Artificial Minds and the Dilemma of Personal Identity

Christian Coseru
Department of Philosophy, College of Charleston
coseruc@cofc.edu

I. Introduction

All diurnal organisms are stirred to action by light, but as entomologists have long known, for nocturnal insects the pull of its radiance can also spell doom. The image of a moth drawn to flame is suggestive of the sort of self-destructive behavior that poets have long used to maximum rhetorical effect: "Thus had the candle singed the moth" proclaims Portia in the *Merchant of Venice*, pitying the fate that befell Arragon. The deceptive lure of artificial lights, however, is no longer mere fodder for the poetic imagination. Ecologists are warning us that species accustomed to navigating by moon or starlight, from beetles to seals and sea turtles, are getting confused, or worse.[1] And medical science is beginning to establish a link between night-time light exposure and various sleep and metabolic disorders.[2] Yet our Promethean ingenuity is a sight to behold: fueled by illumination tropes as metaphors for awareness and understanding, our mastery of light and of the electromagnetic wave spectrum has brought forth technologies, from CT scans and electromicrography to myoelectric sensors and solid-state lighting, that inspire utopian visions of genetically engineered and neurally enhanced post-humans alongside dystopian fantasies of runaway cyborgs. Among these technologies, AI stands out as the shiny new object commanding our attention.

The future envisioned as a result of these technological advancements now includes the prospect of artificial minds, minds that, if they were to come anywhere close to resembling ours, let alone to surpassing human intelligence, would pose serious existential and ethical challenges for humanity. Or so the story goes, if Susan Schneider's compelling foray into AI research is any indication. Indeed, *Artificial You: AI and the Future of Your Mind* (Schneider 2019) offers a philosophically savvy investigation of the debates that animate much of the contemporary fascination with the prospect of machine intelligence, which by some of the more optimistic

Philosophy East & West Volume 74, Number 2 April 2024 281–297

measures could be achieved within the lifetime of most millennials.[3] As books aimed at a general audience pondering the promises and perils of superintelligent AI go, *Artificial You* succeeds precisely where most others falter: it brings conceptual clarity to often vague and nebulous claims by techno optimists about the impending emergence of consciousness from computational complexity, while at the same time offering a sober reminder that the technology underlying, say, artificial neural networks is far more advanced than most AI skeptics realize.

I will not address Schneider's survey of the state of the art in AI research, which is succinct, informative, and for the most part illustrative of the transhumanist ethos she appears to embrace. Instead, I will focus on the problem of personal identity and the seemingly insurmountable challenges it raises for the prospect of radical human enhancement and synthetic consciousness. Specifically, I will argue that conceptions of personal identity rooted in psychological continuity akin to those proposed by Parfit and the Buddha may not provide the sort of grounding that many transhumanists chasing the dream of life extension think that they do if they rest upon ontologies that assume an incompatibility between identity and change. I will also suggest that process ontologies that take change to be primary, such as those that align with contemporary systems biology, offer a better way out of the personal identity dilemma. But the solution in this case, which regards biological organisms as processes rather than things, may constrain the possibility of biologically inspired superintelligent aliens (BISAs), which Schneider (following Bostrom 2014) posits as possibly the most common form of (extraterrestrial) superintelligence in the universe.

II. Enhancement Beyond Recognition

Perhaps the central leitmotif of transhumanism is the idea of enhancement by means of current technologies such as genetic engineering and information technology as well as future ones such as nanotechnology and artificial intelligence. The range of enhancement options includes the extension of the human lifespan, the elimination of disease, and most importantly the augmentation of human intellectual, physical, and affective capacities (Bostrom 2003b). Further down the line, enhancement may move beyond merely augmenting existing human capacities to the integration of biological and artificial systems. A key aspect of the enhancement strategy in this context is the possibility that future technologies may also allow us to choose which intellectual capacities and traits are incorporated in any future iterations of oneself and which are discarded. As such, enhancement raises questions that are central to the problem of personal identity: Is the person post-transformation the same as the person pre-transformation? Can enhancement transform not only the various traits that we associate with

persons, but also the consciousness that grounds a basic sense of self-awareness, which makes experiences appear as if occurring for someone?

Many proponents of human-AI integration fail to appreciate, however, as Schneider herself acknowledges, that "the enhanced being may not be you" (p. 7). In other words, the prospect of consciousness engineering is informed by a naïve view of intelligence as an attribute or capacity that can be controlled and harnessed without any significant alteration of the person or individual whose attribute it is. If anything, the prospect of enhancement reinforces a deeply ingrained sense of self, as exemplified by a persistent concern (some might say obsession) among techno-optimists with life-extension, 'mind-uploading', and the general idea of transferring a self-identical human consciousness from a biological brain to a computer. Indeed, for champions of the computational theory of mind such as Ray Kurzweil, if the mind is nothing but "the program running on the hardware of the brain" (Kurzweil 2005, p. 383), 'uploading' or "scanning the synaptic structure of a particular brain and then implementing the same computations in an electronic medium" (Bostrom 2003a) is a forgone conclusion. The transhumanist manifesto is replete with optimistic reassurances of over-coming biological senescence by creating backup copies of oneself to be perpetually uploaded and/or rebooted as new and ever perfected AI systems with lifespans potentially as vast as that of the universe itself become available (Vita-More 2020). If conscious minds are nothing over and above certain kinds of information patterns (as proposed, inter alia, by the Information Integration Theory (IIT) of consciousness),[4] then survival is a matter of preserving this pattern rather than the medium—organic or synthetic brains—that implements it.

Schneider takes issue with this patternist conception of personal identity, which entails the possibility of upgrading to ever new versions of oneself: from human to human 2.0 to a human merged with AI to just a computational configuration that is continuous with the various iterations of the original human consciousness pattern. Drawing on an illustration from the science fiction novel *Mindscan* by Robert Sawyer, in which mind scanning results in a reduplication of the person rather than the transfer of the original pattern to a new substrate, Schneider rightly identifies the limitations of patternism: the sense of psychological continuity associated with having a particular type of pattern "cannot be *sufficient* for personal identity" (p. 84) since duplicates can enjoy the same sense of psychological continuity despite occurring at a different place and time. Whether patternism can answer the reduplication problem (i.e., the problem that sameness of pattern is not sufficient for sameness of person) depends on whether patternism is merely *necessary* but not *sufficient* for personal identity. If spatiotemporal continuity is indispensable to personal identity, then patternism cannot satisfactorily answer the personal identity problem: making copies of your

mind cannot count as an enhancement of you if your mind "still carries on, and it is subject to the limitations of its substrate" (p. 88).

Can a modified version of patternism fine-tuned to accommodate spatiotemporal continuity do the job? Schneider thinks that a *modified patternism* is bound to fail as well, if 'uploading' does not guarantee survival for the person. Nor will a modified version of patternism that relies on replacement (of neurons with synthetic materials configured to perform the same function) rather than uploading fare any better. Although in this case spatiotemporal continuity is preserved, the composition of the substrate raises its own challenges, even as she admits uncertainty about the technological feasibility of this alternative thought experiment.

The trouble is that questions about the technological feasibility of these thought experiments cannot be answered without first addressing the problem of what the nature of a person is, a matter further complicated by anti-essentialist conceptions of the person, which are taken to be in keeping with scientifically informed accounts of human nature. So, do anti-essentialist conceptions of personal identity such as those rooted in psychological continuity bolster the transhumanist vision of a human-AI interface? And will this interface preserve enough of the attributes and capacities of the individual to address the dilemma of personal identity?

III. Are Persons Reducible?

First, let me clarify the dilemma itself. Because persons persist through time, hence exist for longer than a single moment in time, it is necessary to explain how they can do so. The inability to provide adequate answers to the persistence problem has led metaphysicians to postulate either the existence of enduring entities such as a soul or self or to attribute belief in personal identity to a persistent illusion about the unity of our conscious life. For metaphysicians sympathetic to a reductionist account of persons, yet weary of the ethical implications of illusionism, alternative solutions to the problem must explain how we can have continuity without sameness.

Consider Parfit's vastly influential theory of personal identity, which argues against the commonsensical, non-reductionist view of persons (Parfit [1984] 1987, pp. 214 ff.). According to the non-reductionist view, persons are distinct and discrete entities that exist over and above their bodies and psychological states. Their identity, then, is an irreducible, brute fact of existence, and cannot be explained or described in more basic terms. Whatever persons are, an account of their identity would have to employ person-level descriptive categories of experience.[5] One paradigmatic example for person in this non-reductive sense is the Cartesian Ego. The view that there are such entities as Cartesian Egos or souls is representative of a particular intuition about personal identity, according to which we assume that questions of the sort 'Will I survive the death of my body?' or 'Will I be

the same person if I were to be teleported elsewhere?' must have definitive answers. Regardless of whether or not we have answers to these questions at present, given their implications for personal identity, answers must in principle be available. There must be a way to settle these questions one way or another, perhaps on the basis of our very conception of what personal identity entails. What drives this intuition is the assumption that our identity must in some sense be *determinate*.

If we reject this intuition and allow for the possibility that Cartesian Egos do not exist, then we are in a sense compelled to accept the view of reductionism. One of the advantages of reductionism is that it offers new possibilities for reconceiving the problem of personal identity on both metaphysical and empirical grounds. Since the body is the seat of our physical, affective, and mental lives, we may conceive of persons as (1) bodies or as (2) entities that have bodies, thoughts, and emotions. The first conception can also be understood as an endorsement of one version of the *identity* view (persons just *are* bodies), while the second makes the case for the *ownership* view (persons are the sort of entities that *have* bodies, thoughts, and other kinds of experiences). Whereas the ownership or constitutive view of personal identity can be easily entertained, and may even fit classical conceptions of persons as property-possessors, the identity view of reductionism opens the door for something more radical: eliminative Reductionism, the best example of which is the mind/brain identity theory.

The motivations for the reductionist view in Parfit's case are well known and will not be repeated here: they include the simple and complex teleportation thought experiments, with the latter raising precisely the challenges that Schneider thinks a patternism unable to accommodate spatiotemporal continuity faces. They provide the theoretical framework of the psychological continuity thesis that informs many discussions in the metaphysics of personal identity. But Parfit also appeals to Buddhist reductionism, which articulates something close to a psychological criterion of personal identity. The problem is that Buddhist reductionism is not eliminativist about the underlying principle for personal identity, namely consciousness understood, inter alia, as the capacity for discerning the difference between self and other. The irreducibility of consciousness for Buddhism is not incompatible with a Cartesian view of personal identity as grounded in consciousness (Strawson 2023), even as the prevailing tendency has been to view it as closer to a sort of higher-level trope that grounds an impersonal account[6] of the unity of consciousness (Siderits 1997, 2015; Goodman 2004).

Buddhism is also host to a robust personalist school of thought, which argues that persons, although neither identical to nor different from their constituent bodily parts and mental processes, are nonetheless real. Like Parfit, Buddhist Personalists thought that neither a purely physical nor a purely psychological criterion will suffice for personal identity (Priestley

1999, pp. 81–82). Rather, persons on this account are defined primarily in terms of the subjective and phenomenal character of their conscious mental states. Parfit's eventual rejection of the impersonal description thesis (the thesis that says that we can provide an account of psychological continuity without reference to persons and their phenomenally conscious states) is motivated by the conception that people are not collections of things but primarily agents—"not thoughts and acts" but rather "thinkers and agents" (Parfit [1984] 1987, p. 223). It is as thinkers, specifically as conscious thinkers, that we conceive ourselves as persons, a view that recalls the Lockean thesis that personal identity extends as far as our "consciousness can be extended backwards to any past action or thought" (Locke [1689] 1975, 2.27.9, 335).

Elsewhere I have argued that Buddhist Personalism provides a closer analogue for Parfit's theory of personal identity than Buddhist Reductionism, with its stated mereological nihilist view that there are no such things as composite entities (Coseru 2020a). The psychological continuity criterion of personal identity is plausible only to the extent that I can conceive of myself in dependence upon a sufficiently similar pattern. But the conceivability principle concerns the epistemological, rather than the ontological, dimension of personal identity, which brings up an altogether different set of considerations, specifically about the relations that obtain between self-referential mental states (those that presuppose the notion of oneself as a subject) and self-consciousness. The epistemological dimension is framed by a different set of questions that pertain not to what awareness supervenes on but to its structure and specific properties—namely: What, in particular, accounts for a mental state becoming an instance of self-consciousness? Does self-consciousness require that a referential subject-to-object relation become present to itself as an object? If we can answer these questions, we can make progress in understanding the relation between self-referential mental states (i.e., *de se* states) and self-consciousness. And if we can get clarity about the nature and character of self-consciousness, we are in a better position to understand what it is that makes us persons.

For the Buddhist Personalist, reductive analysis is meant to capture not what persons are *made of*, but rather what human experience is *constituted as*: specifically, as a series of intentional and self-referential mental events. Consider the paradigmatic example of pain: as a sensation, pain is not reducible to the physical substrate, say a finger, in which it is instantiated (nor presumably to a mere physiological response). Rather, pain is constituted as a distinctly qualitative phenomenon whose intentional content cannot be dissociated from its subjective aspect. There is no such thing as generic or impersonal pain (understood strictly in terms of, say, the activation of Aδ- and C- fibers following an intense stimulation of nociceptors) apart from phenomenally foregrounded sensations of some kind: of burning, stinging, or throbbing.

When the Buddhist Personalists insist that a mere functional account of the aggregates will not suffice to explain why an action counts, say, as killing, they draw attention to the specificity and individuality of a given bundle of aggregates, and hence of its actions and consequences: "If the self were absolutely non-existent, then there could not be killing nor would the killer have killed anything. There would be nothing like theft and robbery. . . . [G]ood and bad would yield neither freedom nor bondage; even bondage would have no one bound. There would be neither the doer nor the deed nor any result thereof" (Venkataramanan 1953, p. 177). Indeed, understanding why something is categorized as killing and not simply as the rearranging of material elements presupposes a conception of intentional action that is unintelligible without reference to persons. Unlike clumps of clay arranged in such a way as to resemble human beings, living beings are characterized primarily in terms of their capacity for responsive and intentional action: they can both do things and have things done to them in a way unavailable to insentient objects. Persons, unlike other assemblages of parts made to resemble them in likeness and functionality, are not simply generic unities of aggregates, but agentive and self-disclosing wholes. Persons are what they are by virtue of the fact that their constitutive elements belong together: the heart, lungs, blood, and blood vessels work together as a system that we call the circulatory system; the brain, the spinal cord, and the nerve fibers work together as the nervous system; and the sense organs (sight, hearing, etc.) in concert with the nervous system and the body's motor controls work together as the sensorimotor system.

IV. Natural, Artificial, or Empty? Enhancement and the Personal Identity Dilemma

What does this organismic (or biological) conception of personal identity mean for the prospect of human enhancement by means of brain uploading or by brain chips designed to augment intelligence and fundamentally alter one's cognitive abilities? Is such artificially-enhanced intelligence and digital immortality feasible enough to warrant serious scrutiny? Is the prospect of Artificial General Intelligence (AGI) as imminent as techno-optimists claim?[7] Or should it be dismissed offhand as highly speculative and not based on any real understanding of the biological basis of human intelligence? If biological intelligence entails the capacity for self-generation, self-organization, and self-regulation, it is hard to imagine how such capacities could be replicated in systems whose integrity and functionality are technology-dependent (that is, dependent on industrial assembly and maintanance lines, power grids, and other operational relations that are external to the system).

Consider flying as an analogy for intelligent behavior. Birds fly. Do humans fly? They do, but only on such devices as airplanes. Do the planes and paragliders themselves fly? Yes, but they do not fly by flapping their

wings. So, just because we can metaphorically extend the sense of 'fly' from birds to planes and paragliders, and even develop an abstract theory of flying (based on the laws of aerodynamics), does not mean the flight of planes or gliders is a sensorimotor and cognitive process (the presence of a complex satellite and telecommunication network allowing for autopilot navigation notwithstanding). Likewise, metaphorically extending the sense of 'think' from humans to machines does not mean computational systems think in a way that is structurally and functionally independent of human input and participation. Without user prompts and a design architecture modeled on information retrieval and analysis, there is no machine intelligence.

Schneider's plea for *metaphysical humility* in the face of a seemingly conceivable near-future shopping trip to the Center for Mind Design suggests that she does take seriously a scenario in which, for a civilization that develops the requisite AI technology, the transition from biological to postbiological existence is possible. This scenario is informed by a position in the field of astrobiology, which advances the claim that "members of the most intelligent alien civilizations will be superintelligent AIs" (p. 99). The position is motivated by three highly speculative observations: (1) that it only takes a few hundred years for a civilization to go from biological to postbiological; (2) that the existence of much older alien civilizations cannot be ruled out; and (3) that members of these much older alien civilization are likely to be synthetic rather than biology-based.

Schneider does anticipate various objections to these observations. Nonetheless she thinks that when combined, exponential advancement in chip technology, an understanding of the limitations of the human brain, and the vastness of the time and space scale of the observable universe warrant that we take these observations seriously. This is a perfectly reasonable stance. If alien civilizations exist that are advanced to greater orders of magnitude, then it is plausible that they would have answered many of the problems in physics, cosmology, and astrobiology that our science is yet to encounter, let alone those that currently elude us. But they would have had to answer them *within the bounds of what is naturally or physically possible*. And the idea of synthetic or non-biological consciousness that underscores the case for superintelligent AIs finds little support in current knowledge of the nature of physical and biological systems. Indeed, the very notion of 'Artificial Intelligence' is contested, even by scientists at the forefront of the field. As Jaron Lanier recently argued in a provocatively titled essay ("There is no AI"), for the tech culture, depictions of AI such as one finds in sci-fi blockbuster movies like "The Terminator" and "The Matrix" serve as a sort of "religious mythology." Rather than giving in to quasi-religious visions of an accelerating tech revolution that spells doomsday for humankind, Lanier urges that we view these technologies for what they actually are: tools, not creatures. The real peril lies in mythologizing

the technology, hence his contention that "we can work better under the assumption that there is no AI, for doing so at least makes it more likely that we will "start managing our new technology intelligently" (Lanier 2023).

Despite its critical stance on AI, *Artificial You* reflects much of this mythologizing ethos. Schneider provides compelling, if highly speculative, scenarios of alien thinkers pondering the intractable issues of personal identity generated by cognitive enhancement in a manner not unlike that pursued by philosophers of mind today. We are led to believe there is a good chance that members of such advanced civilizations may well have embraced radical enhancement despite the risk of death, or simply "because they mistakenly believed they found clever solutions to the philosophical puzzles of personal identity" (p. 102). Or they may have pondered the risks but concluded, "based on reflections of alien philosophers who have views akin to the Buddha or Parfit, that there is no real survival anyway" (ibid.). Lacking belief in an enduring self, denizens of such alien civilizations may have opted to upload their minds anyway, ushering in an era of superintelligent AI.

Does the Buddhist no-self view give credence to the sort of trans-humanist scenario in which, despite the lack of a persistent self, personal identity can nonetheless by transferring "the informational structure of the brain from tissue to silicon chips" be preserved (ibid.)? As I noted above, Buddhist reductionism about selves is not incompatible with the psychological continuity thesis that preserves a conception of persons as self-conscious agents. But Buddhism is also host to a strongly illusionist stance. For Mahāyāna Buddhism in particular, this illusionist stance underpins a vast cosmology of innumerable world systems endowed with reality by the yogic power of a buddha or buddhas for the purpose of leading sentient beings to salvation. It is not just selves that are illusory on this account, but the universe in its entirety. Indeed, even the Buddhist path itself, and the process of bringing sentient beings to awakening, is sometimes likened to a magical show. Far from endorsing a view of the mind as implemented by discernible brain patterns that are susceptible to technological replication and enhancement, the radical insight of this Buddhist intellectual tradition is that the mind itself is ungraspable, that it does not withstand analysis. On this scenario the prospect of radical enhancement, and hence of superintelligent AI, is just another manifestation of the failure to grasp the essencelessness of phenomena. It would seem that a no-self view of personal identity does not make things any easier for the champion of radical enhancement, whether that involves brain uploading, merging of humans and AI, or the emergence of biologically inspired superintelligent AIs.

V. Conclusion: The Biological Challenge to AI

Let me conclude with what I think might be the biggest challenge to the sort of enhancement that entertains the prospect of synthetic consciousness: if

persons are processes rather than things, then they cannot be cognitively enhanced by augmenting, replacing or transferring their functioning parts (e.g., brain cells). On a process-ontological account, the biological and cognitive processes that function to keep an individual alive and aware do not exist in the sort of discreet isolation that is assumed to be the case for things. Things can be taken apart, and have their parts rearranged and modified in ways that processes cannot. This does not mean that as a particular kind of organism, humans cannot be differentiated from the processes that constitute them. But if the lessons of systems biology are any indication, this differentiation has no clear demarcation line. As distinct dynamic unities, the complex processes that are constitutive of human existence and experience depend on a constant interaction with that from which they differ: the processes in the environment that sustain them, and which in turn are impacted by them (Maturana and Varela 1980; Moreno and Mossio 2015). And the determining factor in the way an organism persists in its environment is *metabolism* rather than some kind of self-ascribed identity.

Adopting a biologically informed process-ontology perspective on personal identity means recognizing that organic identity is different in kind: it is an identity of *form* rather than of *matter*. What persists is the specific organizational form of the processes by which the organism continues to exist, not its constitutive elements. As organisms, persons endure by undergoing constant change on a fundamentally cellular level, including the change in the neural pathways that realize the cognitive processes of thinking and expressing these very thoughts right now (see Jonas 1966, and discussion in Meincke 2018). Unlike numerical identity, in which different person stages can be said to belong to or be constitutive of the same person and type identity in which some classes of mental states are taken to be identical with some classes of brain states, biological identity is a kind of functional or processual identity. It is precisely this constant metabolic exchange and energetic coupling with its environment that defines an organism's survival and identity. In short, for organisms, identity is processual in the sense that those processes (e.g., blood circulation, cell division, digestion, synaptic firing) that ground it serve as a specific dynamic manifold that persists through time.

The notion that organisms persist by maintaining their form while undergoing constant elemental change is indeed suggestive of the patternist conception of personal identity. As such, it may be argued that it is in keeping with, rather than counter to, the principles that underpin the psychological continuity thesis of personal identity, according to which identity is a function of the persistence of enough of the same brain (on the mind-brain identity view). But AI systems that implement an individual's brain pattern are independent of whatever it is that powers or sustains their functionality (e.g., electricity or an equivalent source of energy). Microchips

or rather their components (transistors, resistors, capacitors, diodes, etc.) do not respond and adapt on their own, either individually or as part of an assembly, in response to electric current flow.[8] Certainly, the logic board ensures that each hardware component, which is designed to perform a specific function, coordinates the complexly choreographed electrical connections that render inputs into the desired outputs. But whether it is translating mechanical pressure or air vibrations into words and images on a screen, executing complex mathematical calculations, or mapping various features of the environment for various purposes, the processes that translate commands into tangible outputs are not adaptive, not even for systems that implement a neural network architecture.

Machine neural networks, for instance, cannot adapt to changing stimuli by changing, on their own, the weights or the activation functions (that can only be done during a training phase). And while they can rewrite and improve their source code overtime, and even optimize firmware instructions, artificial neural networks are constrained by the hardware that implement their functionality: they cannot execute routines for which they lack hardware (e.g., sufficiently dense memory chips), bandwidth (for synchronizing the data with hard drive storage) or enough power. Furthermore, because the design architecture of neural networks is centered on learning, and because training neural networks can deliver unexpected results, engineers must add specific features that allow the system to be managed. An unmanaged system, as successful 'jailbreak' attacks of ChatGPT and other generative AI systems demonstrate, is not only unreliable (prone to artificial hallucination) but also problematic from an ethical and security standpoint as red teaming tests demonstrate (Greshake et al. 2023; Wei et al. 2023). Organisms, by contrast, persist by continuously rebuilding and maintaining themselves through an exchange of matter and energy with the environment, processes that reflect a fundamental concern with survival. Schneider recognizes that "intelligent biological life tends to be primarily concerned with its own survival and reproduction" (p. 114). As a result, any biologically inspired superintelligent systems who have inherited this concern are likewise likely to make survival their primary goal even as they pursue various forms of enhancement. But she downplays the challenge posed by attempting to reverse engineer an organism's self-generating, self-organizing, and self-regulating capacities.

Any AI system whose design and architecture depend on elements and components that lack the autopoetic capacity of organic processes is bound to lack the adaptive intelligence of living systems, raising questions about whether such capacity could even be engineered,[9] let alone be engineered out of any putative biologically inspired superintelligent aliens (BISAs) such that survival may no longer be a concern. Why advanced civilizations should be precisely those that dared to forgo rather than strengthen their

concern with survival in the relentless pursuit of enhancement is not at all clear. Sure, the transhumanist ethos is all about enhancement, leaving the question of survival to ethicists who worry about its impact on individual autonomy and the value of life. And yet, Schneider is right to draw attention to the Parfitian attitude toward the conundrum of personal identity: after all, personal identity might not be what really matters for prudential concern about our future. Rather, what matters is a right combination of relations of psychological continuity and connectedness to meet the requirement for similarity of psychological makeup. Regardless of whether this attitude is strengthened by Buddhist claims about personal identity, it is still the case that, as Schneider in the end concedes, creating consciousness in a different, non-biological, substrate "may not be compatible with the laws of physics" (p. 149).

The specter of artificial minds cast by cutting-edge computers running the neural network algorithms that power AI systems such as ChatGPT and Google Bard may be alluring. But in keeping with the test devised by Portia's father to determine the suitability of her suitors, we are well advised to choose our criteria for signs of intelligence, let alone consciousness, wisely, lest we grant it to some simulacra of intelligence with near endless capacities for deception, in short to a stochastic parrot (Bender et al. 2021, 617)—a system capable of random generation of linguistic patterns drawn from its vast training data, based on probabilistic information about syntactic rules, yet lacking any reference to meaning.

Notes

1 – For a review of the global impact of artificial light on marine ecosystems, see Marangoni et al. (2022).

2 – The metabolic impacts of disruption to the circadian system and sleep due to artificial lighting include, inter alia, gut microbiota dysregulation, immune system deregulation, pancreatic function and adipose tissue impairment, reduced satiety, etc. (Potter et al. 2016; Park et al. 2019).

3 – This accelerating pace of development in AI research does raise real concerns about its broader and possibly disruptive social impact, as recently noted in an open letter calling for pausing giant AI experiments. Indeed, as of August 2023, the "Pause Giant AI Experiments: An Open Letter" posted to the Future of Life Institute (https://futureoflife.org/open-letter/pause-giant-ai-experiments/), has over 33,000 signatories, including many prominent tech leaders such as Elon Musk, Apple co-founder Steve Wozniak, and Pinterest co-founder Evan Sharp.

4 – It is noteworthy that rival theories for the neural correlates of consciousness such as the IIT and Global Neuronal Workspace Theory (GNWT) have yet to deliver experimental results that would settle beyond any reasonable doubt questions about the neural markers of conscious perception, as recently conceded by, inter alia, Christoph Koch (one of the proponents of the IIT theory) at the 26th annual meeting of the Association for the Scientific Study of Consciousness in New York City, June 2023 (Finkel 2023).

5 – One way to understand the difference between the non-reductionist and the reductionist views of personal identity is along the simple/complex divide: the non-reductionist favors the simple, soul, or Cartesian Ego view, whereas the reductionist prefers the complex view that entails relations among physical and psychological states. Holding a soul view, of course, does not necessarily amount to holding a brute-fact view, although in the absence of non-circular criteria for personal identity (of the sort required by the complex view) it is hard to tell them apart. What motivates recent defenders of the simple view (e.g., Baker 2013, Lowe 2013, Nida-Rümelin 2013, Swinburne 2013) is not a commitment to a Cartesian Ego, but rather the notion that a specific, perhaps non-conceptual and pre-reflective, type of self-awareness seems indispensable to framing any account of personal identity. See Coseru 2020 for a detailed discussion.

6 – Whether such an impersonal account of the unity of conscious experience is intelligible, particularly if intentionality is itself an ineliminable dimension of consciousness, is debatable. Buddhists have historically faced a barrage of arguments against the intelligibility of the impersonal description thesis (including from within their ranks), and it is not clear that a master argument for the thesis is available that eschews doctrinal commitment to the no-self view (Arnold 2012, 113f; Coseru 2020b: 128f).

7 – Within nine weeks of the release of OpenAI's ChatGPT to the public in November 2022, the prediction site Metaculus, which tracks forecasters' guesses as to when we should expect for an AGI system to arrive, brought forward the estimated date from 2050 to 2026. By August 2023 it had slipped back to April 2032, still barely a decade away.

8 – Biochip technology, such as microfluidic chips or organ-on-a-chip (OCC) platforms do hold the promise of creating in-vitro human-derived neuronal networks that can respond to electrical stimulation (Azizipour et. al. 2020; Zhao et al. 2020). But while these biochip platforms may help advance our understanding of the complex functionality of brain tissue, so far, they are "inherently unable to

replicate the three-dimension (3D) environmental complexity of the brain" (Muzzi et. al. 2023, 2).

9 – Biochip technology represents an important step in tackling this challenge. A biochip is a miniature system that mimics the in vivo physiological environment of parts of the body or organs, typically by regulating the distribution of cells, gradient of biochemical molecules, and various mechanical stimuli (Chung et al. 2005; Huh et al. 2010; Zamprogno et al. 2021; Zhao, Demirci, Y. Chen, P. Chen 2020). But while biochips-driven micro-devices such as organ-on-a-chip (OOC) platforms (e.g., neural and retinal implants for blind patients) can mimic tissue- and organ-level physiology (e.g., axonal growth, decrease of firing activity), the devices cannot rebuild and maintain themselves.

References

Arnold, Dan. 2012. *Brains, Buddhas, and Believing: The Problem of Intentionality in Classical Buddhist and Cognitive-Scientific Philosophy of Mind*. New York: Columbia University Press.

Azizipour, Neda, Rahi Avazpour, Derek H. Rosenzweig, Mohamad Sawan, and Abdellah Ajji. 2020. "Evolution of Biochip Technology: A Review from Lab-on-a-Chip to Organ-on-a-Chip." *Micromachines* (Basel) 11, no. 6 (June 18): 599. doi: 10.3390/mi11060599.

Baker, Lynne Rudder. 2013. "Personal Identity: A Not-so-simple view." In *Personal Identity: Complex or Simple*?, edited by Georg Gasser and Matthias Stefan, pp. 179–191. Cambridge: Cambridge University Press.

Bender, E. M., T. Gebru, A. McMillan-Major, and M. Mitchell. 2021. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 610–623. New York, NY, USA: Association for Computing Machinery.

Bostrom, Nick. 2003a. "Transhumanist FAQ: A General Introduction," version 2.1. World Transhumanist Association. https://nickbostrom.com/views/transhumanist.pdf.

———. 2003b. "Human Genetic Enhancements: A Transhumanist Perspective." *Journal of Value Inquiry* 37, no. 4:493–506.

———. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Chung, Bong G., L. A. Flanagan, Seog W. Rhee, Philip H. Schwartz, Abraham P. Lee, Edwin S. Monuki, and Noo L. Jeon. 2005. "Human

Neural Stem Cell Growth and Differentiation in a Gradient-Generating Microfluidic Device." *Lab Chip* 5:401–406.

Coseru, Christian. 2020a. "Reasons and Conscious Persons." In *Derek Parfit's Reasons and Persons: An Introduction and Critical Inquiry*, edited by A. Sauchelli, pp. 160–186. London: Routledge.

———. 2020b. "Whose Consciousness: Reflexivity and the Problem of Self-Knowledge." In *Buddhist Philosophy of Consciousness: Tradition and Dialogue*, edited by Mark Siderits, Ching Keng, and John Spackman, pp. 121–153. Leiden: Brill.

Finkel, Elizabeth. 2023. "'Adversarial' Search for Neural Basis of Consciousness Yields First Results." *Science*. doi: 10.1126/science.adj3877.

Goodman, Charles. 2004. "The Treasury of Metaphysics and the Physical World." *The Philosophical Quarterly* 54, no. 216:389–401.

Greshake, Kai, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. 2023. "Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection." *arXiv*. https://doi.org/10.48550/arXiv.2302.12173.

Huh, D., B. D. Matthews, A. Mammoto, M. Montoya-Zavala, H. Y. Hsin, and D. E. Ingber. 2010. "Reconstituting Organ-Level Lung Functions on a Chip." *Science* 328:1662–1668.

Jonas, Hans. 1966. *The Phenomenon of Life: Toward a Philosophical Biology*. Evanston: Northwestern University Press.

Kurzweil, Ray. 2005. *The Singularity is Near: When Humans Transcend Biology*. New York: Viking.

Lanier, Jaron. 2023. "There Is No AI." *The New Yorker*, April 20, 2023.

Locke, John. (1689) 1975. *An Essay Concerning Human Understanding*. Edited with an Introduction by Peter H. Nidditch. Oxford: Oxford University Press.

Lowe, E. J. 2013. "The Probable Simplicity of Personal Identity." In *Personal Identity: Complex or Simple?*, edited by Georg Gasser and Matthias Stefan, pp. 137–155. Cambridge: Cambridge University Press.

Marangoni, Laura F. B., Thomas Davies, Tim Smyth, Airam Rodriguez, Mark Hamann, Cristian Duarte, Kellie Pendoley, Jørgen Berge, Elena Maggi, and Oren Levy. 2022. "Impacts of Artificial Light at Night in Marine Ecosystems—A Review." *Global Change Biology* 28, no. 18:5346–5367.

Maturana, Humbert R., and Francisco J. Varela. 1980. *Autopoiesis and Cognition: The Realization of the Living*. Dordrecht: Springer.

Meincke, Anne Sophie. 2018. "Persons as Biological Processes: A Bio-Processual Way Out of the Personal Identity Dilemma." In *Everything Flows: Toward a Processual Philosophy of Biology*, edited by Daniel J. Nicholson and John Dupré, pp. 357–378. Oxford: Oxford University Press.

Moreno, Alvaro, and Matteo Mossio. 2015. *Biological Autonomy: A Philosophical and Theoretical Enquiry*. Dordrecht: Springer.

Muzzi, Lorenzo, Donatella Di Lisa, Matteo Falappa, Sara Pepe, Alessandro Maccione, Laura Pastorino, Sergio Martinoia, and Monica Frega. 2023. "Human-Derived Cortical Neurospheroids Coupled to Passive, High-Density and 3D MEAs: A Valid Platform for Functional Tests." *Bioengineering (Basel)* 10, no. 4:449. https://doi.org/10.3390/bioengineering10040449.

Nida-Rümelin, Martine. 2013. "The Non-Descriptive Individual Nature of Conscious Beings." In *Personal Identity: Complex or Simple?*, edited by Georg Gasser and Matthias Stefan, pp. 157–176. Cambridge: Cambridge University Press.

Parfit, Derek. (1984) 1987. *Reasons and Persons*. Oxford: Clarendon Press.

Park, Yong-Moon Mark, Alexandra J. White, Chandra L. Jackson, Clarice R. Weinberg, and Dale P. Sandler. 2019. "Association of Exposure to Artificial Light at Night While Sleeping with Risk of Obesity in Women." JAMA Internal Medicine 179, no. 8:1061–1071.

Potter, Gregory D. M., Debra J. Skene, Josephine Arendt, Janet E. Cade, Peter J. Grant, and Laura J. Hardie. 2016. "Circadian Rhythm and Sleep Disruption: Causes, Metabolic Consequences, and Countermeasures." *Endocrine Reviews* 37, no. 6:584–608.

Priestley, Leonard C.D.C. 1999. *Pudgalavāda Buddhism: The Reality of the Indeterminate Self*. Toronto: Centre for South Asian Studies, University of Toronto.

Schneider, Susan. 2019. *Artificial You: AI and the Future of Your Mind*. Princeton: Princeton University Press.

Siderits, Mark. 1997. "Buddhist Reductionism." *Philosophy East and West* 47, no. 4:455–478.

———. 2015. *Personal Identity and Buddhist Philosophy: Empty Persons*. 2nd ed. Aldershot: Ashgate.

Strawson, Galen. 2023. "Descartes and the Buddha—A *Rapprochement*?" In *Reasons and Empty Persons: Mind, Metaphysics, and Morality: Essays in Honor of Mark Siderits*, edited by Christian Coseru, pp. 63–86. Cham: Springer.

Swinburne, Richard. 2013. "How to Determine which is the True Theory of Personal Identity." In *Personal Identity: Complex or Simple?*, edited by Georg Gasser and Matthias Stefan, pp. 105–122. Cambridge: Cambridge University Press.

Venkataramanan, K[rishniah]. 1953. *Sāṃmitīyanikāya Śāstra*. *Visva-Bharati Annals* 5:155–243.

Vita-More, Natasha. 2020. "The Transhumanist Manifesto," vol. 4. https://www.humanityplus.org/the transhumanist-manifesto.

Wei, Alexander, Nika Haghtalab, and Jacob Steinhardt. 2023. "Jailbroken: How Does LLM Safety Training Fail?" *arXiv*. https://doi.org/10.48550/arXiv.2307.02483.

Zhao Y[anan], Demirci U[tkan], Chen Y[un], and Chen P[u]. 2020. "Multi-scale Brain Research on a Microfluidic Chip.' *Lab Chip* 20, no. 9:1531–1543.