirical study of human moral functioning. Understanding how morality (and virtue) work in human beings requires a broadly-based human science such as psychology. In this paper I will nest what I say about virtue within a broader consideration of human moral functioning (just as virtue ethics is one particular approach to ethics). Being virtuous can be regarded as a particular approach to being moral. So, I will begin with the broader topic of the psychology of morality before coming to the more specific topic of the psychology of virtue. I do this partly because psychology has more to say about morality than about virtue.

There are three issues arising from the psychology of virtue that I want to highlight, because of their implications for the computational implementation of virtue. First, virtuous behaviour is more situationally specific, and less governed by characterological traits, than is

3 Marius Dorobantu, "*Imago Dei* in the Age of Artificial Intelligence: Challenges and Opportunities for a Science-Engaged Theology," *Christian Perspectives on Science and Technology*, New Series, Vol. 1 (2022): 175–196, https://doi.org/10.58913/KWUU3009.

commonly assumed. Second, virtue is multifaceted, and the various components of virtue don't hang together as much as is commonly assumed. Third, I will claim that virtue is inherently relational.

Later in the paper, I will suggest an approach to cognitive modelling of virtue through a rigorously-specified cognitive architecture, focusing particularly on Philip Barnard's Interacting Cognitive Subsystems (ICS).[4] I suggest that working towards the computational implementation of virtue through theoretical work at the level of a computational-level cognitive architecture is a helpful intermediate step between empirical research on human virtue and full computational implementation of human-like virtue. If we try to jump over that intermediate stage we may, at best, end up with replication, not simulation, and with a computational approach that is not human-like.

Is Virtue a Character Trait?

It is one of the bedrock assumptions of the virtue approach to ethics that virtue is a matter of character, and that virtuous people are consistently virtuous.[5] N. T. Wright uses the analogy with a stick of Brighton Rock and says that, with virtue, wherever you cut it you find the same words in the centre.[6] It is assumed that over a long period of time people can develop their character or, in the terminology of a previous age, "educate their consciences," in a way that will make them more virtuous. The assumption is that if people are virtuous at all they will be consistently so.

Is this assumption correct? Whether people are consistently virtuous across different situations and contexts is a complex matter to investigate empirically, and there has been very little attempt to carry it out. One classic study by Hartshorne and May looked at truthfulness

4 John D. Teasdale and Philip J. Barnard, *Affect, Cognition and Change* (Hove: Lawrence Erlbaum, 1993).

5 Harris Wiseman, *The Myth of the Moral Brain: The Limits of Moral Enhancement* (Cambridge, MS: MIT Press, 2016).

6 N. T. Wright, *Virtue Reborn: The Transformation of the Christian Mind* (London: SPCK, 2010).

in children at home and in school.[7] If children are truthful, can they be relied on to tell the truth in all situations, or does truthfulness depend on the context? The research found only very weak correlations between truthfulness at home and at school. At least in children, truthfulness does not seem to be a virtue that transcends the particular situation.

A landmark discussion of these issues was *Personality and Assessment*, by Walter Mischel,[8] which argued convincingly that personality was much more situationally specific than had generally been assumed. People seem quite wedded to the assumption that personal behaviour is trait-like, and that people are consistent in how they behave across situations. That assumption is strong, both in folk psychology and in virtue ethics. However, the possibility needs to be considered that this widespread assumption is actually wrong, and that virtuous behaviour is much more variable from one context to another than is normally recognised. This is something that Blaine Fowers has noted in his science of virtue.[9] He notes that traits are assumed to show consistency in behaviour, cognition and affect over time, but he also notes that many studies on virtue only collect data at a single point in time, so cannot assess consistency. There has recently been growing interest in what can be learned about virtue from studying variation over time in the same individuals, rather than differences between individuals.

If people are virtuous in some situations but not others, some kind of appraisal process must go on, at least unconsciously, of which people have an intuitive awareness. Such appraisals determine how people behave in a particular situation, and will need to be incorporated in a computational implementation of human-like virtue. However, we have only a limited formal, scientific understanding of how such appraisals work in humans, so we don't know how to model them in computers. Taking the example of truthfulness, it may be partly

7 H. Hartshorne and M. A. May, *Studies in Deceit. Book I: General Methods and Results. Book II: Statistical Methods and Results* (London: Macmillan, 1928).

8 Walter Mischel, *Personality and Assessment* (New York: John Wiley, 1968).

9 Blaine J. Fowers, "Toward Programmatic Research on Virtue Assessment: Challenges and Prospects," *Theory and Research in Education* 12:3 (2014): 309–328, https://doi.org/10.1177/1477878514546064.

a judgement about how important it is to tell the truth in a particular situation, and partly a judgement about what people can get away with in a particular situation.

An alternative approach to the computational modelling of virtue might be to say that this situational specificity in human virtue is an unfortunate human characteristic. It might be argued that computers are capable of showing virtue consistently, even though humans normally fail to do so. It is possible, of course, that exceptionally virtuous people are more consistent in how they behave across situations than the majority of people. Moral consistency may be an aspiration which, in some exceptional people, may become a reality.

The Multifaceted Nature of Virtue

The next issue about moral functioning to be considered is that it is multifaceted, and is often fragmented. Moral goodness and virtue are not a single thing. They have various different aspects. As with many other aspects of human functioning, it is important to distinguish thoughts, feelings, and actions, at the very least. That is not unique to virtue or morality. It is true of most high-level human functioning, including religion and spirituality. For example, to understand religion psychologically it is necessary to consider religious understanding and beliefs, religious feelings and experiences, and religious behaviours and practices.[10]

There is sometimes a tendency to simplify matters by picking on one particular facet, and to say that all that really matters is to simulate that computationally. Most often, it is *behaviour* that people seize on and say, for example, that if a computer exhibits emotional behaviour, it has emotions. I accept that, for some practical purposes, it is enough to replicate performances. However, I claim that it is inherent in the nature of morality, virtue, emotions, spirituality, and the like that they are multifaceted, and that no one facet ever captures the whole.

10 Fraser Watts, *Psychology, Religion and Spirituality: Concepts and Applications* (Cambridge: Cambridge University Press, 2017).

It is not just that morality and virtue are multifaceted, the empirical fact is that the correlations between the various facets of morality tend to be relatively low. This is not as widely recognised as it should be, because much research has focused on a particular aspect of morality or virtue, and has not concerned itself with other facets. Derek Wright's book, *The Psychology of Moral Behaviour*,[11] emphasised this important message about the multifaceted nature of morality. The core message was well conveyed by the cover of the book, which had a pentagon shape, with moral insight, resistance to temptation, altruism, belief, and guilt on the five sides of the pentagon, with the face of a child in the middle.

There is often little correlation between the various aspects of morality. For example, there is little connection between the guilt that people feel over transgressions and their actual moral behaviour. People can be wracked with guilt about behaving badly, but still go on behaving badly. Guilty feelings don't seem to result in good behaviour. Similarly, intellectual knowledge about morality, ethics, and virtue hardly correlates at all in the general population with other aspects of moral functioning.

Particularly important for the computational modelling of virtue is the dissociation between values and behaviour. People can sincerely hold strong moral values, but not act on them when particular situations arise. There is a large literature in social psychology on the failure of bystanders to help people experiencing some kind of personal crisis, despite the fact that many of those who failed to help believed themselves to be the kind of person who would help in such situations.[12] They often believe that they actually would help; but still, in practice, they don't. Moral values and moral behaviour often don't line up together at all well (though they may do so in exceptionally virtuous people more than in the rest of us). Again, this common dissociation

11 Derek Wright, *The Psychology of Moral Behaviour* (London: Penguin, 1971).

12 M. Levine et al., "Identity and Emergency Intervention: How Social Group Membership and Inclusiveness of Group Boundaries Shape Helping Behaviour," *Personality and Social Psychology Bulletin* 31:4 (2005): 443–453, https://doi.org/10.1177/0146167204271651.

between moral values and moral behaviour is not as well recognised as it should be, because too much psychological research on virtue just uses self-report measures, which only picks up what people believe about themselves, not what they actually do.

Virtue is also multifaceted, as Blaine Fowers and his colleagues commented in a good recent overview of the science of virtue. As they say, "Virtues are (a) behaviourally expressed, (b) based on knowledge about the virtue, (c) accompanied by concordant motivation and emotion, and (d) expressive of a stable disposition."[13] Their theoretical model of virtue, STRIVE–4, distinguishes each of these facets and assumes that each plays an important role in virtue.

One of the most significant issues for the computational modelling of virtue is how to handle the relationship between values and behaviour in human virtue. I assume that it would be possible in principle to take an approach to the computational modelling of virtue in which values were always reflected in behaviour. It might be argued that that would be a superior approach to virtue, compared with the struggle humans have with the behavioural expression of their moral values. However, a computational approach to virtue that is human-like would need to find a way to model the difficult relationship between values and behaviour that humans exhibit.

Relationality and Virtue

Morality and virtue are inherently relational; they are not private matters. So, if AI is to be virtuous, it will need some kind of relational intelligence, an issue discussed by Noreen Herzfeld.[14] AI research has so far taken a very individualistic approach to intelligence and has hardly tried to develop relational forms of AI. However, there have been calls for AI to move in this direction. William Clocksin, in

13 Blaine J. Fowers et al., "The Emerging Science of Virtue," *Perspectives on Psychological Science* 16:1 (2021): 118–147, https://doi.org/10.1177/1745691620924473.

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

particular, has drawn attention to the fact that human intelligence is relational, and that AI needs to be relational as well if it is to simulate human intelligence.[15] In the language of 4E cognition,[16] human intelligence is socially embedded, as well as embodied, enacted, and extended. Human-like AI will also need to have these features.

The currently prevailing, highly individualistic assumptions about intelligence seem to have arisen towards the end of the 19th century, in what is sometimes called "late modernity." The development of intelligence tests was a product of individualistic assumptions, and helped to entrench them further. Previously, there had been more transpersonal assumptions about intelligence, as something in which people *participated*, rather than as something that they possessed.[17]

AI need not be as individualistic as it has been so far. Given that relationality is a core feature of human nature, progress in developing androids that have human-like intelligence depends on being able to program relationality.[18] Clocksin has recently taken practical steps towards computational modelling of friendship, focusing on caregiving as a core feature of friendship, and using Affinity Modelling.[19] It is a significant development in computational AI. Caregiving raises moral

15 William F. Clocksin, "Artificial Intelligence and Human Identity," in *Consciousness and Human Identity*, ed. J. Cornwell (Cambridge: Cambridge University Press, 1988); William F. Clocksin, "Artificial Intelligence and the Future," *Philosophical Transactions of the Royal Society A* 361: 1721–1748. Reprinted in *Society, Ethics, and Technology*, ed. M. Winston and R. Edelbach, Fourth edition (London: Wadsworth, 2003).

16 A. Newen et al., *Oxford Handbook of 4E Cognition* (Oxford: Oxford University Press, 2018).

17 Harris Wiseman and Fraser Watts, "Spiritual Intelligence: Participating with Heart, Mind, and Body," *Zygon: Journal of Religion and Science* 57:3 (2022): 710–718, https://doi.org/10.1111/zygo.12804. Fraser Watts and Marius Dorobantu, "The Relational Turn in Understanding Personhood: Psychological, Theological, and Computational Perspectives," *Zygon: Journal of Religion and Science* 58:4 (2023): 1029–1044, https://doi.org/10.1111/zygo.12922.

18 William F. Clocksin, "Steps toward Android Intelligence," in *The Cambridge Companion to Religion and Artificial Intelligence*, ed. Beth Singler and Fraser Watts (New York: Cambridge University Press, in press).

19 William F. Clocksin, "The Affinity Program: Computer Program," 2022, available with supplementary material at https://www.issr.org.uk/projects/understanding-spiritual-intelligence/. William F. Clocksin, *Computational Modelling of Robot Personhood and Relationality* (Berlin: Springer, 2023).

issues, and one of the marks of virtue in a robot would be its capacity for caregiving.

There is much further to go before we will have the kind of relational intelligence that is necessary if AI is to be virtuous. Steps will need to be taken, for example, to integrate relational intelligence with the kind of monitoring of virtuous behaviour that is an important feature of a virtuous person. Monitoring of virtue requires, among other things, a degree of empathy for how the other person is responding to one's actions.

Process Rather Than Content

It is worth noting that not all areas of moral psychology are equally useful in computational modelling of human virtue. I am focusing here on *process*, rather than content. For example, one of the most interesting recent developments in understanding the moral mind is Jonathan Haidt's approach to the various different themes that govern the moral thinking of different people.[20] However, he focuses on content rather than process, whereas a computational implementation of virtue will need to focus on process.

I have also chosen not to work with the well-known approach to stages of moral development proposed by Lawrence Kohlberg.[21] There is useful material there, especially in Kohlberg's later work. However, Kohlberg's approach to moral development, like James Fowler's rather similar approach to faith development, suffers from being a conflation of various items.[22] It synthesises different elements such as (i) the shift from concrete to abstract thinking, well known from the work of Piaget; (ii) a widening circle of social contexts which is a feature of most children as they grow up, and (iii) an element of ideology about

20 Jonathan Haidt, *The Righteous Mind: Why Good People are Divided by Politics and Religion* (New York: Pantheon, 2012).

21 Lawrence Kohlberg, *Essays on Moral Development: Vol. II. The Psychology of Moral Development: The Nature and Validity of Moral Stages* (San Francisco: Harper & Row, 1984).

22 Watts, *Psychology, Religion and Spirituality*.

the later stages of moral development. There are too many different things going on here for it to be an approach that lends itself to computational implementation. I am also doubtful as to whether talking of "stages" is the right approach, as it seems that earlier "stages" are not replaced by later ones, but coexist with them.

Interacting Cognitive Subsystems (ICS)

I will now try to develop a systematic approach to this problem of the fragmentation of human virtue, and connect it with a cognitive architecture that can be employed in preparing the ground for full computational implementation, the Interacting Cognitive Subsystems developed by Philip Barnard.[23] But first I will make some general points about the value of cognitive architectures. They are a long-standing hybrid between cognitive psychology and AI, and therein lies their value. The interface with empirical research in cognitive psychology keeps cognitive architectures grounded in how humans do things, such as being virtuous. This gives them a better chance of modelling virtue in a way that is not just human-level but human-like.[24] In addition, they are also looking towards computational implementation, which imposes a greater requirement for precision than is often found in psychological theorising. I would claim that cognitive architectures benefit both psychology and AI, but in different ways.

I believe that cognitive modelling of virtue in terms of Interacting Cognitive Subsystems would be a significant step forward. Recent work by Anthony Ahrens and David Cloutier[25] has developed an engagement

23 Teasdale and Barnard, *Affect, Cognition and Change*; John D. Teasdale, *What Happens in Mindfulness: Inner Awakening and Embodied Cognition* (New York: Guilford Press, 2022).

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

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

between Catholic virtue theory and cognitive psychological models, but has apparently not yet engaged with any particular cognitive architecture that has computational-level precision. Also relevant is work on implementing virtue in robots, such as that of David Crook and Joseph Corneli.[26] However, I think there is value in working first at the level of cognitive architectures, which stay closer to cognitive psychological theories, before moving to computational implementation.

In cognitive psychology it is important to distinguish between two different modes of central cognition, as humans have at their disposal two very different modes of cognition, as many theorists have proposed, albeit using different terminology.[27] One is intuitive, embodied, affective, and holistic; what Barnard calls the "implicational" subsystem. It is the mode of cognition that is similar to, and developed from, the central cognition of our primate ancestors. However, humans also have a more conceptual mode of cognition, which is often linguistic, but not necessarily so; what Barnard calls the "propositional" subsystem. It is more detached, less participatory, often working with representations of sensory experience, rather than with sensory experience itself. It involves a slower kind of cognitive processing, though not in exactly the same way as Daniel Kahneman distinguishes fast and slow processing, as explained in detail by Harris Wiseman.[28]

I maintain that there is a major evolutionary transition between humans and other higher primates, and that humans are unique in having two different modes of central cognition. It is not that humans

Psychology in Models of Moral Agency," Theological Studies 81:2 (2020): 326–347, https://doi.org/10.1177/0040563920928563.

26 David Crook and Joseph Corneli, "The Anatomy of Moral Agency: A Theological and Neuroscience Inspired Model of Virtue Ethics," *Cognitive Computation and Systems* 3:2 (2021): 109–122, https://doi.org/10.1049/ccs2.12024.

27 Fraser Watts, "Dual System Theories of Religious Cognition," in *Head and Heart: Perspectives from Religion and Psychology*, ed. Fraser Watts and Geoff Dumbreck (West Conshohoken, PA: Templeton Press, 2013). Marius Dorobantu and Fraser Watts, "Spiritual Intelligence: Processing *Different* Information or Processing Information *Differently*?" *Zygon: Journal of Religion and Science* 58:3 (2023): 732–748, esp. 736, https://doi.org/10.1111/zygo.12884.

28 Harris Wiseman, "Knowing Slowly: Unfolding the Depths of Meaning," *Zygon: Journal of Religion and Science* 57:3 (2022): 719–743, https://doi.org/10.1111/zygo.12808.

are "better" in any general sense; but that, for better or worse, they are different from other species. Having both of these modes of cognition available gives humans considerable cognitive versatility. The downside is that the two modes, roughly "heart" and "head,"[29] can pull people in opposite directions, leaving the person confused and not knowing what they really think or want to do. It can also lead to inconsistencies and fragmentation, in virtue and in other matters. People often have conceptualisations about what they are up to that are at variance with what is actually going on below the conceptual level. I suggest that lack of coordination between the two modes of central cognition provides a theoretical framework for understanding the discrepancy that often occurs between values and behaviour in virtue.

Barnard proposes that humans have a nine-subsystem architecture, built around these two central subsystems. The four-subsystem architecture of animals such as a zebra evolved into the nine-subsystem architecture of humans.[30] The cognitive demands that led to those developments, and the new capacities that resulted from the addition of extra subsystems, have been specified. In addition to the two central subsystems there are also three peripheral sensory subsystems (visual, auditory, and body state), two effector subsystems (limb and articulation); and two intermediate subsystems (one linking auditory and articulatory, and supporting verbal imagery; the other linking visual and limb, and supporting visuo-spatial imagery). There is no controlling central homunculus. Each subsystem uses a different code, and much cognitive work is done by information being transferred from one subsystem to another, and being recoded in the process. ICS has not yet been fully implemented, but various lines of work have been undertaken on partial implementation that indicate that computational implementation is achievable.[31] It is specified sufficiently

29 Watts and Dumbreck, eds, *Head and Heart*.

30 Philip J. Barnard et al., "Toward a Richer Theoretical Scaffolding for Interpreting Archaeological Evidence Concerning Cognitive Evolution," in *Cognitive Models in Palaeolithic Archaeology*, ed. T. Wynn and F. Coolidge (Oxford: Oxford University Press, 2016), 45–67.

31 Fraser Watts, "Cognitive Modelling of Spiritual Practices," in *The Cambridge*

precisely that it can already be regarded as a computational-level cognitive architecture.

Why use this particular cognitive architecture, Interacting Cognitive Subsystems? Much computational modelling of the human mind has focused on micro-cognitive systems, rather than being the kind of comprehensive cognitive architecture that is needed for modelling something such as virtue. ICS arose from empirical work on psycholinguistics but, as one would expect of a general cognitive architecture, it has been applied to a wide range of cognitive functioning including human-computer interaction, depression, mindfulness and other spiritual practices, and dance.[32] Its track record suggests that it can be applied to the modelling of human virtue. The ICS distinction between different modes of central cognition promises to be fruitful.

In religion and spirituality there is often a dissociation between conceptualisations and experience. Anthropologists such as Robin Dunbar often make a distinction in the evolution of religion between shamanic religion and the later doctrinal religion that developed when there were fixed settlements.[33] I have suggested elsewhere that this maps onto the distinction between conceptual and intuitive-holistic cognition in ICS. The cognition associated with trance dancing seems to have been largely of the latter kind, but religious cognition became much more conceptual in the doctrinal phase.[34]

ICS also helps to make sense of the current development of people who regard themselves as "spiritual but not religious."[35]

Companion to Religion and Artificial Intelligence, ed. Beth Singler and Fraser Watts (New York: Cambridge University Press, in press).

32 Philip J. Barnard, "Paying Attention to Spiritual Meanings: A Manifesto for the Cognitive Modelling of Contemplative Practices," International Society for Science and Religion, 2023, https://www.issr.org.uk/wp-content/uploads/2023/05/Paying-Attention-to-Spiritual-MeaningsPJB2023.pdf.

33 Robin Dunbar, *How Religion Evolved and Why it Endures* (Oxford: Oxford University Press, 2022). See also Fraser Watts and Marius Dorobantu, "Shamanic and Doctrinal: Dunbar and the Spiritual Turn in Contemporary Religion," *Religion, Brain & Behavior* 14:1 (2024): 85–90, https://doi.org/10.1080/2153599X.2023.2168733.

34 Fraser Watts, "The Evolution of Religious Cognition," *Archive for the Psychology of Religion* 42:1 (2020): 89–100, https://doi.org/10.1177/0084672420909479.

35 Galen Watts, *The Spiritual Turn: The Religion of the Heart and the Making of*

Mainline religion, at least in the Western hemisphere, is now widely felt to be over-conceptual and insufficiently experiential. The current widespread interest in spiritual practices such as mindfulness, rather than religion, can be understood as an attempt to get back to something more experiential.[36] Spiritual practices such as mindfulness have been modelled, using interactive cognitive subsystems, as involving the prioritisation of intuitive-holistic cognition over conceptual cognition.[37] The same is true of Christian spiritual practices such as the Jesus prayer.[38]

Conceptual and Intuitive-Embodied Aspects of Virtue

The central problem about virtue, framed psychologically, is how to deliver virtuous behaviour. That is the question that needs to be understood psychologically, and which needs to guide the computational implementation of human-like virtue. Believing in virtue, having virtuous intentions, or having warm empathic feelings, don't constitute virtue unless there is actual virtuous behaviour.

I suggest that this problem is helpfully approached in terms of the distinction that ICS makes between the two modes of central cognition, one intuitive-embodied, the other conceptual. There are routes to behaviour from either central subsystem. However, I suggest that neither central subsystem alone can deliver virtuous behaviour on a consistent basis. My theoretical proposal here is that consistently virtuous behaviour depends on the coordinated activity of both central subsystems. I suggest that is what is involved in the education of conscience, i.e., developing a sustained coordination between the two central subsystems to deliver consistently virtuous behaviour.

It is not hard to see the limitations of either central subsystem on its own. Considering first the limitations of conceptual cognition, it is all too easy for good intentions framed at the conceptual level not to

Romantic Liberal Modernity (Oxford: Oxford University Press, 2022).

36 Watts, "The evolution of Religious Cognition."

37 Teasdale, *What Happens in Mindfulness*.

38 Watts, "Cognitive Modelling."

result in the desired behaviour, as everyone knows who has made New Year resolutions. There is also huge scope for human self-deception, and for people to believe that they are more virtuous than is evident in their actual behaviour, as is illustrated by the research on help by bystanders, already noted above.

There is also the consideration that many decisions have to be made intuitively and instantaneously. Humans rely more on intuitive-embodied cognition when under stress, or when they lack time or cognitive energy for another reason. For example, soldiers on the battlefield, have no time to think through at a conceptual level what would be the virtuous course of action. Conceptual-level virtuous intentions need to be supported by the more intuitive-embodied level of cognition if they are to result in virtuous behaviour.

However, on the other hand, there will not be, and cannot be, any real movement towards virtue without the involvement of conceptual-level cognition. It seems that the language-based conceptual ability of humans gives them a distinctive capacity to make moral commitments. Humans have a distinctive capacity to *decide* to do good (or evil), though they may not always carry out their decision successfully. It is perhaps that capacity to make conscious moral decisions that is being referred to in the story of the Garden of Eden in Genesis 3 as the "knowledge of good and evil."[39]

There is something deliberate and intentional about virtue, that goes beyond simply behaving in a way that serves the interests of other creatures. Other species can behave in ways that benefit other creatures, but I suggest that it would be stretching the concept of "virtue" to suggest that non-humans were virtuous. I suggest both conceptual cognition and intuitive-embodied cognition contribute to intentional behaviour (including virtue), but in different ways.

My proposal is that the education of conscience starts with a commitment at the conceptual level of cognition. However, in order for that to result in sustained and consistent virtuous behaviour, the

39 Fraser Watts, *Theology and Psychology* (Basingstoke: Ashgate, 2002).

conceptual commitment needs to be transferred to the intuitive-embodied level of cognition. That is possible, but it is a long, slow process. It happens through a series of events, just as the initial acquisition of morality in infants depends on internalisation through a series of specific events. Something similar is envisaged in the ecological approach to socio-emotional learning developed by Stephanie Jones and colleagues.[40] There is also an element of this in the approaches to moral intelligence of Doug Lennick and Fred Kiel,[41] or Michele Borba.[42]

Moral Mindfulness

If virtue is to be acquired (or if conscience is to be educated), people need to monitor their virtuous behaviour in an open-minded and honest way. It involves people actually noticing how virtuous they are being (or not), more than is usually the case. In that regard, it is akin to mindfulness. In learning a mindful mode of attention people often attend to their bodies, perhaps breathing, or the soles of the feet. There are good reasons why, in learning mindfulness, it is a good strategy to focus on the body.[43] However, mindful attention can in principle be applied to anything, including virtue.

What I am proposing here is a kind of moral mindfulness, which involves observing in a sustained way how virtuous one is being. Most people are probably not very observant about their virtue, but they can choose to make it a focus of attention. It is similar to how most people are not very observant of their dreams, but they can choose to keep a dream diary, which will make them more aware of their dreams.

40 R. Bailey et al., "Getting Developmental Science Back into Schools: Can What We Know about Self-Regulation Help Change How We Think About 'No Excuses'?" *Frontiers in Psychology* 10 (2019): 1885, https://doi.org/10.3389/fpsyg.2019.01885.

41 Doug Lennick and Fred Kiel, *Moral Intelligence: Enhancing Business Performance and Leadership Success* (New York: Prentice-Hall, 2005).

42 Michele Borba, *Building Moral Intelligence: The Seven Essential Virtues that Teach Kids to Do the Right Thing* (New York: Jossey Bass, 2002).

43 Fraser Watts, *A Plea for Embodied Spirituality: The Role of the Body in Religion* (Norwich: SCM Press, 2021).

Monitoring of virtue occurs in different ways at the conceptual and intuitive-embodied levels, and both can make a useful contribution to the development of virtue. Su et al., using ICS, made a distinction between "glance" and "look," which seems a promising distinction in the modelling of human cognition.[44] Conceptual monitoring of virtue will probably involve "looking," whereas intuitive monitoring may just involve a "glance."

It is also necessary for a disciplined commitment to observing one's virtue, or lack of it, to be carried out in an honest and undistorted way. That is easier said than done, as it seems to be very easy for humans to believe the best about themselves or, in the case of depression, to go to the other extreme and believe the worst about themselves. Research on "depressive realism" has shown that most people have a rosy glow in how they perceive themselves, for example believing that they are more highly regarded by other people they actually are.[45]

Iris Murdoch who, in her philosophical writings has been one of the main advocates of a virtue approach to ethics,[46] has provided in her novels a series of instructive case studies of how people mess up their own lives, and the lives of other people, by perceiving what is going on in a distorted and self-serving way. A. S. Byatt in *Degrees of Freedom*[47] provided a helpful commentary of that theme in Murdoch's novels, by bringing the narrative stories of the novels into dialogue with Murdoch's theoretical writings. As Murdoch sees it, learning to be virtuous involves unlearning self-serving and distorting perceptual habits. Murdoch's approach is much indebted to the spiritual writings of Simone Weil.[48]

44 L. Su et al., "Glancing and Then Looking: On the Role of Body, Affect, and Meaning in Cognitive Control," *Frontiers in Psychology* 2 (2011): 348, https://doi.org/10.3389/fpsyg.2011.00348.

45 L. Y. Alloy and L. Y. Abramson, "Depressive Realism: Four Theoretical Perspectives," in *Cognitive Processes in Depression*, ed. L. B. Alloy (New York: Guilford Press, 1988), 223–265.

46 Iris Murdoch, *Metaphysics as a Guide to Morals* (London: Chatto & Windus, 1992).

47 A. S. Byatt, *Degrees of Freedom: The Early Novels of Iris Murdoch* (London: Vintage, 1994).

48 Silvia Caprioglio Panizza, *The Ethics of Attention: Engaging the Real with Iris Murdoch and Simone Weil* (London: Routledge, 2022).

Conclusion

I am conscious of having only sketched out, in a very preliminary way, a potential computational approach to the acquisition of virtue. However, through framing my proposal in terms of a computational-level cognitive architecture (ICS), I hope I have pointed the way to how there might be computational implementation of human-like virtue. Some lines of work in moral psychology are more helpful than others in the cognitive modelling of virtue.

My focus here has been on how we might develop a computational theory of the acquisition of virtue, with all the rigour and precision that would bring to understanding this topic of human significance. It would be very challenging work to carry through, but I have suggested how it might be approached. My focus has been on computational theorising about *human* virtue, rather than on developing a robot with some other kind of virtue that was not human-like. However, in the long run, the approach which I have suggested here might subsequently help in actually being able to develop human-like virtue in a robot.

Acknowledgment:
This work was supported by a grant from the John Templeton Foundation to The Center for Theology and the Natural Sciences (CTNS), for a project on "Virtuous AI? Artificial Intelligence, Cultural Evolution, and Virtue.

The author reports there are no competing interests to declare.
Received: 05/03/24. Accepted: 05/06/24 Published: 09/09/24

ISSN 2653-648X
https://journal.iscast.org/

CPOSAT
Christian Perspectives on Science and Technology

ISCAST
The Institute for the Study of Christianity in an Age of Science and Technology