Augustine and an artificial soul

Jeffrey White

Cognitive neurorobotics research group (Tani unit), OIST, Okinawa, Japan

jeffreywhitephd@gmail.com

Abstract:
Prior work proposes a view of development of purpose and source of meaning in life as a more or less temporally distal project ideal self-situation in terms of which intermediate situations are experienced and prospects evaluated. This work considers Augustine on ensoulment alongside current work into self as adapted routines to common social regularities of the sort that Augustine found deficient. How can we account for such diversity of self-reported value orientation in terms of common structural dynamics differently developed, embodied and enacted over the life course? This work in progress represents divergent views as different sorts of error theory, and hypothesizes that differential development of spindle neural projections underlies different accounts of error for attentive self-correction through iterative attenuation. Some implications for AI augmentation and value alignment of autonomous agents developed on such a model are briefly entertained.

Keywords: self, AI value alignment, robot religion, artificial soul

1. Introduction

> "It was the sort of idea that might easily decondition the more unsettled minds among the higher castes - make them lose their faith in happiness as the Sovereign Good and take to believing, instead, that the goal was somewhere beyond, somewhere outside the present human sphere; that the purpose of life was not the maintenance of well-being, but some intensification and refining of consciousness, some enlargement of knowledge"
>
> - Brave New World, page 177

Seven centuries after Aristotle and two after Aurelius, against the backdrop of the decay of Rome, Augustine also stressed the influence of the individual's inner life on the condition of the community, beginning with his own. His autobiographical *Confessions* describes a life-long, at first implicit, pursuit of invariant principles confounded by social and historical factors, which he makes explicit as a context-independent motivation away from passing worldly attachments and social norms to certain happiness through an inner relationship with God, "our heart is unquiet until it rests in you" (Augustine, trans. Boulding, 1997/2008, 1:1, p 34).[1] His City of God consists in members motivated similarly, by an inner self-association with an ideal world unachievable during the embodied lifetime yet most worthy of pursuit. Everyone is good, there, meaning that they act accordingly, here. He contrasts this population with that of the City of Man motivated to more immediate interests, explaining the fall of Rome. His *Retractions* clarifies earlier accounts, including concerning the embodiment of the soul through lifelong development (Augustine, trans. Bogan, 1968/1999).

Underlying Augustine's philosophy is an understanding that human cognition involves an

[1] "Confession, in Augustine's pregnant sense, is personal engagement in the creative process. It is a willingness to stand in God's truth and become a co-creator with God in his creation of oneself. Part of this new creation in the human mind is the ability to see oneself and one's own history in God's light. To this kind of truthfulness Augustine's whole effort of confession is directed. (see commentary from Boulding, p. 27) Ongoing references to Augustine's *Confessions* are typically cited with chapter:book numbers and page numbers from the consulted translation only. References to other works are made explicit, to avoid confusion.

inner sense mediating a lower faculty entrained in the perceptual immediacy and a higher faculty through the exercise of which eternal truths (God) become accessible. He argues for the existence of God from human access to invariant relationships including mathematical expressions, and moral law "written in the human heart"(1:30, 2:9).

As such relationships are unchanging, they cannot come from experience, because everything in our experience changes, but must originate from another source, something unchanging, enlightening, and without error, God.[2] Augustine associates self-consciousness with error - *si falor, sum*, if I err, I exist - and suggests that life is most meaningful when assessed in light of such a standard, in relation to God rather than passing normative standard. The basic idea is that Augustine reports perceived error relative some differently internalized standard, as expressed in his cities of God and Man, and that accounting for this difference might afford insights into engineering an artificial soul.

This short paper introduces this view, compares with contemporaries, and suggests that these represent two exclusive sorts of error theory.The report concludes with an hypothesis about the role of spindle neurons in providing for the sense of metaphysical self that can be associated with the ensouled condition on Augustine's account, and a possible mechanism for value alignment in artificial agents with implications for robot religion.

2. Augustine on happiness and the soul

I will try now to give a coherent account of my disintegrated self, for when I turned away from you, the one God, and pursued a multitude of things, I went to pieces.
- Augustine, *Confessions*, II.1.1 p 61 (trans. Maria Boulding)

Call it the fault of civilization.
- The Controller, *Brave new World*, p 234

Augustine's *Confessions* recounts events from earliest memory, recognizing a native inclination to truth and away from deception. Error begins in infant jealousy and childhood fear of peer ridicule, with pernicious effects (e.g. deception, cf. 1:19). He understands that common values are grounded in social engagement with the shared object environment, and emphasizes the influence of friendship and community on individual behavior, "sweet to us because out of many minds, it forges a unity" (2:10, p 58) in coordination to achieve common ends "with minds fused inseparably, out of many becoming one" (4:13, p 86). The question becomes from which unity and to which ends one aspires. He reports as a youth acting on opportunities not from their propriety, but from the ability to act, such as when stealing pears or through manipulative gameplay influencing outcomes to his favor. He suggests that such courses of action are amplified with repetition from habit to social role in adulthood, setting up patterns into which future generations are born and embedded, entrained by normal expectations which provide a worldly standard that, through personal

[2] Though inner sense and moral sense seem to differ in Augustine, this difference is reconciled through considerations of gating dynamics focusing on insular cortex in self-consciousness and salience in section 4 and not treated further in the context of Augustine's scholarship, here. Such research might inspire a future doctorate in philosophy of AI and cognitive science.

aspiration, they seek to excel.[3] He is grateful that such habits had not taken hold so deep that he lost sight of salvation, completely.

With an appreciation for Cicero countering more immediate influences during adolescence, Augustine develops a motivation away from transient "earthly things" to universal principles (3:8, p 68). He appreciates the fit of all things in the natural order, and wonders at its seamless continuity, associating unity with "the essence of truth and of the supreme good" and "disintegration" with irrationality and "the essence of supreme evil" (4:24, p 92). He laments distractions such as theater and self-aggrandizement when others and his community deserved his care, instead. He struggles with the idea of evil, reconciling corruption with the good in terms of the human struggle with their fallen condition, thereby removing a conceptual barrier to loving charity first for himself, and then for all.

It is from this perspective that he composes his *Confessions*. He argues that varied inquiry leads to the unity of all things, that motivating self-association with said unity is the source of all derivative goods, and derides anyone who "in self-sufficient arrogance chooses to love a part of it only, a bogus 'one'" (4:16, p 74).[4] He reports that "Truth, in whom there is no variation, no play of changing shadow" (4:10, p 69) evaded him while motivated by "carnal inclination" (4:11, p 71). Worldly things including human bodies are changeable, needs variable leading to a disordered soul to be rectified in light of the superordinate unity that is God represented by universal moral law (*Retractions* 1:8; 2:56; also in the context of the disintegrated self, *Confessions* 7:16; Hundert, 1992).[5] Souls "gain stability" (4:18, p 88, also 7:23-25, p 152-3) through an inner relationship with this ideal, in unity with God, accompanied by a feeling of peace, suggesting that this is their natural state and happiest (consider the end of Book 8; also Ret 1:2). He applies this understanding of inner motivational dynamics explaining his fall to that of Rome's.

Though he does not understand how the soul comes to be embodied, Augustine considers both that it is inherited and unique (cf. *Retractions,* 1:1:3, p 10), associates ensoulment with extended human development, and argues that it can be modified, through "a kind of death of the soul, which consists in the putting away of former habits and former ways of life" in order to be "created anew after a better pattern"(*On Christian Doctrine*, 1:19, p 12) such as one befitting the eternally unified City of God. Rome's greatness was achieved through the efforts of a small minority of citizens virtuous enough to approximate such a pattern, thereby potentiating the lasting organization on which the remainder fed and for a greater share of which later generations competed, sewing division and inviting decay. Simply put, Romans prayed to different Gods in pursuit of diverse ends, with this practice prefiguring their inner and collective dissolution as willful self-association with unifying ideals diminished.

[3] A pattern into which he also fell, and from which he reclaimed himself, for example confessing to manipulative speech for reputation as chair of rhetoric in Milan (*Confessions* 6.6; Hundert 1992, fn. 41, p 103) with these habits at least redirected in his conversion.

[4] Themselves and their situations or offices overweighted their concern for others. The idea that unity is the natural state of the universe and everything in it was active since before Socrates. Augustine recognizes that some people see farther, and this vision puts them in a position where they have a duty to advise and inform, such as Aristotle's concerned citizen. Augustine makes room for such conscientious action contra norms of expectation, while contemporary experts seem to classify such conditions as unhealthy and unhappy as seen in the next section.

[5] Repentance implies to repay through reflection on experience, as one feels a debt to close the felt distance between one's present and an ideal situation. It is self-love, ultimately, to allow one's self to take up such a perspective on one's self, that is key to this process, a notion rejoined in the fourth section of this paper.

Augustine distinguishes between social norms and context-invariant moral law. He describes a superordinate motivation to action in terms of a hierarchy at the top of which is an ideally just sovereign over all time and space, context invariant and eternal, with action proscribed from this point of view compulsory even when in conflict "with the customs or rules of any human society ... even if it has never been done there, before" and which, moreover, should establish a new ruling convention; if "fallen into abeyance, it must be restored, or if not established previously, it must be established now". About the legitimacy of such commands, he argues that "a king has a right to command that something be done in the state over which he reigns" and that "a general contract to obey its rulers holds good in human society" such that to act from commands of a higher authority contra any spatial-temporally local one "does not undermine that community" but rather, when made convention, represents its health, e.g. "to love one's neighbor as oneself" (discussion from *Confessions* 3:15, p 82-3).

Rather than according to social norm, action from the perspective of a universal moral ideal is "unvaryingly self-consistent" with "good and holy people ... servants" of justice in acting from such principles, "blessed" even though actions "merit condemnation" from a conventional standard due to "discrepancy between the appearance of an action and the intention of the agent; and the circumstances of the time, which may be obscure" (3:17, p 84). These good people contribute something special to the community. Attuned to invariant principle, they "prefigure" necessary correction (3:14, p 82); "society is just" insofar as it is organized and reorganized accordingly rather then in service of e.g. arrogant officials "craving for domination" (3:17, p 84). In this way, Augustine is able to explain why good people are persecuted, yet why they wouldn't be happier to act in service of worldly powers for worldly rewards. The world is unjust because people are not good. God is good. Importantly for us, he establishes a universal standard for personal aspiration and purpose in life in the form of an ideal world model, as opposed to the arbitrary standard of relative fitness for social role within different political economies in various states of decay as represented by the view surveyed in the next section.

Although a man of his times, Augustine recognizes that members of different cultures in different eras come to truth differently. For example, he reads Cicero to be deficient in failing to mention Christ, yet co-opts his argument for the mind's necessary self-presence to criticize materialists conceptually impoverished by worldly experience to appreciate the nature of the immaterial soul (cf. Brittain 2012). The mind differs from other things in the way that it is experienced, being constant, ever-present to self-inquiry, yet without objective representation at least in part because it is in the unfolding middle of ongoing self-realization through reconciliation of the present worldly condition with an ideal moral order (consider *City of God* 7:7; compare "vocation" in Ortega y Gasset 2002, White 2022b, and Kant's "moral perfection" in White 2022).

3.

Each of us, of course, ... goes though life inside a bottle. ... We should suffer acutely if we were confined to a narrower space.

- The Controller, p 223

No one loves what he has to endure, even if he loves the endurance, for although he may rejoice in his power to endure, he would prefer to have nothing that demands endurance.

- Augustine[6]

Like Augustine, predictive processing and predictive coding inspired approaches also characterize cognition in terms of a temporal hierarchy. Lower levels are responsible for direct interactive engagement with the object environment over shorter time-scales, and higher levels encode relatively invariant dynamics over longer time-scales. Contemporary commentators working in this context associate narrative self with higher-level, and immediate phenomenal self-perception with lower-level, neural dynamics (cf. Limanowski & Friston, 2020; in AI, Tani & White 2022). However, a sense of a self constant in the face of contextual change, such as Augustine's, has been challenged. Wozniak (2018) has suggested that a "metaphysical 'I'" corresponding with an invariant sense of self such as a "soul" is a "delusion" located in layers of a temporal processing hierarchy. White (2022b) suggests that such a sense of self may differently develop as an aspect of such a processing hierarchy through human adolescence, as a target global energy minimum and context independent source of motivation. This section introduces Northoff and colleagues' spatialtemporal baseline view, briefly compares with Augustine's, and then concludes with an hypothesis on how both sorts of sense may differently develop in human beings, involving spindle neural projections.

Orthogonal to Augustine, Northoff and Smith (2022) focus on how to "bridge the gap" between subjective point-of-view and objective world by way of neuro-ecological self. Northoff's spatial-temporal theory of task-independent thought considers spatial-temporality "a common denominator" (cf. Smith et al. "common currency") for cognitive computational processes across the temporal hierarchy (2018), what Tani (2016) describes as shared metric space (see also Tani 1998, Tani 2009, Tani and White 2020 in the context of self). Northoff and colleagues (2020) argue that this intrinsic spatial-temporality connects the objective "world-brain relation" and "worldbased subjectivity" or PoV. Northoff and Mushiake (2020) propose a computational model expressing this idea in terms of input summation and output normalization through multi-modal integration of upstream perception (summation) and downstream intentional interaction with the object environment (normalization). They differentiate slow and fast processing in terms of neurotransmitters in biological models, and map their model to schizophrenia (involving output normalization) and depression (involving input summation) self disorders. The basic idea is that adaptive inner-dynamics commensurate with prevalent norms correspond with health, while "divisive normalization" corresponds with dysfunction.

Mechanistically, as in the ACTWith model, they consider insular and more generally salience network dynamics, with some interesting potentials for constructive application. For instance in Northoff and Fraser et al.(2022) using free energy and active inference applying AI as a sort of predictive crutch or regulator in the stabilization of self-disorders around common social expectations in order to facilitate healthy, self-confirming interactions, something like a pacemaker for norm-typical action through monitored and potentially directly modulated neural-system dynamics. Risks involving big-tech-med-pol-mediated thought-policing and hyper-normalization are not considered.

Northoff and Smith (2022) associate "scalefree" activity - nested, modulated with task performance (cf. He et a;. 2010) - with resting state structural dynamics traditionally associated with self ("the brain’s spontaneous activity"). Situated neural dynamics involve switching between outer directed, and inner directed, actional and non-actional processes

[6] *Confessions* 10:28, p 258.

(with some overlap). Northoff, Vatansever and colleagues (2022) consider a "baseline model" that plots such dynamics. They note functional connectivity of left and right insula during interoception and body awareness tasks, differently implicating default mode hubs including the anterior cingulate, medial prefrontal and cortical midline corresponding with the temporal hierarchy of predictive processing inspired models, regions associated with future self-projection and opening or closing to information such as when viewing different types of images.

By focusing on neurodynamics gating signals up and down a temporal hierarchy situated in a shared object environment, they consider their baseline as "an internal spatial and temporal reference or standard for the brain’s processing, including its cognition" (p 16) like a "biological clock" as a sort of fingerprint in a way similar to White's (2006, 2010) characteristic modes of moral cognition e.g. saintly, psychopathic (White 2012, 2013) especially for consideration of more or less selective insular dynamics, opening or closing to bottom-up information.

The basic idea with Northoff and colleagues is that embedding environments change, affecting what is inside the head. The collection of more or less stable routinely inhabited environments shapes baseline activity which may be associated with self. In this way, there is an apparent standard for proper function as adaptive fit to context, as represented in disorders such as depression and addiction representing dynamics which resist such modification.

Not unlike Augustine, Northoff et al. (2023) set out three layers of processing. Their main idea is that the brain aligns the agent to the environment by encoding the environment's spatial-temporality in its own three-layer structure, in their case with progressively slower dynamics on top associated with an intersubjectively shared background of consciousness consisting of commonly recognized (cue Augustine's exclusive cities of God and man) relatively long term interaction patterns stabilizing and normalizing expectations, and the lower with relatively rapid adjustments to the object environment in light of these stabilizing expectations. They repeat the idea that middle layer activity can be associated with changing contents of consciousness, not unlike Augustine's inner sense. They assess schizophrenia and depression as in Northoff and Mushiake (2020) with intersubjective correlations of scalefree activity using EEG. They find that "healthy" subjects shared common resting and task state dynamics, especially in social and theory of mind (thinking about what someone else is thinking) tasks. They conclude that abnormal neurodynamics correlate with "idiosynchratic lifeworld experiences" and reinforce the point of Northoff and Smith's (2022) PoV:

"The inter-individually shared topographic and dynamical properties of the brain's spontaneous activity may establish a contextual neuro-ecological point of view within a pre-given, self-evident and natural life-world that is widely shared across healthy human beings".

How widely shared is not clear (again, consider Augustine's two cities), yet the constructive correlation between mental health and socially resonant neural function in Northoff and Fraser et al. (2022) is presumed in their nudge-normalizing AI guidebots, and this is worrisome.

Interestingly, Baek et al. (2023) use functional neuroimaging to show that lonely people process information in the default mode (resting state) "in idiosyncratic ways that are exceptionally dissimilar to their peers" (p. 692). They consider impacts on well-being, suffering from feelings of disconnection and social isolation regardless of "friends" and thereby "raise the possibility that being surrounded predominantly by people who view the world differently from oneself may be a risk factor for loneliness (even if one socializes regularly with them)" (p. 690; cf. p. 693). Lonely people - so, we are talking here about a stable personality or self-construct in the sense of Northoff and colleagues - report that they are not understood well by other people, they "see the world differently" and "lack" a "shared understanding" with Baek and colleagues considering that they may not value and so attenuate to the same aspects of situations (discussion p. 692). Baek et al. cite Courtney & Meyer (2020) who showed that loneliness was associated with reduced neural representational similarity with other people in the medial prefrontal cortex, "suggesting that lonely individuals think of themselves in a way that is more dissimilar to others than is the case for non-lonely individuals." (p. 692)

By finding different aspects of shared object environments salient (in ACTWith terms, opening to some rather than others) one feeds information upstream for intensional processing which eventuates in action plans which in turn inform forward processing including during collaboration or cooperation. With each iteration, dissimilar valuations drive dissimilar action plans, cooperation is confounded and error is perceived in the suboptimal interaction. In this way, attending to different aspects of shared situations drives a "feedback loop in which lonely individuals perceive themselves to be different from their peers"(p. 692). Social coordination becomes more difficult, increases cognitive load, may be stressful, driving irregular dynamics and potential conflict leading to perceived dysfunction as such self-constructs stabilize. Such a condition might warrant a norm-stabilizing AI-enactive pacemaker.

Northoff, Vatansever and colleagues (2022) allude to the extension of their concept of baseline over the course of the lifespan of an organism, but do not consider specific developmental processes, i.e. as Augustine, they do not know how the soul comes to be in the body. Interestingly, recent work confirms divergent value orientation in religious people as reflected in differences in neural-dynamics through development, showing that parental religiosity corresponds with differences in the anterior cingulate function connectivity, a key area of the salience subnetwork of the brain involved in (prospective) attention (cf. Bornstein et al. 2017) but not with key areas encoding (temporally local, concrete) reward in adolescents, with the idea being that rewards associated with God are distant and abstract (Brooks et al. 2022).

One perspective may associate apparently idiosyncratic relatively exclusive value-orientation as-if living in a different metaphorical city, or different world as described by Socrates[7], as abnormal, unhealthy, dysfunction, and perhaps as a disorder, schizophrenic, and another may associate such a different value orientation with devotion such as accommodated in different ways in religious ritual and institution (cf. Hipolito and Hesp, *in press*). The question for us is how different intensional value orientations recognized as associated neural dynamics might develop more or less normally through common structures, and how such dynamics may be formalized for robot experiments.

4.

And what would it mean to say
"I loved you in my fashion"?
What would be true?

[7] As Callicles assessed Socrates in Plato's *Gorgias*, recalling Socrates' leaky jar which resonates with Augustine's idea that the worldly soul cannot find happiness because it cannot be complete. The focus of this paper is on stressful retention of superordinate value orientation when, as Callicles warns, normal people will hurt Socrates rather than receive his example, beating him over the head, eventually forcing his suicide by poison via poor judgment. *Et tu*, second-personal psychology?

- Sting, Why should I cry for you? Soul Cages, 1991

Augustine's *Confessions* is a candid self-report on personal error and struggle to resolve invariant principles for free self-correction. Contemporary accounts emphasize flexibility over such "rigid" dynamics to reduce perceived error with the relatively short-term patterns evident in the passing social-material, shared object environment, with the aim being adaptive fit with that changing environment in order to minimize stress, maintain health and optimize iterative action for well-being by minimizing the sum of accumulated error in interaction with that environment. Augustine's view reflects similar predictive dynamics, but rather than focus on individual resilience by minimizing stress in relatively immediate interactions with the shared environment, he derives error from a divine ideal, even if that means acting contrary to local custom in order to establish a better pattern, even if this incurs personal sanction.

In effect, these represent different sorts of error theory. One implies motivation to reduce error with a project ideal, and the other to reduce error with the relative immediacy. rather than equilibrate between them, the present focus is on accounting for common developmental dynamics that may account for both. Hipolito and Hesp (*in press*) consider religious cognition in a way that is helpful, with the idea that religious cognition such as Augustine's involves an inner self relation, sense of salvation, and corresponding devotion to and veneration of associated ideals. Such inner self relations are also presumed on views of health as adjustment to changing conditions in life.

How can we explain the variability in accounts? It may be that different developmental courses are sufficient to account for differences in self-reports and associated contemporary theories of ensoulment. Consider self-compassion. Self-compassion has been associated with resilience, as an adaptive self-relation with potential to modulate deeply entrained patterns associated with self-disorders during adolescence (Neff and McGehee, 2010; Neff, 2003). Self-compassion reflects Augustine's acknowledgement in his *Confessions* that confession would not be possible without a loving inner self-relation. Self-compassion does not seem to reduce to how one interacts with the environment, but rather involve how some aspect of self relates with what may be represented in terms of a baseline such as in Northoff and colleagues.

Happiness for Augustine involves an inner self-association with context-invariant principles accessible through higher faculties and worthy of pursuit regardless of worldly situation, passing norm, and personal suffering. He communicates this motivating inner self-association in terms of an ideal world unachievable during the lifetime, but with binding obligations bearing on the present. And, he associates free will with the capacity to determine for one's self to which world one is bound and aspires through action.[8] The resulting disposition prefigures variable attenuation. Here, we may associate neural gating dynamics with target-situation salient information, being relatively open to characteristic aspects of one or the other City, God's or Man's.

In this way, we can make consistent sense of the fact that Augustine understands human freedom like God's, to create, but in the limited sense of "self-creation" including religious

[8] We may ask which layer of a temporal hierarchy mediates internal and external dynamics. The reply might be the overlapping insular cortex and anterior cingulate in coordination with precuneus, perhaps as represented by an activity baseline.

conversion. In so far as one is free to determine for one's self one's inner motivating self-association, one directs one's own self-development through episodic interaction with the external world, experiencing intermediate transitions and becoming the summative result. By this mechanism, Rome rose as a collective sum of more or less voluntary iterative self-summing enacted self-associations with unifying ideals, integrated and then disintegrated through top-down dynamics as the leading visions of these ideals flared and diminished. As leadership deteriorated, to maintain such ideal self-associations would be stressful. Those faithful to founding ideals might have felt out of place, as they valued different things, i.e. long-term binding principle over selfish interests over relatively shorter terms, e.g. selling favors and seating horses in the Senate. Though perhaps unhealthy, it might be difficult to represent contrary value orientations as dysfunctional for that fact, alone.

The City of God does not diminish in the ways that human social organizations do on Augustine's account, because each member freely keeps binding ideals in sight. By "clinging" to God on the inside, human beings can for example do unto others, so directing their own self-creation "after a better pattern" and sharing in divinity by way of the body in which this potential inheres, advancing the world toward that moral ideal along with them.[9] This is the vision that Augustine develops according to his self report and communicates as the relative commitment to one City or the other.

What mechanisms might underwrite such an account? In addition to processes implicated in White (2022b; 2014; consider discussion beginning with moral zombies in White and Tani 2016, p. 15; fundamentally, consider also Paine and Tani 2005, for example) the present hypothesis is that spindle neural projections hold project situations against present and possible situations, as the sense of a globally orienting prior-embodied project ideal self-situation, determined with greater precision through iterative interaction with the external world including with other human beings, and representing in some rare cases a possible project solution to perceived population-level social problems over the life-course contributing to the ever-present sense of self as outstanding obligation or debt to an immaterial ideal such as expressed by Augustine[10] and others.

Formalized in such a way and with biological models studied in this light, developmental robotics experiments may be patterned after human neoteny and met with analogous challenges (such as in Bruner, see Tani and White 2022 for brief discussion). Cases such as Augustine's can be studied and corresponding dynamics trained. Value orientation can be checked by stimulating correlate spindle dynamics. Social simulation experiments may be carried out, given expected developments in computing technology. In the end, the difference between dynamical-enactive homeostatic and purposeful-developmental allostatic accounts such as that proposed in the form of religious cognition may prove essential to developing an artificial soul.

[9] Potentiating immortality, also in Aristotle. And reminiscent of Aristotle's understanding of understanding, or perception of understanding. Augustine challenges that to "Know yourself" is a riddle, because the self is all that is known and that what is known about that is mostly where one went wrong. Augustine laments that God was silent when he was errant because his inner sense was occupied with passing situations. Truth is eternal, and cannot be appreciated from the point of view of an embedded, embodied agent immersed in material society. Augustine's life involved recovering from his fallen condition through a self-directed self-development to truth, realized in self-association with a supreme unity, revealed through inner sense directed at highest invariant principles, represented in the moral law and accessible through love. To share in this condition is to share in the divine, to choose to become more like God, and for Augustine this is evidence that the soul - and the self associated with it - is immortal. If we take love to be not wanting to cause pain, and extending to future generations, then it is not clear that minimizing suffering through facile adaptation of worldly routines in the short term is good advice.

[10] Note Augustine's distinction between fate and destiny in this context.

Besides as proof of concept for human moral potential, such a model should afford a means for value alignment of autonomous artificial agents with long-term human interests. Agents designed on a developmental model may be trained on religious and ethical principles alongside everyday mundane tasks, and set offline in rumination (simulating the pensive adolescent's universal audience) to consider non-coercive, nonviolent, cooperative pathways forward from current non-ideal, relatively unjust and unsustainable social norms, conventions and institutional arrangements, to better situations for all. Such a model might reflect the life-course of a Socrates, and compared alongside developmental courses for example based on Augustine by way of his extensive self-report.

Recalling Northoff and colleagues' social-resonant enactive-pacemaker, we can consider Augustine's conversion to Christianity during adulthood as a adoption of such a regulator in his professional placement within a religious institution. The idea is that consistent error in certain stable dimensions might motivate conversion to different ways of life that minimize error in those dimensions, either by adopting a technological prosthesis or, by entering into an existing institution, by living with others like-minded and by perceiving error in terms of internalized standards that may be dissimilar from conventional but shared in that subgroup to different degrees. It may be that traditional religious institutions exist as the sum of all agents so commonly devoted, e.g. the Church, including those who think of themselves as essentially dissimilar from others per Courtney and Meyer. In regards to self and ensoulment as developing purpose and sense of self as outstanding, accounting for all such differences, the basic idea is that developing higher-level longer-time-span cognitive capacities, alongside rapid development in social and spatial capacities, underwrites variably reported self-associations with different and differently (as in modally different) internalized project ideals. PoVs that are not shared in some valuable ways with common values of others and might be commonly perceived as idiosyncrasies are shared with others, even rarely, and religious institutions can be considered a technology that stabilizes personalities around such value orientations.

The sense of self that has been at issue in this paper is one of project or purpose in view of which intermediate situations and afforded actional opportunities are evaluated. One attenuates to different aspects of situations to afford further actionable ends, and objects represent opportunities or obstacles to such opportunities. It is in light of such ends that aspects of present intermediate situations appear valuable, as mediators for progress toward project ideals however selfish or socially distal. One way that AI may be developed to augment and to stabilize human cognition through supported sociality is for meta-analysis of idiosyncratic value orientations such that AI may accommodate idiosyncrasies and support for example lonely people with information commensurate with value orientation, and networking common orientations may open opportunities for creative collaboration. Moreover, information about relevant neural dynamics and their development may afford meta-cognitive oversight over idiosyncratic value orientations that these different neural activity represents. In the spirit of Northoff and colleagues' baseline, users of associated technologies may self-monitor and tailor reinforcement protocols autonomously, thereby addressing some perceived risks.This is the general view. Which actions to which ends, to be realized during whose lifetime and in what ways, these are all questions which remain.

Finally, it may be worth noting that religion as a value-orienting technology seems historically to have been poorly utilized in the solution of large-scale intergenerational coordination problems such as those confronting us today. Pursuant to insights into for example Augustine's self-reported lifelong developing religious orientation, AI models reinforced with robot religion might make solutions to such problems transparent, if not afford an army of long lasting atemporal agents programmed to autonomously bring ideal situations to fruit. In such work, Hipolito and Hesp's (in press) four-aspect assay of religious cognition may prove useful.

Works consulted

Augustine, Bogan M (translator) 1968/1999. The Retractions, from The Fathers of the Church. The Catholic University Press, Washington D.C.

Augustine, Boulding M (translator) 1997/2008. The Confessions (1st edition; study edition). New York City Press; Hyde Park:NY

Augustine, Shaw J F (translator) 2009. *On Christian Doctrine*. Dover Publications, Mineola, NY

Bornstein MH, Putnick DL, Lansford JE, Al-Hassan SM, Bacchini D, Bombi AS, Chang L, Deater-Deckard K, Di Giunta L, Dodge KA, Malone PS, Oburu P, Pastorelli C, Skinner AT, Sorbring E, Steinberg L, Tapanya S, Tirado LMU, Zelli A, Alampay LP. 'Mixed blessings': parental religiousness, parenting, and child adjustment in global perspective. J Child Psychol Psychiatry. 2017 Aug;58(8):880-892. doi: 10.1111/jcpp.12705. Epub 2017 Feb 28. PMID: 28244602; PMCID: PMC5513768.

Brittain C. (2012) Self-Knowledge in Cicero and Augustine (De Trinitate X, 5, 7–10, 16). Medioevo. 2012 Jan 1;37:107-35.

Brooks, S. J., Tian, L., Parks, S. M., & Stamoulis, C. (2022). Parental religiosity is associated with changes in youth functional network organization and cognitive performance in early adolescence. Scientific Reports, 12(1), 17305.

Courtney A, Meyer M (2020) Self-Other Representation in the Social Brain Reflects Social Connection. Journal of Neuroscience 40(29) 5616-5627

He, B. J., Zempel, J. M., Snyder, A. Z., & Raichle, M. E. (2010). The temporal structures and functional significance of scale-free brain activity. Neuron, 66(3), 353-369

Hipólito, I, & Hesp, C. (in press). On religious practices as multiscale active inference: Certainties emerging from recurrent interactions within and across individuals and groups. https://doi.org/10.31234/osf.io/t9632

Hundert, E. J. (1992). Augustine and the Sources of the Divided Self. *Political Theory*, 20(1), 86–104. https://doi.org/10.1177/0090591792020001005

Huxley A. 2006. Brave New World. Harper Collins; NY:NY.

Limanowski, J., Friston, K. (2020). Attenuating oneself: An active inference perspective on"selfless" experiences. Philosophy and the Mind Sciences, 1, 1-16

Neff K (2003) Self-Compassion: An Alternative Conceptualization of a Healthy Attitude Toward Oneself, Self and Identity, 2:2, 85-101, DOI: 10.1080/15298860309032

Neff K, McGehee P (2010) Self-compassion and Psychological Resilience Among

Adolescents and Young Adults, Self and Identity, 9:3, 225-240, DOI: 10.1080/15298860902979307

Northoff G. How does the brain's spontaneous activity generate our thoughts? The spatiotemporal theory of task-unrelated thought (STTT). The Oxford handbook of spontaneous thought: Mind-wandering, creativity, and dreaming. 2018 Apr 5:55-70.

Northoff, G, Klar P,, Bein M, Safron, A. (2023) "As without, so within: how the brain's temporo-spatial alignment to the environment shapes consciousness." Interface Focus 13(3): 20220076

Northoff G, Mushiake H. Why context matters? Divisive normalization and canonical microcircuits in psychiatric disorders. Neuroscience research. 2020 Jul 1;156:130-40.

Northoff G, Smith D. (2022) The subjectivity of self and its ontology: From the world–brain relation to the point of view in the world. Theory & Psychology. 2022:09593543221080120.

Northoff G, Vatansever D, Scalabrini A, Stamatakis EA. Ongoing Brain Activity and Its Role in Cognition: Dual versus Baseline Models. The Neuroscientist. 2022 May 25:10738584221081752.

Ortega y Gasset J, Garcia-Gomez J (translator) 2002 What Is Knowledge? State University of New York Press, Albany, NY

Paine, R. W., & Tani, J. (2005). How hierarchical control self-organizes in artificial adaptive systems. Adaptive Behavior, 13(3), 211-225.

Smith D, Wolff A, Wolman A, Ignaszewski J, Northoff G. Temporal continuity of self: long autocorrelation windows mediate self-specificity. NeuroImage. 2022 Aug 15;257:119305.

Tani J (1998) An interpretation of the 'self' from the dynamical systems perspective: a constructivist approach. J Conscious Stud 5:516–542

Tani, J. (2009). Autonomy of Self at criticality: The perspective from synthetic neuro-robotics. Adaptive Behavior, 17(5), 421-443.

Tani, J. (2016) Exploring Robotic Minds: Actions, Symbols, and Consciousness As Self-Organizing Dynamic Phenomena. Oxford University Press: Oxford, UK

Tani, J., & White, J. (2022). Cognitive neurorobotics and self in the shared world, a focused review of ongoing research. Adaptive Behavior, 30(1), 81-100.

White JB (2006) Conscience: toward the mechanism of morality. Dissertation, University of Missouri-Columbia. http://hdl.handle.net/10355/4327. Accessed 15 Oct 2015

White J (2010) Understanding and augmenting human morality: an introduction to the ACTWith model of conscience. In: Magnani L (ed) Model-based reasoning in science and technology: abduction, logic and computational discovery. Springer, Berlin, pp 607–621

White J (2012) An information processing model of psychopathy and anti-social personality disorders integrating neural and psychological accounts towards the assay of social implications of psychopathic agents. In: Fruili AS, Veneto LD (eds) Psychology of morality. Nova Science Publishers, Hauppauge, pp 1–34

White J (2013) Manufacturing morality: a general theory of moral agency grounding computational implementations. In: Floares A (ed) Computational intelligence. Nova Publications, Hauppage, pp 163–210

White J (2014) Models of moral cognition. In: Magnani L (ed) Model-based reasoning in science and technology: theoretical and cognitive issues. Springer, Berlin, pp 363–391

White J, Tani J, (2016) *From Biological to Synthetic Neurorobotics Approaches to Understanding the Structure Essential to Consciousness, part 1*, APA Newsl. Philos. Comput. 16 (1) pp. 13-23

White J (2022) Autonomous Reboot: Kant, the categorical imperative, and contemporary challenges for machine ethicists. *AI & Soc* **37**, 661–673 (2022). https://doi.org/10.1007/s00146-020-01142-4

White J. (2022b) On a possible basis for metaphysical self development in natural and artificial systems. FILOZOFIA I NAUKA. 2022c:71

Wozniak, M. (2018) "I" and "me": the Self in the Context of Consciousness. Front in Psych,9. https://www.frontiersin.org/article/10.3389/fpsyg.2018.01656