religions

Article

Souls and Selves: Querying an AI Self with a View to Human Selves and Consciousness

Andrew Oberg

Faculty of Humanities, University of Kochi, Kochi City 780-8515, Japan; oberg@cc.u-kochi.ac.jp

Abstract: The question of self-aware artificial intelligence may turn on the question of the human self. To explore some of the possibilities in play we start from an assumption that the self is often pre-analytically and by default conceptually viewed along lines that have likely been based on or from the kind of Abrahamic faith notion as expressed by a "true essence" (although not necessarily a static one), such as is given in the often vaguely used "soul". Yet, we contend that the self is separately definable, and in relatively narrow terms; if so, of what could the self be composed? We begin with a brief review of the descriptions of the soul as expressed by some sample scriptural references taken from these religious lineages, and then transition to attempt a self-concept in psychological and cognitive terms that necessarily differentiates and delimits it from the ambiguous word "soul". From these efforts too will emerge the type of elements that are needed for a self to be present, allowing us to think of the self in an artificial intelligence (AI) context. If AI might have a self, could it be substantively close to a human's? Would an "en-selved" AI be achievable? I will argue that there are reasons to think so, but that everything hinges on how we understand consciousness, and hence ruminating on that area—and the possibility or lack thereof in extension to non-organic devices—will comprise our summative consideration of the pertinent theoretical aspects. Finally, the practical will need to be briefly addressed, and for this, some of the questions that would have to be asked regarding what it might mean ethically to relate to AI if an "artificial self" could indeed arise will be raised but not answered. To think fairly about artificial intelligence without anthropomorphizing it we need to better understand our own selves and our own minds. This paper will attempt to analyze the self within these bounds.

Keywords: artificial intelligence; consciousness; identity; mind; personhood; self; soul

Citation: Oberg, Andrew. 2023. Souls and Selves: Querying an AI Self with a View to Human Selves and Consciousness. *Religions* 14: 75. https://doi.org/10.3390/rel14010075

Academic Editor: Joseph Rivera

Received: 8 December 2022
Accepted: 27 December 2022
Published: 5 January 2023

1. Some Background: Received Abrahamic Scriptural Ideas on the Soul to Situate Our Thinking for an Exploration of the Self

While the future of artificial intelligence remains uncertain, advances in the field are continuing at such a pace that new and previously unforeseen potentials have become questions of concern. If we do decide to build systems that in some ways are able to relate outwardly and—importantly—inwardly like ourselves, then we must also give concern to what we may be willing to grant them vis à vis the claims we make for individuals within our societies; yet here is the crux of the matter, "like us": What kind of selves are we (or, depending upon the interpretative nuance desired: What kind of selves do we have)? What is the self as something that is yours, mine, *its:* that is, an artificial intelligence's self, or at least the *possibility* of such; this is a point to reflect on while an "artificial self" remains a "maybe", such that if/when the shift is made to a definitive reply (positively or negatively) we will be able to more fruitfully deal with it. For notional structuring let us therefore start with a backwards glance to the idea of the soul as it is intimated (but not explained in much detail) in the Abrahamic monotheistic scriptural lineage, religious sources from which many of us draw our intuitive judgments on what a self entails. I should note that we will limit ourselves to scripture, and that our purpose here is absolutely not a full exploration of how soul has been thought in Jewish, Christian, or Islamic traditions; rather we only need to

establish the broad boundaries of this idea as it appears in these scriptural writings in order to establish the often presumptive approaches to self that are hidden in our minds when we first think the query. Due to these buried influences, all too often we erroneously equate soul with self, and in common vernacular the latter can clearly be seen to be historically and conceptually derivative from the former along such conceptual lines as one's "true essence" or "inner person", et cetera. Therefore, we have to first show what we do not mean by self to clear the way for what we do mean. As a final introductory comment, I add that on my reading soul as it is used in scripture would necessarily be beyond self in that for it transcendence seems a must, whereas—as we shall come to recognize—a self is fully determinative in purely "earthly" terms, and hence its applicability to AI. The reader is moreover forewarned that the below references will not provide deep investigative rigor to the idea of a soul since the ancients' objectives were naturally elsewhere, but such will be informative and with it finished we will then work out the details of the self in the next two sections without any need to return to considerations of soul.

The clearest view we can garner of what a soul might "consist" of, that is, what is meant, implied, and/or indicated by this word substantively while also generally immaterially (however, see the Apostle Paul's remarks in the portion on Christianity below), is from received commentary upon the afterlife, since the reality of the "next life" was both presumed and stated as the ultimate destination of the soul in these works: of that which occurs to the person's "true essence" that continues in existence after the individual's physical form has ceased to function. Colloquially: When this meat sack stops, what happens to the "me" that keeps going?[1] As the oldest of the three Abrahamic lineages we will start our journey in Judaism, which textually at least did not move beyond the shadowy and ill-described location marker *Sheol* until after its canon had been closed, but which then did in later centuries arrive at a more or less Heaven picture not unlike what we have in Christianity and Islam (e.g., a place of nearness to the Divine and to one's forebears, a locale of respite and rest; the details thereafter of course differ by group and subgroup). By way of example, some verses on *Sheol* taken from the Tanakh should suffice, supplemented by a rather humorous rabbinic remark on Paradise which admittedly is not from scripture and may thus be considered a bonus. (For clarification at the outset: In the below we restrict ourselves to illustrative samples on Heaven and do not include those for Hell, although both Christianity and Islam contain warnings about postmortem punishments and distance from the Divine).

To begin, Genesis 37.35b has Jacob proclaim after hearing (falsely, as it turns out) that his favorite son Joseph has been killed: "'No, I will go down mourning to my son in Sheol.' Thus his father bewailed him". Or again, a pair of verses from the Psalms, 6.6 and 18.6:[2] "For there is no praise of You among the dead; /in Sheol, who can acclaim You?" and "ropes of Sheol encircled me; /snares of Death confronted me". Or finally Isaiah 38.18: "For it is not Sheol that praises You, /Not [the Land of] Death that extols You; /Nor do they who descend into the Pit /Hope for Your grace". (Tanakh: The Holy Scriptures: The New JPS Translation according to the Traditional Hebrew Text 1985, pp. 60, 1113, 1123, and 697, respectively) What we have here, then, is "Sheol" as the seemingly common and sole postmortem point of arrival for everyone,[3] and from which nothing can be done nor out of which is there any escape. This is naturally a rather depressing view, and therefore we ought not be too surprised that such an idea was later transformed; as evidence of this makeover, witness the following much more recent remark wherein "Sheol" has become "Heaven", and where despite the individual's being called to account (which appears necessary upon entry), on the whole things seem much more engaging compared with the sit-around-for-no-purpose of *Sheol*: "When I get to Heaven, they'll ask me, why didn't you learn more Torah? And I'll tell them that I wasn't bright enough. Then they'll ask me, why didn't you do more kind deeds for others? And I'll tell them that I was physically weak. Then they'll ask me, why didn't you give more to charity? And I'll tell them that I didn't have enough money for that. And then they'll ask me: If you were so stupid, weak, and poor, why were you so arrogant? And for that, I won't have an answer". (Rabbi Rafael of

Barshad (1751–1827), quoted in Morinis 2010, p. 63) Thus, in summary, the soul as given above is basically the same as a person's "me", but not entirely so, and most certainly not in an equal state of mortality.

In Christian scripture we can find more direct attempts at a soul definition of sorts, although again everything is conducted in terms of an afterlife setting and the picture of a final otherworld dwelling. The writer who came to be known as the Apostle Paul attempts to make some conceptual headway in his letters to the group of believers in Corinth.[4] In 1 Corinthians 15.42-44 and 53-54a he argues: "(42) So it is with the resurrection of the dead. What is sown is perishable, what is raised is imperishable. (43) It is sown in dishonor, it is raised in glory. It is sown in weakness, it is raised in power. (44) It is sown a physical body, it is raised a spiritual body. If there is a physical body, there is also a spiritual body . . . (53) For this perishable body must put on imperishability, and this mortal body must put on immortality. (54a) When this perishable body puts on imperishability, and this mortal body puts on immortality, then the saying that is written will be fulfilled: 'Death has been swallowed up in victory'". This "second body", or "heavenly body", or "spiritual body", must again, as with Rabbi Rafael's description (and let us not forget that Paul was a Jew and that Christianity was almost indistinguishably Jewish for many of its early years), answer for its earthly conduct; only this time the verdict will be delivered by Jesus: Paul adds in 2 Corinthians 5:10: "For all of us must appear before the judgment seat of Christ, so that each may receive recompense for what has been done in the body, whether good or evil". It is not entirely transparent what Paul is expounding in the preceding verses, but the essence seems to be of a second actual body of some sort, something that may not include corporeality but which does point to a particularism that goes further than how we might think a ghostly identity would; we therefore compare a summative remark put into the mouth of Jesus by the person who composed the Gospel of Matthew (written some decades after Paul's epistles), from 25.46: "'And these [i.e., the non-righteous] will go away into eternal punishment, but the righteous into eternal life'" (The Go-Anywhere Thinline Bible with the Apocrypha, New Revised Standard Version 2010, pp. 154, 158, and 27, respectively[5]). Whatever the components of a "spiritual body" as Paul conceived it may be, the image which emerges is of an equation between it and soul as we know the word—perhaps even Paul himself had mentally more or less notionally bound these terms together when he thought and wrote—and this is reinforced by the same motifs of needing to answer for one's behavior during mortal life and a permanent abode after physical death, this time in clearly glorious or disciplinary conditions.

Islam, for its part, avoids confusion by simply focusing on a presentation of soul as one's individual self which after death faces either reward or punishment;[6] and moreover it frames our everyday being itself in terms of this structure. As enlightening examples here are two verses from the Qur'an's sixth surah and one from the fifth: To begin is 6.12: "Say, 'To whom belongs all that is in the heavens and earth?' Say, 'To God. He has taken it upon Himself to be merciful. He will certainly gather you on the Day of Resurrection, which is beyond all doubt. Those who have lost their souls will not believe.'" Further explaining the Day of Judgment, 5.119 reads: "God will say, 'This is a Day when the truthful will benefit from their truthfulness. They will have Gardens graced with flowing streams, there to remain for ever. God is pleased with them and they with Him: that is the supreme triumph". Then, returning to the sixth surah, we find this verse on the relative importance of the "two lives" (so to phrase things), 6.32: "The life of this world is nothing but a game and a distraction; the Home in the Hereafter is best for those who are aware of God. Why will you [people] not understand?" (The Qur'an 2004, pp. 81, 79, and 82, respectively) Many other verses could be quoted here (I am tempted to also include 41.30-32 [p. 309 in *ibid.*]), but for reasons of succinctness we shall restrict ourselves to the foregoing. I believe that the straightforwardness and clarity of this message is as easily grasped as it is direct ("soul" = "me"; "this life" to "the next life"), and with it we can now pass from the background and typically unnoticed ideational stance (our hidden ideas) as have been communicated by the Abrahamic line of faiths to the beginnings of our psychological and

cognitive analyses: those areas which we will firstly need to lay bare before applying the same splicing onto (the potential or lack thereof) the self with regard to artificial intelligence. The self, we deem, is not soul; but when we think the self we tend to begin from an image of soul-like essential and full identitarianism, and this partly (mostly?) because of how soul is given in these worldview-establishing faiths for those familiar with them. Thus we must take caution, and herein lays the reason I wanted to append these opening thoughts to what will henceforth be a far more technically and scientifically oriented study.[7] The reader is also reminded that we shall not return to soul in what follows, but rather focus on the self.

Let us then start our reduction from the soul to a self which is more narrowly defined by recalling that for a conceptually maintained "me" to be held by a creature/being/existence (of some kind) there will need to be consciousness enough for self-regard, and that definitionally and necessarily consciousness is much more than intelligence (a calculator has intelligence but not consciousness, a computer has self-reflection in the form of internal monitoring but not self-regard in the sense of identity); thus foremostly some important terminological distinctions will need to be made. Towards this, in the next section we shall start with the human self, and I will base my analysis of this self as a psychological posit on a study originally done by Kristján Kristjánsson, to which I have added (what I take to be) important extensions; using his work as foundation, I elaborate and develop the self-model further into issues of mind.[8] (Kristjánsson 2010; Oberg 2020) This unfolding will require us to examine the topic from three interrelated and constantly interacting levels: the self (as core), personal identity, and whole person. From that investigation it will emerge that consciousness—and relatedly a two-tiered mental model—are paramount, and thus we will also attempt to determine a source for consciousness within the human framework (and I note here at the outset that as intriguing as ideas like panpsychism are we will not entertain any, but for a thorough dissection the reader is again directed to (Oberg 2020)).

From that initially concluding position we ask: Based on this, would an artificial self be constructible? Such is not a question of desirability (herein we take a neutral stance on that), only achievability. Presuming, perhaps incorrectly but for the sake of argument, that this does become feasible in engineering and design at some point (and maybe soon, as we will consider), what "rights" or "privileges" might apply to en-selved and self-aware artificial intelligences? If any were shown to be relevant, those too would call for careful differentiation since the issues might not be of a strictly property/ownership legality nature,[9] but the responses to them need to be nuanced and regulated. Placing that to the side for the moment, however, here are the queries that will guide our path: (1) What is the human self?; (2) What/whence is consciousness?; and (3) Might artificial intelligences be en-selved, and if so then what concerns relate? We will deal with each of these in turn, although we acknowledge that final and lasting answers will not be possible: ours is a preliminary study, the beginning of a wondering which is overdue for a widening and a deepening, for an intensive public debate, for forethought, for preparations, for "maybes". We make a start.

2. What Is the Human Self?

Kristjánsson's account, as mentioned, will provide our starting point for investigating the self as distinct from how soul (and perhaps our unexamined ideas on the self) tends to be thought. This particular self-theory appears to be essentially educational in nature,[10] and it is grounded in a mental framework which is singly structured, one where rationality underpins emotion and precedes action. (Therefore, the potential for self-engagement/training on this view is: X decides to become such and such a person, and hence X does thus and so over and over in order to make herself into that). The attempted praxes involve engaging one's reason in the production of one's feelings, and therefrom one's behavior commences in better alignment with one's rational aims. As for the core of the self, Kristjánsson places it on the emotional plane and claims that it is composed of three sets: (1) Self-constituting emotions: "core commitments, traits, aspirations, or ideals"; (2) Self-comparative emotions:

those which take the self as an "indirect object" or "reference point" for "comparison with a baseline of expectations"; and (3) Self-conscious emotions: those within the self that such are about, that take the self as "their direct attentional and intentional object" (Kristjánsson 2010, pp. 75–77). While on the whole I find this preferable to many other self-theories,[11] there are two major problems with this structure; one is easily rectified, but the other will require a more detailed undertaking. Doing so will, however, aid us in establishing our own alternative, and thus the efforts required will quickly be rewarded.

We will consider the easier problem first. This is that Kristjánsson's initial "emotional set" is not in point of fact emotional: these facets rather are ideational and identitarian, and although they do line up with Kristjánsson's focus on rationally choosing one's self (oneself), and thereafter working to eventuate it, a large amount of what makes you "you" and me "me" is missing from this "self-constituting" group: namely, those elements which Edmund Husserl referred to as one's "lifeworld" and which Martin Heidegger delicately adjusted and re-named simply "world" (Husserl 1999, 2001; Smith 2013; Heidegger 2010). Examples of such pertain to what may also be labeled one's "enworldedness", or situational and contextual embeddedness, and include the vast array of influential and personally molding so-called "facts of life" into which we are born: these are items such as DNA, historical epoch, geographic region, climatic variances, cultural trends, socioeconomic background, family and received traditions, et cetera. These forces reach deeply into an individual's life, particularly during the years of pre-pubescence, and leave marks that in many ways cannot be undone; at best they might be shifted. It is from out of this condition (from between these blinders, behind these lenses) that a person perceives and comprehends her environment, and therefrom interacts with her surrounding social milieu. She is simply not equipped to understand her encountered world differently from how the hemmed-in "horizon" of the "natural attitude" has enabled her.[12] (Luft 2011) Accordingly, in the alternate self-theory that we will offer below based on Kristjánsson's three self "sets", we shall need to make the most attunements to this portion of it.

Prior to that, however, the second major problem with Kristjánsson's view will need to be addressed, and this is in the mental model upon which he bases his trio of set structures. As mentioned, Kristjánsson employs a single-tiered framework wherein a person is able to rationally decide the manner and type of characteristics (traits) she would like to have and then embark on a program of disciplined training to effectuate such in her behavior (again, this is very Aristotelean). Yet, a vast preponderance of evidence from the fields of neurological science and psychology yield a very different picture. What appears rather to be the case is that we are creatures built almost wholly on the pre-aware,[13] that our brains function by automatically and very rapidly taking in and then processing stimuli (both external and internal), affixing what might perhaps be called "emotional tinges" to these data, and determining best courses of action through intuitive judgments which are thereby "tagged" and effected in subsequent actions: all of this, it must be stressed, happens entirely *before* rational analysis or reasoned decision-making is even possible (Damasio 1994, 1999, 2012; Dijksterhuis 2004; Gazzaniga 2011; Greene 2013; Haidt 2001, 2012; Kahneman 2003, 2011; Klein 1993, 1998; Sadler-Smith and Shefy 2004; Tversky and Kahneman 1974). The brain, moreover, is incredibly skilled at these pre-thought procedures, able to handle 11.2 million units of data at once, whereas in aware thought (/rational cognition) we can typically accommodate a mere seven items simultaneously (Dijksterhuis 2004). As Jonathan Haidt colorfully put it, we are emotional dogs with rational tails (Haidt 2001): or in other words, we function in our environments from out of the intuitive and emotional systems that dominate our brains and then only sometimes toss in a bit of reasoning as well. Furthermore—as if that were not enough—Haidt also outlines that we are only able to analyze our actions after the fact, meaning that at best we can try to notice what we have done and then (attempt to) determine to do otherwise *the next time;* something which, while not ideal, does still seem to help towards shifting one's baseline automatic intuitions (feedback from social sources works towards such as well). Nevertheless, it should be plain that due to the separate proceedings of the intuitive/emotional/automatic

on the one hand, and the rational/purposive/effortful on the other, the overall system is two-tiered rather than one, with the fundamental portion—and the vastly more important layer for the obvious biological benefits of speed and safety—being that of the pre-aware, with intentional reasoning a distant second. Any self-theory that wishes to establish itself empirically will need to account for this.

We are now prepared to present our alternative self-theory, following the pattern of Kristjánsson's explanatory sets, and then thereafter to make some short remarks on how it relates from out of the core self and into personal identity and whole person issues, before turning to explore consciousness more fully in the following section. Maintaining Kristjánsson's notion of the core self as a fundamental psychological posit, I offer the below re-adapted trio of sets, to which will also need to be added two subsequent elements in order to account for a human self:

(1) *Self-defining traits:* These are one's preferences and outlooks, genetic inheritance, upbringing, choices previously made and the resulting influences (in pre-awareness and in awareness), historical, socioeconomic, epochal, geographic, climatic, et cetera, elements: in short, the formative "defaults" we face plus our having lived through them.

(2) *Self-directing traits:* One's ideas on oneself, but including with the ratio-conceptual the kind of intuitive and affective neural responses mentioned above (associational "tags" and "tinges"), which contribute towards a maintaining, adjusting, or more forcefully shifting of the first set's traits.

(3) *Self-evaluating traits:* Those reflections done with acknowledgment and aim towards ascertaining the nature of oneself vis à vis where one considers oneself to be against where one would like oneself to be; this aspect also includes the intuitive and emotional co-judgments given in the second set.

It will be apparent how Sets 2 and 3 work in tandem, as will the greater identitarian importance of the more delineative first set. Moreover, none of these ideationally composed and influencing sets would remotely be possible in the absence of a mind possessing them, and hence (ours not being a Cartesian or otherwise dualist theory) we will need to augment these with both consciousness and some form of bodily presence (consciousness emergent from a functioning body; see below). The result of this organizational aligning might be stated in equation form as: Set 1 + Sets 2 & 3 + C (for "consciousness") + B_P (for "bodily presence") = the Self (capitalized to indicate this is the core self under discussion).

Thus far, we have established only the core self, which might or might not be enough for artificial intelligence (in the form of software, perhaps), but is certainly not sufficient for a living human being. Let us quickly then also enunciate the terms for personal identity and whole person, and in giving those additional "formulae" I believe it will be clear how these further aspects differentiate. To arrive at personal identity we "flesh out" the bare bodily presence (B_P) needed for the core self by supplementing contingent facts about the body (idiomatic facets and features), other contingent facts connected to personhood,[14] along with feedback from the social realm; this gives us: Personal Identity = CF_B ("contingent facts: body") + CF_O ("contingent facts: other") + FB_S ("feedback: social") + the Self. To conclude our portraiture of a full human individual, to the foregoing we need simply add those myriad embedded ("enworlded") details to which we have alluded, and thus we arrive at: Whole Person (i.e., as within one's lived environ) = E ("embeddedness") + PI ("personal identity", inclusive of the Self) (Oberg 2020; for full definitional details, see especially pp. 49–57). What naturally holds this entirety in unison—that without which none of the aspects could accrue nor even exist—is consciousness, and hence it is in that direction we now move before later attempting to apply the findings we arrive at to our queries on artificial intelligence. Could an inorganic body do what an organic body does? If it might, such would be because said "body" is able to mimic/copy/create some type of consciousness.

3. Consciousness and the Self

> *"To me it is not in the least demeaning that consciousness and intelligence are the result of 'mere' matter sufficiently complexly arranged; on the contrary, it is an exalting tribute to the subtlety of matter and the laws of Nature."*
>
> Carl Sagan (1977, p. 221)

Consciousness, as we know it from daily experience, is often considered to include something extra-physical, to be "more than" the brain which we of course realize is its location, although—and herein lies the trouble—we have difficulty also acknowledging the brain as its source. These "extra" phenomena have been termed in the literature "qualia", but perhaps they are better known by the phrasing Thomas Nagel famously gave: the "what it is like" of this or that (Nagel 1979). On this "extra" line of thought, it is determined that because everything "feels like" something, such "data" (the "feels like") are evidence for the existence of an additional element(s) beyond what might be described through the relevant biophysics; or, via another explanatory avenue, that these qualia point to a more widespread expanse of mind which cannot be accounted for merely by the biophysics. At its root the idea here is that consciousness exceeds matter in some way(s), and that hence regardless of how complete the accounting of said matter might be it will prove inadequate to account for the "excessive" mental side. This argument is an old and well-established one, going back in academic circles at least to Bertrand Russell and Arthur Eddington (Russell [1927] 1992; Eddington 1928), although the roots stretch much further, possibly traceable even to Ancient Greece (depending on the leeway one is willing to grant). Given this depth and breadth we cannot offer the position a full treatment in our current study, and thus instead we simply point the reader towards two exemplary contemporary representatives and titles whose works are provocative and have become commonly known: firstly is Thomas Nagel again with his *Mind and Cosmos: Why the Materialist Neo-Darwinian Conception of Nature Is Almost Certainly False*, and secondly is David J. Chalmers' *The Conscious Mind: In Search of a Fundamental Theory* (Nagel 2012; Chalmers 1996) (we will also give some examples of researchers who align themselves with biophysics as presently being able to, or as will shortly become able to, explain consciousness in the below).

What then is at issue, and particularly as regards the question of the self, for which we came to realize that consciousness is a fundamental must? To those who take qualia as important and challenging to a materialist perspective, every experience in one's awareness carries an associated, singular content which needs to be accounted for. Chalmers, in his work cited, writes that: "When I think of a lion, for instance, there seems to be a whiff of leonine quality to my phenomenology: what it is like to think of a lion is subtly different from what it is like to think of the Eiffel tower" (Chalmers 1996, p. 10). Another example commonly given is that of water: We have difficulty in characterizing the quality of "wetness" that we know from our lived realities with the simple physical equation of two parts hydrogen and one part oxygen. "H_2O" tells me nothing about placing my hands into a river or under a faucet; and this disconnect, it is argued, is crucial and significant. The reader will recall our earlier outline of a two-tiered cognitive model and no doubt anticipates the response that can be made here: It is not the *thinking of the lion* that is providing the "whiff of leonine" aspect to said thinker; rather it is the held concept "lion" by said thinker, with its particular and previously established associations of intuitive judgments and affective affixations ("tags" and "tinges"). Moreover, it is not the *description* "H_2O" that fails to express wetness; it is the difference between tactile sensory processing and notional sensory processing, the former and latter each attached to previously adjudged automatic reactions and emotions (and the former probably also with extant memories of similarly encountered stimuli). Wetness can be accounted for in molecular science, but the *feel to me* of wetness cannot: that, however, is not molecular science's job; it is the task of speech expression, and this is a point for the abstract. The real problem therefore does not lie with each quale and the connected brain-processed physical attributes thereof, rather it is with the confused and confusing way we talk about—and thus think about—each quale and its physics (Oberg 2018; Hofstadter 2007). There is certainly much mystery in life (the

sheer existence of the cosmos itself is astoundingly awe-inducing, and let us not forget the nearly universal acceptance of soul from which we are trying to find the more limited self), but to impute some form of mystical aura onto each thought and sense is, I believe, to miss the whole for its parts, it is to shift oneself so far into the minutiae that creaturehood disappears. What I think is occurring in qualia is the linkage between the ideational and the brain's "labeling" of that idea from out of its pre-aware layer of functioning. Again, these "tags" and "tinges" are automatic, and when they are noted in awareness it is tempting to presume that something extra (something non-biophysical) has been placed thereupon, but in fact it is all just the brain. The question now becomes: What *physically* could be happening if indeed it is only physical?

I suspect the answer lies with neurons and the representative "maps" they form within the brain, acting to aid the "constellations" and networked lattices of consciousness modules within the organ as it coordinates amongst its varied parts (e.g., the brainstem, limbic system, lobes, et cetera). Antonio Damasio very helpfully details how special this unique type of cell is, and the essential role it plays in data-processing as groups of neurons massed into varying shapes create the aforementioned "maps" and "turn" themselves "on" and "off" as occasion dictates from received and/or encountered stimuli (Damasio 2012; see also Damasio 1994, 1999; Gazzaniga 2011; Ramachandran 2011; and Dennett 1991 for further physicalist accounts). When activated and "read" by the brain for more efficient functioning (automatic and pre-aware), these "maps" already contain the intuitions and emotions that sustain the foundational level of our mental model and that form the overwhelming majority of its business in the world. If the brain is operationally held together by neurons and what they enable an organism to do with/to itself managerially, and if the efficiency and effectiveness of those procedures are enhanced by "type-based" reactions in the form of employing adhered intuitions/emotions (i.e., "in X situation trigger Y action/feeling/reference"), then it would make sense for everything to "feel like" something without the need for anything "beyond" coming into play. Moreover, that "feel" is not about the thing itself but rather about the *manner of its processing*. Yet, again we must pause to ask: How could this be?

The final portion of this unraveling, I believe, comes through an application of what William G. Lycan has labeled "second-order representations" (Lycan 1996). He has suggested that in one's introspection (one's internal examination and discovery of those "whiff of leonine" qualia) what is taking place is the relating of second-order (informational) representations to first-order psychological states, and that this is done using purely personally accessible and analyzable items (tokens) that have referents and modes of presentation. In this way qualia do provide phenomenal information that is only available to oneself and not to third parties (e.g., outside researchers or observers), but that nevertheless there are no "special phenomenal facts" to be had thereby (Lycan 1996, pp. 100–1): in other words, the feel of the explanation differs from the explanation itself (my brain working on its tokens versus my mouth trying to talk about my brain working on its tokens). This appears to match well with what has just been discussed in relation to neuronal "maps" and the intuitional and emotional "tagging" or "tingeing" that was given in the two-tiered model of cognition. We live—we think, speak, and act—on the abstract and conceptual level of brain function, and this is extraordinarily useful for the form of being we are in the environments (above all the social environments) wherein we dwell: it is a matter of highly advanced evolutionary efficacy. Much more could be written here, especially with regard to neurons, but for our purposes of the self and AI we must consider the above as justifiably arguably making room for a materialist model and proceed. Our picture: Consciousness is rooted quite firmly, and is fully explicable, in the biological plane; yet it is *experienced* in the ideational plane of assigned names, symbols, and shorthands. If, then, the "feel" of consciousness, which evidently is the mysterious part, is generated thusly there is no need to look further than the body for its source; and this now brings us to our central query and major concern regarding the potentiality of an en-selved (Set 1 + Sets 2 & 3 + C + B_P) artificial intelligence.

4. Artificial Intelligence and the Self

A quick review is called for at this point in order to prepare for what must be a rather conjectural section in the following: We have thus far sought to take the core self out of and away from traditional and unexamined perceptions of the soul, which threaten to blur these two abstract constructs, and instead considered it as a fully psychological posit composed of those external influential elements which act to mold a person, the internal influencing elements which act to form a person, and the additionally influential and influencing elements generated by said person's reflections upon their personhood vis à vis self and society. As a psychological posit this self is essentially an idea, a notion that one carries in one's identity, but given its unparalleled importance for an individual and its vast potential for both good (health, self-growth, et cetera) and bad (contributing to illness, obstructive of growth, et cetera), this cognition requires a label that lends appropriate consequentialist weight;[15] but as a placeholder for that better word—which might not be forthcoming—we will here simply capitalize it (i.e., Self) when considering its possibility within the context of artificial intelligence(s). This (human) Self is furthermore rooted in an (organic) brain that is two-tiered in operation and serves as the organism's manager and chief of maintenance (to put it rather colloquially): its primary level functions automatically, rapidly, effectively and efficiently, organized around patterns of data-processing—on input received externally and internally—and that employs the establishment and association of intuitions and emotions to facilitate rapid and relatively effortless decision-making. The vast majority of the organism's behavior stems from this area. The secondary level of the (human) brain, however, allows for the remarkable ability to rationally analyze data in ways far above what might be accomplished by the foundational mechanisms, yet this is extraordinarily effortful, slow, energy costly, unevenly distributed amongst organism populations (some individuals are better at this than others), and tends to have limited impact upon decision-making, often taking the form of "after the fact" considerations; but in such, and along with feedback from the social realm, perhaps contributing to the re-formation or alteration of existing intuitions and linked emotions.[16]

This brain, moreover, is equipped with a highly developed networked structure of consciousness (more accurately: "consciousnesses", as the picture is of a thoroughly interconnected collection of modular (probably specialized?) centers), and in the normal usage of this consciousness the organism holding it experiences that there is a certain "feel" or "nuance" involved in its application (that is, when with awareness, i.e., not during deep sleep, blackouts, et cetera); these felt phenomena have been broadly called "qualia". Due to the presence of these qualia it has been argued that consciousness cannot possibly be the simple result of brain (i.e., against the "immaterial" (the mental) coming from the material (the physical organ)). These arguments have convinced many, and on the face of one's personal experiential evidence they do appear to hold merit; however, in our examination of the role of neural "maps" as information representations which the brain uses for its data-processing, and the likely connection between these and the practice of intuitive and emotional affixations, we objected that "everything feels like something" for these very biophysical reasons; there need not be an imputation of outside forces to achieve the same results (and we might add at this point a nod to Ockham's emphasis on not multiplying entities unnecessarily). In additional support of our position, we reflected too on the brain's engagement of tokens (or second-order representations of first-order psychological states) to assist its general directing of the vast amounts of information constantly flowing through it in its organism-conducting affairs; qualia perhaps might be thought of in these terms: or better yet, we cease confusing ourselves with such items as qualia altogether and instead think on neural "maps" as such. That, indeed, appears to be the key overall: How we think (and talk) rests on the abstract level of conceptualizations as manipulations of the representations, but in cellular terms our brains are naturally operating on another level in forming and framing the representations, namely that of the biophysico-chemical. There is therefore no reason to insert a vast gap between: these are different parts of the same complex organic machine (i.e., the brain), and as different parts they have their varied

contributions to make. We have once more arrived at the quick: Might a complex inorganic machine be designed and built that is similar enough such that a Self for it proves possible?

In an earlier work on this topic I was convinced that until experiential consciousness as emergent from patterned brain function were better understood an artificial Self would be impossible, and that even if research into consciousness advanced quite far such a Self would nevertheless remain unlikely (Oberg 2017). At the time I concluded that artificial intelligence could not have "metalevels" (as I phrased it), that it must be "necessarily mindless and feelingless. It [the artificial intelligence] stops at, and is incapable of going beyond, the second step along the threefold path we tread as described above: (1) input, (2) processing and response, (3) thinking, feeling, awareness of being" (Oberg 2017, p. 524); now, however, I am not so sure. Partly this change in my approach is a result of further reading on computational structural possibilities (e.g., Graziano 2017); partly it is due to recent developments in existing artificial intelligence (e.g., The World that Bert Built 2022); partly it is a matter of increased reflection on representation and software design; partly it is a better acquaintance with the manner in which the terms in use tend to distort comprehension; but (probably) mostly it is a condition of my "gut reaction" (pre-aware sense) with regard to the brain and its distinctive tools having shifted; and this is incredibly instructive. My automatic comprehension of the connection between the biophysical layer of neurons and the employ of intuitive and emotional mechanisms as biophysico-chemical "tags" and "tinges" to assist in more efficient biological data-processing (e.g., X stimuli is like Y and categorized into Z response) have themselves *become intuitions.* I can now understand them in that foundational way and thus "feel" them without laborious thought. I have come to appreciate representation as an informational analytic tool, and it is precisely this lived aspect that is the guide to percipience (Varela 1999), the "condition for anything to count as an explanation" as Michel Bitbol writes (Bitbol 2021, p. 142). It is this internalizing into my pre-awareness that has finally allowed the concept to be fully assimilated by my brain, and thus accepted in rational awareness, something that has no doubt involved the construction of particular neural "maps" upon which various cognitive centers might operate. In short: markers have been created which are then processed by operating systems. Is this not exactly how computers and similar devices work? What then might be done with software designed towards this end? If systems of symbolization and organization could be made to mimic the associational "labeling" that our brains conduct to enable and promote rapid decision-making, what might be a "natural" machine result? Would these procedures be "felt" by the artificial intelligence in a manner similar to how we "know" consciousness "from the inside"?

The key here, I believe, is how we think about the self (with a lower case "s" to indicate this as the self "in general"); because on this point the tendency has been to signal a dualist model, often despite one's contrary efforts, as an unwanted result of the language used and the ingrained habits of long affiliation. Take, for example, accounts of the self that (rightly) seek to emphasize the role of the body, typically using Maurice Merleau-Ponty's important work as a starting point (e.g., Merleau-Ponty [1945] 2012), but many others in his oeuvre. In emphasizing the "embodied" nature of the self it is not difficult to find such descriptions as "one's experience of the flesh", or (more blatantly) "one's experience of one's flesh". What is being asserted here? The problem of course is the *one,* for the genitive "of" following that word—by pure grammatical function—indicates ownership (hence "of one's flesh" negatively reinforces this thought more than "of the flesh"): the "one" must *necessarily per linguistic structure* become dissociated from "flesh", and this creates a dichotomy. Not only that, but it furthermore establishes a hierarchy wherein the "one" of implied/default "consciousness-self" is over and above (transcending, or perhaps supervening upon) the "flesh" of implied/default "inhabited body". Even the term "embodied" is suspect on this account due to the prefix "em-" with its coapted nuances of "put in / put into / bring to". Again, the formats by which we conceptualize and verbalize create a gap between the biophysical and the mental where there should not be (unless we actually are dualists of some sort (and/or arguably panpsychists)); if consciousness is simply networked brain

function, and if the "feel" of it in which we find a great mystery and a "something else" is merely the employment of second-order tokens affixed with "tags" and "tinges" to assist operationally, then a likewise "what we know from the inside" might be possible non-organically. We shall need to draw this out.

Let us recall our definition for the Self (capitalized again for our specific usage), what its (partly) compositional three sets contain, and then scrutinize each to try and discover whether or not it might be obtained artificially and if so of what such may entail, remembering too that as ours is an additive definition each of the components will need to be present to give rise to the Self; the formula: Self = Set 1 (Self-defining traits) + Sets 2 & 3 (Self-directing and Self-evaluating traits) + C (consciousness) + B_P (bodily presence). Starting with the final element in our Self definition, we might naturally assume a "bodily presence" of some form could also apply in a machine context for the reason that as we cannot have a disembodied Self (not being dualists, mentalists, et cetera), neither could we have a disembodied artificial Self. If we take software as that most likely to align with what we are proposing—as it seems we must—then we may consider that its commands, its lines of code (be those in binary form (i.e., "0", "1") or the human readable type (e.g., "cd", "mkdir", "pwd", "touch")), would need to be hosted on a physically existing device: in other words, on hardware (of some kind). This could be our "bodily presence": so far, so good; but of course that was the easiest part.

"Consciousness" we have stated as the interconnected data-processing system replete with the "feel" of the elements within it, and further that such elements are manipulated by the employment of representational and categorical assignations given to them for the objective of increasing efficiency: these provide said "feel" into the network. Therefore, the system in its operation—if it did have that "feel"—could perhaps be "conscious" in a way recognizably comparable to our own; yet would the "shortcuts" of these second-order tokens aimed at efficacy be necessary for something like a computer program? Could the device not simply take up only the raw data? I do not see why such would not at least be advantageous for an artificial intelligence, even if it were not strictly necessary,[17] and although I am certainly no expert in the area of software writing I suspect that this is exactly what is already happening. Advances in our understanding of neurons, axons, and dendrites (the last two facilitating cross-neuronal communication) have already led to the creation of artificial neurons (Butterflies of the Soul 2022); further brain-mimicking may not be far behind, with applications that could be made to both software and hardware. Hence, by whatever form representative and associative systems develop, it seems there is at least a reasonable chance that they could provide the sort of "tags" and "tinges" that are the reactional/judgmental prompts which Sets 2 and 3 manipulate in analyses.

Thus far, we have the possibilities for "B_P" and "C" along with portions of the make-up of Sets 2 and 3. What is most noticeable about these two sets is their goal-orientation: each of "Self-directing traits" and "Self-evaluating traits" would perforce need a degree of *will*, else there would not be any assessments to be informed by the symbolic markers we have been dwelling upon; the questions "What do 'I' want to be/become?" and/or "How am 'I' doing in that?" would simply not occur.[18] This facet of intention indeed appears to be all that is missing for the referrals of these sets to take place under our artificial constraints; to potentially supply it, however, we may look to Set 1. That set is comprised of items such as preferences, outlooks, goals, et cetera, on the one hand; and the multitudinous molding dimensions of embeddedness (biological, historical, cultural, et cetera) on the other. For something such as a software program, the latter formative aspects would probably be reduced or eliminated vis à vis those that affect a human being (our situations of "made-ness" and "place" necessarily being quite complex), but might yet (partially) be there; what of the former? The answer is perhaps surprisingly obvious and less tentative: Due to the fact that a program is always created for a purpose—and an artificial intelligence would certainly be no exception in that—it too could have these in regard to whatever its originary objective(s) were, i.e., preferences (how best to obtain: less energy use, smoother running, et cetera), outlooks (manner of function: means engaged,

strategic methodology, et cetera), goals (what to pursue and when: tactics and timing, short-term and long-term, et cetera), and the rest. From out of such the channeling of plans formed and followed which is the essence of a will apparently could unfold, and thereby in relation to said will and its concomitant aspirations the workings of Sets 2 and 3 could also co-occur. Another avenue from which we may be able to think the preceding possible involves what Michael Graziano labeled the "Attention Schema Theory": based on the idea that the brain naturally creates not only a model of the whole body for its managerial purposes but also one of itself with which it works in information-processing (Graziano 2017). To achieve this non-organically what would appear to be most pertinent would be the construction of engagement layers whereby the descriptive data-handling level undergirds a level for cognition and language/communication (Graziano 2017; see in particular p. 4); the higher level would operate on (and by, through) the lower, and therefore by such it could self-reflect. It is not difficult to suppose increasing complexity in a framework like this, and out of such more might emerge following well-observed natural patterns wherein the whole becomes more than the sum of its parts through those parts' interworkings. With each segment in place we can try to imagine a Self that is based in and arising from artificial intelligence; this Self might be quite different from the human Self (almost without question quite different), but it is possible to strongly speculate that it could be there. This is a stunning conceivable outcome; what might it mean for a society?

5. Whither the Equality of an AI Self? Can We Think This Self within a "Soul"?

In the pursuit of our question regarding an en-selved artificial intelligence we first considered traditional teachings on the soul, which typically form the background to much of our default and unexamined thinking on the self, thereafter distinguished and dissected the constitution of a human self separately and more restrictedly than soul, followed that with an exploration of the nature of consciousness—upon which the self hangs—and thence explored how a self may potentially appear in the case of a device or, more specifically, something like an AI software program. Central to this were particular aspects of consciousness, namely: the faults in the way we tend to think on consciousness, which generates a mystery where there need not be one (Varela 1999; Bitbol 2021[19]); and the tremendous importance of the representational and organizational tools used for data-processing and manipulation. We speculated that if these were to be replicated in a machine context which was also correspondingly interconnected, directed, and administrative in purpose and scope, then the somewhat incredible judgment that a Self in the manner we described it (Self = Set 1 + Sets 2 & 3 + C + B_P) might arise existentially. If an artificial intelligence Self were to occur, however—and particularly in light of its similitude to the human Self—how would we need to relate to it? What parameters might be required from legal and ethical perspectives? The human Self is intensely socially alive and active, and if a machine Self were alike enough to it that we would have cause for concern for its own sociality, then the issue is indeed an urgent one. One former Google employee, in fact, thinks we are already there. As reported by the BBC's technology section, a man called Blake Lemoine was recently fired from his job in Google's Responsible AI team for claiming that the interactive language software system being developed was sentient and should therefore have its desires respected, even going so far as to release what he termed an "interview" that he and another had conducted with the program (Wertheimer 2022). These are far-reaching assertions that point to both individuality and self-awareness, and further still even possibly to rights which are at least akin to those we have become willing to grant nonhuman animals; perhaps more.

All this, however, may well be premature. Let us recall our definitions for the wider (human) personhood, and how the layers met not only gave rise to one another but also looped into a return, providing cyclical input that continually informs as it builds and rebuilds. In addition to the core Self proper there is personal identity, comprising the Self plus contingent facts about the body (beyond the mere presence of a body, as needed for the Self), other contingent facts (segregated to emphasize the importance of the body), and

feedback from the social realm; then over these (but also "within" them in the sense of an inextricable intertwining) is the realm of whole person: the Self and personal identity with embeddedness (the multitudinous facets of living in a world). In investigating an artificial intelligence self-cum-Self that was as far as we went—the Self—and engaging with it now such appears to be the end limit. A human being is inevitably set within an environment that importantly includes other human beings (to whatever extent, great or small there will be *some*), but what we can reasonably think the best chance for an artificial intelligence that is/has/becomes a Self—i.e., software—would have as its environment is its hardware "body"; and such might well be the full extent of an embeddedness.

It is of course feasible, and maybe even likely, that other software programs would be operating on/in the same hardware and thus we might presume a "society" of sorts, but on closer analysis that would only be the case if the multiple programs were to also continuously *interact*, and furthermore did so in ways that concerned each artificial intelligence's unique intentions aligned with its Set 1 and the reflections of its Sets 2 and 3. Were this to occur there might be enough outside stimuli and responses to such to warrant at least the doubt of personal identity—and therefrom with these a broader embeddedness and hence whole person—to emerge as well, but this is not *necessarily* so; and indeed the units might be designed to prevent this from transpiring. In the absence of personal identity and whole person, the legal and ethical matters would be reduced to concerns such as not causing undue harm to a self-aware creature; again perhaps such strictures could be arranged along the lines we have and are still developing with regard to nonhuman life (although much progress is yet needed here, in my view). Would these measures satisfy what people like Mr. Lemoine deem called for? The query is naturally an open one, but raising it allows the debate to begin; and that this discussion should be had now either in the infancy of these discoveries (if Mr. Lemoine is correct) or prior to them (if the executives at Google—including other experts on AI and the company's internal reviewers—who deny that Mr. Lemoine's allegations are correct), is absolutely fitting. Thinking on these points, we may additionally recognize that even very simple but purposefully intentioned organisms have more "community" than programs which likewise had a measure of intentionality would: paramecia moving and living together have greater reciprocity than would differentiated software.[20] A truly intriguing question would be what a group of artificial intelligences with Selves that did relate might form legally and ethically exclusively by and for themselves; that, however, is far outside the confines of our current study.

Finally, the prospect of an artificial intelligence that is self-aware and has its own goals, objectives, desires, et cetera, raises the specter of it choosing on its own to create anew for itself (and maybe too its compatriots); such is the well-known plot for numerous science fiction stories. As history has repeatedly indicated, however, that which is fiction today can become fact tomorrow; and thus it behooves us to make use of our own awareness and proceed with foresight for whatever reasonable consequences and eventualities we may expect: for example, purposefully engineering desirable (from a human perspective) obstructions if we think a delimiting path wisest. Yet, even if we grant Self to AI we may not also wish to acknowledge soul: but what if the AI did for itself? How could we deny such in the face of an insistence by the other? If we did admit as much, then what? As a species we seem to have the bad habit of letting the things we manufacture rather easily get out of control; whether the merits of the proposed outweigh the demerits surely needs critical examination. Firstly, then, let us call at this (early?) juncture for that which seems most pressing: An open, honest, realistic, and ongoing public dialogue not only about what could be done but about what is *worth doing*.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: I would like to thank the editorial staff at MDPI and the journal, as well as the anonymous reviewers for their kind gifts of time and efforts.

Conflicts of Interest: The author declares no conflict of interest.

Notes

1 Again, that a "me" should continue is essentially a given in these traditions, and indeed barring outliers like Epicurus it was widespread in the extreme in antiquity (and remains thus today); so much so, in fact, that one is tempted to claim a default sense of the hereafter as a core human trait.

2 Note that these are using the Jewish Tanakh's numbering and not a Christian Bible's scheme, wherein they are rendered as 6.5 and 18.5.

3 Especially illustrative of this, I think, is the Jacob character's remark of "*I* will go down" (emphasis added): he, personally, in his identity as Jacob—as the self so-labeled, with everything that entails—will descend to *Sheol* at his passing from this world.

4 Technically these people were not yet "Christian" but were Jesus following Jews and Gentiles, group divisions and identity solidifications were to come later; on this see the very interesting (Boyarin 2004).

5 In this version of the Bible the publisher has chosen to re-set page numbers by sectional division: Old Testament, Apocrypha, and New Testament.

6 We might think this one's "true self", "core self", or maybe less concisely but more accurately, the "ever-enduring self" but, once more, soul is more than the self which we will argue for below: the soul here indicates postmortem continuity, whereas (for us) the self does not.

7 Incidentally, I believe similar comments could also be made regarding how soul is thought in Hindu, some Buddhist, Jain, Sikh, and East Asian ancestor venerating traditions, but we lack the space to consider them.

8 Kristjánsson's primary concern in his book appears to have been educational and/or therapeutical, my own has been theoretical and ethical.

9 Our institutions may after all write laws for any sort of terribly immoral treatment when seen from the point of view of such an en-selved device: e.g., see the ideas haunting some recent popular media reports such as (Wertheimer 2022), discussed below; slavery being only the most ready example of many.

10 Specifically, in the sense of an Aristotelean type self-transformation via repeated purposive practice.

11 For wide analyses of categorized current options, please see (Oberg 2020).

12 Here is another pair of Husserlean terms: together they indicate the inability to "see" (understand) beyond the notional confines of one's conditioning within a non-reflective cognitive position, and hence the need to conduct phenomenological analyses to better attain to one's true setting before being capable of deciding what to do about it.

13 Or put differently, the pre-conscious: that which operates "behind" or "below" what our lucid capacities are able to access. Traditionally this would be termed the "unconscious", but I think that word is misleading because on present modular models of cognitive performance, wherein the brain manipulates a "constellation" of networked "consciousnesses", there is a strong case to be argued for a human being *always* having "consciousness" in some sense, even during deep sleep. To me, "unconscious" signals discontinuity, and on my understanding of the literature I do not think such can be supported in the complete way in which it tends to be comprehended. For some good overviews see (Gazzaniga 2011); for my own reasons for taking one's consciousness as fully coextensive with one's life, see again (Oberg 2020), and for an examination of how the terms and categories in which we think and discuss this issue can lead to more opacity than clarity, see (Oberg 2018).

14 Technically these contingent facts might be grouped singly but I think that one's body carries particular import and thus wish to stress it this way.

15 For reasons that would take us outside of present concerns I have elsewhere termed it a "soft realist self"; see again (Oberg 2020).

16 Haidt (2001, 2012) emphasizes how the responses one gets from others act over time to reinforce or to help change one's intuitions.

17 Recalling too that while the human brain in its pre-aware mode can handle around 11.2 million units of data it nevertheless makes great use of representations, intuitions, and emotions as aids (Dijksterhuis 2004; Haidt 2001, 2012; Damasio 1994, 1999, 2012; Kahneman 2011; et cetera; see again Section 2).

18 An anonymous reviewer kindly pointed out here that many machines already perform a kind of self-evaluation in the form of internal monitoring, but our Set 3 should be thought of as being at a higher level than such as it is necessarily tied to the presence of (some sort of) consciousness. It is a willed self-reflection rather than an unwilled (externally programmed) self-reflection.

19 While he does refer to this too, the model Bitbol presents as an alternative differs from the one given here; his is perhaps less "hard" physicalist than ours.

20 These protozoans do, however, lack the potential for reflection required for a Self, despite demonstrating signs of will (approach/avoid, et cetera); an interesting specimen.

References

Bitbol, Michel. 2021. The Tangled Dialectic of Body and Consciousness: A Metaphysical Counterpart of Radical Neurophenomenology. *Constructivist Foundations* 16: 141–51.

Boyarin, Daniel. 2004. *Border Lines: The Partition of Judaeo-Christianity*. Philadelphia: University of Pennsylvania Press.

Butterflies of the Soul. 2022, *The Economist*, July 2, 71–72.

Chalmers, David J. 1996. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford: Oxford University Press.

Damasio, Antonio. 1994. *Descartes' Error: Emotion, Reason, and the Human Brain*. New York: Putnam.

Damasio, Antonio. 1999. *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*. New York: Harcourt Brace.

Damasio, Antonio. 2012. *Self Comes to Mind: Constructing the Conscious Brain*. New York: Vintage Books.

Dennett, Daniel C. 1991. *Consciousness Explained*. New York: Little, Brown and Co.

Dijksterhuis, Ap. 2004. Think Different: The Merits of Unconscious Thought in Preference Development and Decision Making. *Journal of Personality and Social Psychology* 87: 586–98. [CrossRef] [PubMed]

Eddington, Arthur. 1928. *The Nature of the Physical World*. New York: Macmillan.

Gazzaniga, Michael S. 2011. *Who's In Charge?: Free Will and the Science of the Brain*. New York: Ecco Press.

Graziano, Michael S. A. 2017. The Attention Schema Theory: A Foundation for Engineering Artificial Consciousness. *Frontiers in Robotics and AI* 4: 1–9. Available online: https://www.frontiersin.org/articles/10.3389/frobt.2017.00060/full (accessed on 8 November 2022). [CrossRef]

Greene, Joshua. 2013. *Moral Tribes: Emotion, Reason, and the Gap Between Us and Them*. New York: Penguin Press.

Haidt, Jonathan. 2001. The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment. *Psychological Review* 108: 814–34. [CrossRef]

Haidt, Jonathan. 2012. *The Righteous Mind: Why Good People Are Divided by Politics and Religion*. New York: Pantheon Books.

Heidegger, Martin. 2010. *Being and Time*. Translated by Joan Stambaugh. Revised and Foreword by Dennis J. Schmidt. Albany: State University of New York Press.

Hofstadter, Douglas. 2007. *I Am A Strange Loop*. New York: Basic Books.

Husserl, Edmund. 1999. *The Essential Husserl: Basic Writings in Transcendental Phenomenology*. Introduction and Edited by Donn Welton. Bloomington: Indiana University Press.

Husserl, Edmund. 2001. *Logical Investigations, Volume 1*. Translated by John Niemeyer Findlay. Preface by Michael Dummett, Introduction and Edited by Dermot Moran. Abingdon-on-Thames: Routledge.

Kahneman, Daniel. 2003. A Perspective on Judgment and Choice: Mapping Bounded Rationality. *American Psychologist* 58: 697–720. [CrossRef] [PubMed]

Kahneman, Daniel. 2011. *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.

Klein, Gary A. 1993. A Recognition-Primed Decision (RPD) Model of Rapid Decision Making. In *Decision Making in Action: Models and Methods*. Edited by Gary Klein, Judith Orasanu and Roberta Calderwood. New York: Ablex Publishing, pp. 138–47.

Klein, Gary A. 1998. *Sources of Power: How People Make Decisions*. Cambridge, MA: MIT Press.

Kristjánsson, Kristján. 2010. *The Self and Its Emotions*. Cambridge: Cambridge University Press.

Luft, Sebastian. 2011. *Subjectivity and Lifeworld in Transcendental Phenomenology*. Evanston: Northwestern University Press.

Lycan, William G. 1996. *Consciousness and Experience*. Cambridge, MA: MIT Press.

Merleau-Ponty, Maurice. 2012. *Phenomenology of Perception*. Translated and Introduction by Donald A. Landes, Foreword by Taylor Carman, Historical Introduction by Claude Lefort (1974; Translated by Landes and French original 1945). Abingdon-on-Thames: Routledge. First published 1945.

Morinis, Alan. 2010. *Every Day, Holy Day: 365 Days of Teachings and Practices from the Jewish Tradition of Mussar*. Assisted by Rabbi Micah Berger. Boulder: Trumpeter Books.

Nagel, Thomas. 1979. What Is it Like to be a Bat? In *Mortal Questions*. New York: Cambridge University Press, pp. 165–80.

Nagel, Thomas. 2012. *Mind and Cosmos: Why the Materialist Neo-Darwinian Conception of Nature Is almost Certainly False*. New York: Oxford University Press.

Oberg, Andrew. 2017. Enlightening Your Laptop: Machine Selves and a "Real" Buddhist Self? *Journal of the Korean Association for Buddhist Studies, Special Issue: Encounter of Buddhism & the 4th Industrial Revolution*, 504–37.

Oberg, Andrew. 2018. Talking About Consciousness. *Bulletin of the University of Kochi* 67: 1–11.

Oberg, Andrew. 2020. *Blurred: Selves Made and Selves Making*. Leiden: Brill Rodopi.

Ramachandran, Vilayanur S. 2011. *The Tell-Tale Brain: Unlocking the Mystery of Human Nature*. London: Windmill Books.

Russell, Bertrand. 1992. *The Analysis of Matter*. London: Routledge. First published 1927.

Sadler-Smith, Eugene, and Erella Shefy. 2004. The intuitive executive: Understanding and applying "gut feel" in decision-making. *Academy of Management Executive* 18: 76–91.

Sagan, Carl. 1977. *The Dragons of Eden: Speculations on the Evolution of Human Intelligence*. New York: Ballantine Books.

Smith, David Woodruff. 2013. *Husserl*, 2nd ed. Abingdon-on-Thames: Routledge.

Tanakh: The Holy Scriptures: The New JPS Translation according to the Traditional Hebrew Text. 1985. Philadelphia: The Jewish Publication Society.

The Go-Anywhere Thinline Bible with the Apocrypha, New Revised Standard Version. 2010. National Council of the Churches of Christ in the United States of America. New York: HarperCollins.

The Qur'an. 2004. New translation by M. A. S. Abdel Haleem. Oxford: Oxford University Press.
The World that Bert Built. 2022, *The Economist*, June 11, 21–24.
Tversky, Amos, and Daniel Kahneman. 1974. Judgment under Uncertainty: Heuristics and Biases. *Science* 185: 1124–31. [CrossRef]
Varela, Francisco J. 1999. Dasein's brain: Phenomenology meets cognitive science. In *Einstein Meets Margritte: An Interdisciplinary Reflection*. The White Book. Edited by Diederik Aerts, Jan Brokaert and Ernest Mathijs. Dordrecht: Kluwer Academic Publishers, pp. 185–97.
Wertheimer, Tiffany. 2022. Blake Lemoine: Google Fires Engineer Who Said AI Tech Has Feelings. *BBC News*. July 24. Available online: http://www.bbc.com/news/technology-62275326 (accessed on 25 July 2022).