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.

Christian Perspectives on Science and Technology, New Series, Vol. 3,
Special Issue: Artificial and Spiritual Intelligence (2024), https://doi.org/10.58913/OWHV7073

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, a

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).

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. In

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

6 Biblical quotations are taken from the King James Version.

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

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.

> we have a sense of incompleteness and deep hunger for a kind of fulfilment that cannot be achieved solely within ourselves. Unlike 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

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?"

is entangled with 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

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.

rationalistic aspiration of the Enlightenment, suspicious of emotions, 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 duty to forget,

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.

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).