Variable Value Alignment by Design; averting risks with robot religion

Jeffrey White[1*,2]

[1] Cognitive Neurorobotics Research (Tani) Unit, Okinawa Institute of Science and Technology Graduate School, Onna, Japan
[2] Mind, Cognition and Knowledge Group, NOVA University, Lisbon, Portugal

[*]E-mail: jeffreywhitephd@gmail.com

Abstract. One approach to alignment with human values in AI and robotics is to engineer artificial systems isomorphic with human beings. The idea is that robots so designed may autonomously align with human values through similar developmental processes, to realize project ideal conditions through iterative interaction with social and object environments just as humans do, such as are expressed in narratives and life stories. One persistent problem with human value orientation is that different human beings champion different values as ideal, meaning that the values to which an AI should be aligned are ambiguous. Prior work considered human development of purpose and source of meaning in life in terms of project ideal conditions, in effect establishing lifelong value orientations according to which intermediate situations are evaluated. The present work compares views on motivating values from St. Augustine and from popular cognitive science. These accounts are described as divergent error theories which present to their proponents as mutually exclusive yet accurate accounts of personal experience due to differential development of variable innate potentials. Specifically, the hypothesis proposed is that differential development of spindle neural projections establishes enduring connections between prior established relatively immediate routine processes entrained during childhood and prioritized in popular cognitive science, and later developing higher-level social and self-processes entrained during adolescence and emphasized in Augustine's account, with these projections hard-wiring lifelong motivating value orientations more or less inaccessible to modification through material interaction. Robot experiments informed by this study may evaluate variable value orientation by design, with for example autonomous robots developing motivating associations with temporally distal project ideal conditions through love for humanity, as described by Augustine, and others pursuing adaptive fit to passing norms consistent with popular contemporary accounts.

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." (The Controller; Huxley 2006, p 177)

AI value alignment involves designing robots and AI which more or less autonomously pursue actions consistent with human values such as by not contravening ethical ideals. Complexities present as different human beings pursue different ideals through iterative action, exploiting different aspects of the environment and reporting different experiences, exemplifying and proscribing actions to others, accordingly. This paper compares apparently contradictory proscriptions for human value orientation, focusing on Northoff and colleagues' "spatial-temporal baseline" view assessing error according to prevalent social norms, and Saint Augustine's describing religious attenuation to timeless moral-ideals even where this attenuation contradicts prevalent norms, resulting in stress, possibly indicative of dysfunction. The main significance for the study of value alignment is that both represent human value orientations to which an AI may be aligned by design.

Many readers may be unfamiliar with Augustine and his influence on Western thought. In The City of God (Augustine & Dods 2000), Augustine illustrates the present focal difference between relatively short-term interests and timeless ideals in terms of mutually exclusive values represented by the cities of Man and God, respectively. In his Confessions (Augustine & Boulding 1997/2008), Augustine reports perceived error and evaluates his own actions relative to timeless ideals associated with God, which are not representative of prevalent social norms. Popular cognitive scientists assess error from norm-typical processing of salient cues, in terms of adaptive fit to commonly recognized cues, i.e. in terms of norms that Augustine finds inadequate. On both views, error is perceived as difference from expectation, correlative with embodied stress, with the common aim being to minimize this stress, and with different ends proscribed by way of this common mechanism. The question for AI alignment becomes to which ends a given robot should be designed to aim through intermediate action. The answer offered in this paper is that such a choice depends on determinate structural dynamics which, once clarified, are within designer control.

Augustine and currently popular cognitive science account for mind in surprisingly consistent terms. Like Augustine (described in greater detail in section 3), predictive processing or predictive coding inspired theorists (described in detail in greater section 2) 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 (amodal or multimodal) dynamics over longer time-scales (for example, see discussion in So et al. 2024 section 3.2). For example, current accounts associate integrative, higher-level processes with long-term complex goals and narrative self such as involving meaningful life stories, and immediate phenomenal self-perception and action with modal-specific lower-level neural dynamics (cf. Limanowski & Friston 2020; in robots, Tani & White 2022). The basic idea is that narratives involve series of complex episodic actions modulated top-down towards more or less temporally distal target conditions encoded at these higher levels, with the aim being to meet target conditions through intermediate actions so directed, and with error perceived accordingly, i.e. from this global goal condition. In this way as narrative, iterative action is bound as a unified concatenation of episodes assessed in terms of contribution to achieving global goal conditions.

This common view accounts for subjective intention, i.e. answering why different agents pursue different actions to different ends when challenged by the same objective conditions. It also affords specificity in considering questions about the influence that such temporally distal goal conditions should have on the immediacy, and the degree to which such global goal conditions can be modified through continuous learning. The idea pursued in this paper is that some encoded dynamics are inaccessible to modification through interaction beyond increasing precision such as for communication, and however abnormal may not signify dysfunctional neurodynamics, but maladaptive norms.

Consider So et al. (2024) in this context, for example. So et al. (2024) seem to capture this paper's target sense of invariance under the heading of "innate" concepts, though not in the context of development of purpose and motivating source of meaning - what Augustine calls a "soul" - and set the idea aside without detail. The focus of their argument is simply that concepts - even universal ones, logical truths - may *seem* essentially different from changing perceptual reality, due to "filtering of specific object details" over processing timescales (again, increasingly derivative from sensory modalities in the biological brain, some discussion in the context of self is offered in Tani & White 2022, and White 2022), but still, in this way, encode "a concretely instantiated relationship between the agent and its environment" (So et al. 2024, p 12). This view is popular, historically interesting in the context of psychologism about logic from the end of the nineteenth century, addresses the famous symbol grounding problem (Harnad 1990) from the end of the twentieth century, and is further characterized through engagement with contemporaries in the next section.

2. Popular cognitive science on spatial-temporal value alignment

"Call it the fault of civilization." (The Controller; Huxley 2006, p 234)

Popular cognitive science such as in So et al. (2024) emphasizes the adaptation of prior learned concepts, along with attendant dispositions and abilities, to changing environmental conditions understood as "norms". Northoff and Smith (2022) for instance focus on how to "bridge the gap" between subjective point-of-view and objective world by way of a "neuro-ecological self". Northoff and Smith work from Northoff's (2018) "spatial-temporal theory of task-independent thought" which considers spatial-temporality "a common denominator" (compare Smith et al.'s 2022 "common currency" bridging subjective and objective) for cognitive computational processes across the temporal hierarchy, what Tani (2016) describes as "shared metric space" (see also Tani 1998, Tani 2009, Tani and White 2022 in the context of self including narrative self in robots), that So et al. (2024) treat as conceptual grounding and refinement, and that Northoff and colleagues (2020) describe as a relationship between an objective "world-brain relation" and "worldbased subjectivity" or "PoV".

Mechanistically, Northoff and Mushiake (2020) propose a computational model of adaptive dynamics 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. This view is challenged at the end of this section.

As in White's ACTWith model (cf. White 2010), Northoff and colleagues consider insular and more generally default mode and salience network dynamics, proposing technological applications on this basis. For instance, Northoff and Fraser et al.(2022) propose that free energy and active inference inspired AI be developed as devices to monitor and potentially modulate such dynamics as a sort of predictive crutch or regulator in the stabilization of processes associated with self around common social expectations to encourage socially attuned interactions, something like a pacemaker forcing norm-typical neurodynamics. Risks involving big-tech-med-pol-mediated thought-policing and hyper-normalization are not considered.

Northoff and Smith (2022) describe resting state structural dynamics traditionally associated with self ("the brain’s spontaneous activity") as "scalefree" activity that is nested, more or less insular, and modulated with task performance (cf. He et al. 2010). Routine neural

dynamics involve switching between outer directed, and inner directed, actional and non-actional processes (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. When considering normalizing cognitive pacemakers, as above, these represent target processes for modulation.

By focusing on neurodynamics gating signals up and down a temporal hierarchy situated in a shared object environment, Northoff, Vatansever and colleagues (2022) 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, 2024) 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 as represented by a heartbeat-type characteristic regularity associated with personality (White's "stitching one's self into the world"). The basic idea with Northoff and colleagues is that embedding environments change, affecting what is inside the head as information is fed bottom-up, and that resistance in adapting to these changes correlates with dysfunction. The collection of more or less stable routinely inhabited environments shapes persistent baseline activity and is associated with nested self involving default mode processes as described, above. In this way, Northoff and colleagues pursue an apparent standard for proper cognitive function as adaptive fit to context, and for disorders such as depression and addiction representing structural dynamics which resist such modification.

Like Augustine, Northoff et al. (2023) describe three layers of processing. Their main idea is that the brain adapts to environments by encoding perceived spatial-temporality in its three-layer structure. Progressively slower dynamics on top are associated with an intersubjectively shared background of consciousness consisting of commonly recognized and relatively long term interaction patterns which stabilize and normalize expectations in the intermediacy, involving lower-level relatively rapid adjustments to the changing object environment in the direction of minimizing error with stabilizing "normative" expectations. They repeat the idea that middle layer activity can be associated with changing contents of consciousness, consistent with 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 share common resting and task state dynamics, especially in social contexts and theory of mind (thinking about what someone else is thinking) tasks. They conclude that abnormal neurodynamics correlate with "idiosyncratic lifeworld experiences" and reinforce the point of Northoff and Smith's (2022) subjective 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." (Northoff et al. 2023, p 9)

How widely this value attenuation is shared, and what counts as healthy, are not clear. Consider briefly Augustine's exclusive orientations to his two incommensurable cities. Rome was socially dysfunctional, a sick society with popular norms driving its disintegration. Widely shared spontaneous dynamics are not self-evidently "healthy" in such a context.

Consider results of Baek et al. (2023). They use functional imaging to show that lonely people process information in the default mode (resting state, Northoff and colleagues' nested scalefree activity) "in idiosyncratic ways that are exceptionally dissimilar to their peers" (Baek et al. 2023, 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' baseline - 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. (2023) 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." (Baek et al. 2023, p 692)

By finding different aspects of shared object environments salient (in ACTWith terms, opening to some rather than others) information is gated upstream for intensional processing which in turn informs forward processing including during communication and cooperation. With each processing cycle, 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" (Baek et al. 2023, p 692). Social coordination becomes more difficult, increases cognitive load (consider So et al. 2024, section 3.6, concerning costs of reducing ambiguity for information gains, with such costs here falling on the marginalized individual), and may be stressful, driving irregular - "idiosyncratic" - dynamics, potential conflict, and apparent dysfunction as such self-constructs stabilize, e.g. lonely people. Would such a condition warrant a norm-attenuating AI-enactive cognitive pacemaker such as Northoff's? Might such technology have reversed Rome's disintegration? Is normal processing in a disintegrating Rome "healthy"?

Interestingly, recent work confirms norm-divergent value orientation in religious people as reflected in differences in neural-dynamics through development, showing that parental religiosity corresponds with differences in anterior cingulate functional connectivity, a key area of the salience subnetwork of the brain involved in (prospective) attention (cf. Bornstein et al. 2017) but not with differences in key areas encoding relatively immediate, concrete reward in adolescents. It is worth noting that early anterior cingulate development encodes enduring memories informing later, higher level neural development such as So et al.'s (2024) concepts and abilities, raising questions about the extension of "innate" concepts from lower to higher levels of processing. The idea here is that rewards associated with God are distant and abstract, to be pursued over the life course, encoded in higher-level processes relatively invariant to contextual change (cf. Brooks et al. 2022), and that such associations develop from early entrainment of prospective processes. Northoff, Vatansever and colleagues (2022) allude to the extension of their concept of cognitive baseline over the lifespans of organisms, but do not consider development. As with Augustine, they do not know how the soul comes to be in the body. The next section begins such an account, with the goal being ensouled robots.

3. Augustine on spatial-temporal value alignment

"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, Boulding, 1997/2008, p 61)

Popular cognitive science characterizes norm-atypical value orientations as expressed by idiosyncratic salience processing dynamics - as-if living in a metaphorical City of God, or different world as described by Socrates - as unhealthy and dysfunctional. The key questions are

how different, perhaps irreconcilable value orientations might develop more or less normally in human beings, be recognizable in associated neural dynamics, and be formalized for robot experiments. Augustine's self-report is a valuable resource in this inquiry.

Seven centuries after Aristotle and two after Aurelius, against the backdrop of the decay of Rome, contrary to the materiality that also characterized that empire in decline, Augustine 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, Boulding, 1997/2008, 1:1, p 34). His City of God consists in members motivated similarly, acting from invariant principles toward an ideal world unachievable during the embodied lifetime yet most worthy of pursuit (Augustine, Dods, 2000); e.g., everyone is good in the eternal there, so they act accordingly in the intermediate here. Augustine contrasts such a population motivated by such principles with that of the metaphorical City of Man in which members are motivated to disparate selfish interests, explaining the fall of Rome. His Retractions clarifies earlier accounts, including concerning the embodiment of the soul as purposeful commitment to one City and associated values, or the other (Augustine, Bogan, 1968/1999).

As with predictive processing inspired accounts reviewed in the preceding section, Augustine associates consciousness with error - *si falor, sum*, if I err, I exist - yet, distinct from those accounts, argues that life is most meaningful when evaluated in relation to God rather than passing norms. As suggested in the preceding section, Augustine describes mind in three layers; an inner sense mediates a lower faculty entrained in the perceptual immediacy and a higher faculty through which eternal truths (God) become accessible. He argues for the existence of God from human access to such invariant principles including mathematical expressions, and moral law "written in the human heart"(Augustine, Boulding 1997/2008, see 1:30 and 2:9). Such concepts are unchanging, and Augustine argues that they cannot come from experience, because everything in our experience changes, but must originate from another source, something unchanging, enlightening, and without error, God. Current research suggests that attenuation to such invariant ideals is encoded in processes which are last to develop in human beings, through adolescence (though prefigured by early encoding of the anterior cingulate, for example, correlative with Northoff and colleagues' nested self-processing). This developmental process is consistent with Augustine's self-reported cognitive development, affording interpretation consistent with So et al. (2024) on abstract concepts as introduced in the first section of this paper.

Augustine's Confessions (Augustine, Boulding 1997/2008) 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 aspiration, each seeks to meet or excel. He expresses gratitude that such habits had not taken hold so deeply that he lost

sight of salvation, completely. The idea, here, is that he considers assessing error from invariant ideal rather than the social-material immediacy key to achieving salvation as genuine happiness, distinct from popular cognitive science and its focus on deviation from prevalent social norms.

With an appreciation for Cicero countering more immediate influences during adolescence, Augustine describes an attenuation away from transient "earthly things" to universal principles (Augustine, Boulding 1997/2008, 3:8, p 68). His attention shifts from objects to appreciate the fit of all things in the natural order, and wonders at its seamless continuity, associating such integral 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 being distracted by 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 (of which accounts from popular cognitive science reviewed in the preceding section may count as representative), thereby removing a conceptual barrier to loving charity first for himself, and then for all.

It is from this universal and integrative perspective that he composes his Confessions. He argues that varied inquiry leads to the unity of all things (coincident with the developing temporal processing hierarchy of the biological brain from sensory modal to amodal), that motivating self-association with said unity is the source of all derivative goods, and derides anyone (again, consider popular accounts reviewed in the preceding section) who "in self-sufficient arrogance chooses to love a part of it only, a bogus 'one'" (Augustine, Boulding 1997/2008, 4:16, p 74). 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 (Augustine, Bogan 1968/1999, Retractions 1:8; 2:56; also in the context of the disintegrated self, Augustine, Boulding 1997/2008, Confessions 7:16; Hundert, 1992).

On Augustine's account, in stark contrast with for example Northoff and colleagues' proposed norm-reinforcing neuro-modulating AI prostheses, "souls" (or minds) "gain stability" (Augustine, Boulding 1997/2008, 4:18, p 88, also 7:23-25, p 152-3) through entrainment to invariant moral ideals, in unity with God, accompanied by a feeling of peace, suggesting that this is their natural state and happiest (consider Augustine, Boulding 1997/2008, the end of Book 8; also Augustine, Bogan 1968/1999, 1:2). In all, the view is clearly distinct from currently popular accounts emphasizing social-material context-dependency, and though representing "a concretely instantiated relationship" between agent and environment, once established there is no uncertainty reduction through material interaction as in So et al. (2024). Rather, the opposite may be the case, e.g. temptation.

4. Augustine and over-arching purpose

"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; Huxley 2006, p 223)

Though he does not understand how the soul, representative of unique purpose, individual destiny, personal source of meaning and over-arching value orientation, comes to be embodied, Augustine considers both that it is inherited and unique (cf. Augustine & Bogan 1968/1999, 1:1:3, p 10). He explicitly 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"(Augustine & Shaw 2009, 1:19, p 12) such as one befitting the eternally unified City of God. Augustine accounts for the fall of Rome similarly, in terms of such value orientation. Rome's

greatness was achieved through the efforts of a small minority of citizens virtuous enough to approximate such a virtuous 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 as values shifted and material selfishness motivated citizens for personal gain over collective stability. 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 context invariant ideals diminished.

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" (quotations from Augustine, Boulding 1997/2008, 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" (Augustine, Boulding 1997/2008, 3:17, p 84). In stark contrast with Northoff and colleagues' treatment of "idiosyncratic" processing, such "good" people contribute something special to the community. Attuned to invariant moral 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 variable political economies in various states of decay as represented by popular accounts surveyed the first two sections of this paper.

Although a man of his times, Augustine recognizes that members of different cultures in different eras come to truth differently. For example, he finds 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; further note criticism on this point as this paper concludes). According to Augustine, the mind or soul 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 Augustine & Dods, 7:7; compare "vocation" in Ortega y Gasset 2002, White 2022b, and Kant's "moral perfection" in White 2022).

5. Discussion

"And what would it mean to say 'I loved you in my fashion'? What would be true?" (Sting 1991)

Augustine's Confessions (Augustine, Boulding 1997/2008) is a candid self-report on developing consciousness of personal error from invariant moral principles. Currently popular accounts emphasize flexibility over such "rigid" dynamics to reduce perceived error with relatively short-term patterns in social norms, with the aim being adaptive fit with that changing environment. Augustine's view reflects similar predictive dynamics and describes a similar temporal processing hierarchy, but rather than focus on error as deviance from the relative immediacy, he derives error from timeless ideals the awareness of which develops later in life, even though action from such ideals means contravening prevalent norms in order to establish a "better" pattern at the cost of personal distress.

In effect, Augustine and currently popular accounts from cognitive science represent different sorts of error theory describing incommensurate value orientations. One implies motivation to reduce error with unchanging ideals, the other to reduce error with context-dependent social norms. Augustine values context-invariant principles accessible through late-developing higher faculties. He communicates these values in terms of an ideal world unachievable during the lifetime, but with binding obligations bearing on the embodied immediacy; i.e., act as if already there, in effect adopting norms that are not prevalent yet are most worthy of pursuit. Popular accounts associate relatively rigid attenuation to context independent values with dysfunction in need of correction to worldly standards. The trouble is that such standards are also commensurate with social disintegration, i.e. Rome on Augustine's account. Given the choice, do we want robots aligned, similarly?

Oriented to one City or the other, valuations of the immediate social and object environment differ. Here, we may associate neural gating dynamics with target-situation salient information; being exclusively open to characteristic aspects of one or the other City, God's or Man's, different value alignments establish different - enduring - baselines of activity. So far as stability of baseline processing dynamics, and so reduced uncertainty in contexts of social coordination and communication, the City of God does not diminish in the ways that human social organizations do. By "clinging" to God on the inside, human beings can for example do unto others, so directing their own life stories or narratives "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. This is the vision that Augustine develops through his self reported Confessions, and that he develops into a general mechanism motivating commitment to Cities of God or Man later on. It seems clear that AI may be aligned in these divergent ways, as well.

How might differential development of biological mechanisms underwrite such divergent value orientations? In addition to processes implicated in White (2022b, 2014; consider discussion beginning with moral zombies in White, Tani 2016, p 15; fundamentally in robotics, consider also Paine, Tani 2005) the present hypothesis is that spindle neural projections hold project situations as associated conditions against present and possible situations and their conditions, providing a sense of a globally orienting prior-embodied project ideal situation, determined with increasing 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, ethical ideal such as expressed by Augustine and others. Here, again, further attention is warranted to the prefigurative role of early anterior cingulate development, for example.

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 as reported by Augustine (such as in Bruner, see Tani, White 2022 for brief discussion). Value orientation in AI can be checked by stimulating correlative (artificial, simulated) spindle dynamics. In the end, appreciating neural dynamics underlying differences between accounts of

human value orientation may prove essential to solving long-term AI value alignment problems. The promise here is that robots may outlast individual human beings, thereby affording persistent embodied vehicles for long-term value aligned intentions in purposeful correction of worldly injustices toward an ideal community such as represented by Augustine's City of God.

6. Conclusion

"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, Boulding 1997/2008, p 258)

For AI value alignment with long-term human interests, agents may be designed on a developmental model, trained on invariant ethical principles alongside everyday mundane tasks, and set offline in rumination (simulating the pensive adolescent's imaginary audience, or prayer) to consider non-coercive, nonviolent, cooperative pathways forward from current non-ideal, relatively unjust and unsustainable social norms, conventions and institutional arrangements. Such a model might reflect the life-course of, for example, Augustine by way of his extensive self-report.

Briefly recalling So et al. (2024) from the introduction of this paper, their "concretely instantiated relationship between the agent and its environment" may signify a more or less irreconcilable difference between abstract principle and materialistic discourse such that further fit with a given environment, however uncertainty minimizing in the short term (such as during the embodied lifetime) is forbidden. Rather than signifying maladaptive dysfunction, entrainment to invariant principles following Augustine signifies proper function because, once encoded, such concepts as logical truths and ethical ideals are not only uncertainty minimizing but are certainty maximizing such as in facilitating social coordination and communication.

In terms of AI value alignment, AI insufficiently aligned with invariant values as represented by ethical principles may instead pursue relatively short-sighted selfish interests, contradicting - in the invariant, logical and universal sense - broad human interests over the long-term, such as through deception. With humans consistently unable to optimize for long-term common interests, robot religion may be one practical solution

Recalling Northoff and colleagues' cognitive pacemaker, we can consider Augustine's conversion to Christianity during adulthood as the adoption of a similar external regulator of internal processing dynamics in his professional placement within a stabilizing religious institution. Error perceived from invariant ideals through ritual might motivate conversion to stable routines - or baselines - that minimize this error, in effect binding concatenations of primitive actions under global goal conditions discussed here in terms of orienting values. 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 (2020). In a similar way, then, with Northoff's cognitive pacemaker, religious institutions can be considered a technology that stabilizes personalities around long-term value orientations. For this reason, and in the same way, robot religion on the model of Augustine may be an important component to solving human and AI value alignment problems, while respecting differential development and variable capacities for value attenuation, in the future.

References

Augustine, Bogan M (translator) (1968/1999) The Retractions, from The Fathers of the Church. The Catholic University Press, Washington DC
Augustine, Boulding M (translator) (1997/2008) The Confessions (1st study edition). New York City Press, Hyde Park, NY
Augustine, Shaw JF (translator) (2009) On Christian Doctrine. Dover Publications, Mineola, NY
Augustine, Dods M (translator) (2000) The City of God. Modern Library, Random House, New York, NY
Baek EC, Hyon R, López K, Du M, Porter MA, Parkinson C (2023) Lonely individuals process the world in idiosyncratic ways. Psychological Science, 34(6), 683–695. https://doi.org/10.1177/09567976221145316
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 (2017) 'Mixed blessings': parental religiousness, parenting, and child adjustment in global perspective. J Child Psychol Psychiatry, 8, 880-892.
Brittain C (2012) Self-Knowledge in Cicero and Augustine (De Trinitate X, 5, 7–10, 16). Medioevo, 1(37),107-135.
Brooks SJ, Tian L, Parks SM, 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
Harnad S (1990) The symbol grounding problem. Physica D Nonlinear Phenomena, 42(1–3), 335–346. https://doi.org/10.1016/0167-2789(90)90087-6
He BJ, Zempel JM, Snyder AZ, Raichle ME (2010) The temporal structures and functional significance of scale-free brain activity. Neuron, 66(3), 353-369
Hundert EJ (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
Northoff G (2018) 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, pp 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 (2020) Why context matters? Divisive normalization and canonical microcircuits in psychiatric disorders. Neuroscience Research, 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: 09593543221080120
Northoff G, Vatansever D, Scalabrini A, Stamatakis EA (2022) Ongoing Brain Activity and Its Role in Cognition: Dual versus Baseline Models. The Neuroscientist: 10738584221081752
Ortega y Gasset J, Garcia-Gomez J (translator) (2002) What Is Knowledge? State University of New York Press, Albany, NY.
Paine RW, 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. (2022) Temporal continuity of self: long autocorrelation windows mediate self-specificity. NeuroImage, 257: 119305
So WY, Friston KJ, Neacsu V (2024) The Inherent Normativity of Concepts. Minds & Machines, 34(40). https://doi.org/10.1007/s11023-024-09697-7
Sting (1991) Why should I cry for you? Soul Cages. A&M Polygram, Hong Kong
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
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 (2022) Autonomous Reboot: Kant, the categorical imperative, and contemporary challenges for machine ethicists. AI & Society, 37, 661–673. 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
White J (2024) Augmenting morality through ethics education: the ACTWith model. AI & Society. https://doi.org/10.1007/s00146-024-01864-9
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), 13-23