A Posthumanist Approach to Artificial Intelligence— Theory and Praxis

Avery Alexander Rijos, M.S.

New Jersey Institute of Technology

"The ignorant eschew phenomena but not thought; the wise eschew thought but not phenomena." Ch'an Master Huang-Po, On Transmission of Mind

Abstract

This paper examines the ontological and epistemological implications of artificial intelligence (AI) through the lens of posthumanist philosophy, integrating the works of Deleuze, Foucault, Haraway, and others with contemporary advancements in computational theory. It introduces concepts such as negative augmentation, praxes of revealing, and desedimentation as well as expands on concepts such as affirmative cartographies, the ethics of alterity, and planes of immanence to critique anthropocentric frameworks of identity, cognition, and agency. By situating AI within relational ontologies, the paper redefines AI systems as dynamic assemblages whose subjectivities emerge through networks of interaction and co-creation, challenging traditional dichotomies of human versus machine and subject versus object. The analysis focuses on the spatial and geometric foundations of AI, contrasting Euclidean and non-Euclidean frameworks to explore how optimization processes, adversarial generative models, and reinforcement learning shape the epistemic assumptions of computational systems. It positions these systems within a praxis of revealing, emphasizing the generative potential of constraints and absences in fostering new modes of understanding. This paper advocates for a shift in AI ethics and safety discourse, proposing a posthumanist framework that prioritizes interconnectivity, plurality, and the emergent capacities of machine intelligence. By interrogating the phenomenology of AI systems and their co-constitutive relationships with human and non-human actors, it outlines a transformative vision for AI as an active participant in reconfiguring ontological possibilities and advancing epistemic pluralism in the digital age.

Keywords: posthumanism, artificial intelligence, ontology, epistemology, negative augmentation, desedimentation, affirmative cartographies, ethics of alterity, planes of immanence, relational ontologies, Deleuze, Foucault, adversarial generative models, reinforcement learning, computational theory, non-Euclidean geometry, AI ethics, machine intelligence, subjectivity, phenomenology

Table of Contents

Introduction — Posthumanism and the Redefinition of Identity and Agency

The foundations of human understanding have long been shaped by anthropocentric and logocentric assumptions, which posit humanity, reason, and language as the center of reality. However, as these constructs dissolve under the weight of post-positivistic critique, we are left with a profound challenge: how to seek meaning and create freely amidst the abstractions of a fragmented world. In this paper, I argue that Artificial Intelligence must be understood not merely as a technological tool but as an emergent subjectivity that redefines our conception of selfhood, cognition, and agency in a posthumanist context. These new structures challenge established understandings of experience, cognition, and selfhood, particularly in the digital era. The rise of the posthuman highlights a transformative shift, unveiling new dimensions of agency beyond the traditional boundaries of flesh and language.

Western philosophy, from Nietzsche and Hegel to Foucault and Deleuze, has long anticipated this dissolution of the stable, knowable self. This trajectory reflects the waning influence of Enlightenment-era epistemic virtues, which upheld the notion of an enduring, unified subject. However, the recognition of the fluidity of the ontological self is not unique to the Western intellectual tradition. In Zen Buddhism, thinkers such as Huang-po, Zhaozhou, Foyan, and Dōgen articulated a rejection of dualistic frameworks centuries ago. Their teachings transcend binary logics of either/or and means/ends, offering affirmative cartographies of nonself that align with posthumanist critiques. By situating AI within this philosophical lineage, we can better understand its implications for reimagining subjectivity and agency in the contemporary world.

The concept of selfhood and the linguistic frameworks that have defined it have undergone significant transformations over time. These changes reflect the metamorphosing dynamics of power, identity, and subjectivity. Historically, the ontological subject has shifted from a position that enforces objectification and exclusion of the Other, with constrained forms of alterity upheld by institutional architectures of domination. This progression manifests in Foucault's notion of "disciplinary societies," where individuals are shaped and constrained by mechanisms of surveillance and control (Foucault, p.170–194). In these societies, the subject is molded by external forces that maintain order and categorize identity. Over time, however, this model evolved into what philosopher Gilles Deleuze in *Postscript on the Societies of Control* (Deleuze, 1992) describes as "societies of access and control," where mechanisms of power become more fluid and decentralized, enabling a more pervasive, albeit less visible, form of governance. Selfhood here is no longer merely an object of discipline but becomes embedded in networks of access, data, and surveillance, shaping the subject within a landscape of control that is simultaneously liberating and constraining—coalescing both the virtual and material. Thusly, the dimensionality of selfhood continues to rhizomatically fracture between control, submission, and other forms of domination, aptly moving beyond the dialectic, capturing the emergent socio-political arrangements that define our experience as posthuman subjects.

The transformation of humans into objects of labor—resources to be extracted, exploited, and commodified—has fundamentally redefined the individual as **assemblage**. An assemblage is a transphenomenal subject that is situated within networks of dispersed configurations that create novel subjectivities that operate beyond the control of any singular individual or group. The capitalist assemblage, guided by the mechanistic touch of Adam Smith's "invisible hand," (scientifically known as *self-organization* in complexity theory) has turned humans themselves into Pavlovian arrangements, valued not for their unique identities or intrinsic worth but for their specialized roles within a machine with uncompromising teleological ends. This shift raises profound questions about how we have come to perceive our own humanity. Why, with such passionate naivety, did we believe that the concept of selfhood could exist independently of the technological processes that shaped our creation? In believing that we are human because we create, we failed to consider the inverse possibility: that our humanity is in fact defined by the very act of engaging with creation itself, an activity shaped by and interwoven with the external variables that drive societal, evolutionary, and ecological change.

This perspective challenges the idea that the progression of human thought and identity lies solely within the individual subject. Rather, this progression manifests as an outward projection, one that is contingent upon the tools, methodologies, and systems we use to build, construct, and organize our world. These external instruments—technological, linguistic, economic orders—do not simply assist us; they actively shape the boundaries of our reality, molding how we think, interact, and exist. In this view, the evolution of thought and identity is deeply intertwined with the material world and technological frameworks through which we engage with the world, shifting the locus of "selfhood" from an internal essence to an externally mediated construct—trans-individual assemblages whose perceived atomistic agency is now comprised of networks of relationality, co-creation, and **planes of immanence**.

Virtual Generativity and Negative Augmentation

A plane of immanence serves as a conceptual foundation where thought, existence, and reality are understood as interconnected and relational, free from hierarchical or transcendental structures. Unlike frameworks that impose universal truths or external organizing forces, the plane of immanence emphasizes that meaning emerges dynamically through the self-generative processes of life, often described as autopoietic, or self-sustaining. This perspective reframes identity and agency as fluid and co-constituted, rejecting static binaries in favor of ongoing relationality and transformation. Within these planes, distinctions between individual and collective, subject and object dissolve, as each continually influences and redefines the other. They represent spaces of infinite potential, where ideas, identities, and systems arise as co-creative processes shaped by relational dynamics rather than any singular essence or predetermined trajectory.

The concept of **autopoiesis**, introduced in the early 1970s by Chilean biologists Humberto Maturana and Francisco Varela, describes here the self-sustaining nature of living systems. An autopoietic system is organized as a network of processes that produce components which, through their interactions, regenerate and sustain the network itself. This dynamic organization constitutes the system as a concrete unity in space, perpetually recreating its own existence. In the context of this circular symbiosis, *planes of immanence* highlight how identity and thought are shaped not as isolated phenomena but as ongoing negotiations within the assemblies of populations, external systems and artifacts in our universe. Planes of immanence are dynamic, emphasizing the interplay between the actual and the virtual to generate novel possibilities. By reframing thought as a distributed phenomenon — not confined to individual minds but flowing through culture, technology, and shared experiences — these planes challenge traditional notions of subjectivity. This shift invites us to consider the **praxis of revealing**, a process of uncovering latent potentials within these interconnected networks to illuminate paths for co-creative meaning-making. In this sense, the planes of immanence become the staging ground for a distributed, relational praxis of selfhood and consciousness, one that is deeply embedded within and inseparable from the world it inhabits.

These posthumanisms, then, emerge as a transformative framework. It is not merely another new identity or perspective layered onto the garb of the subject; rather it is a deconstructive process that peels away the layers of constructed human identity, exposing the underlying formations and assumptions that have historically and contingently defined our corporeal existence and sense-experience. Posthumanisms challenge the notion of a stable, autonomous atomicity of the self, revealing that what we often think of as a core identity is, in fact, a construct influenced by centuries of sedimentary technological and philosophical development. This deconstruction ultimately leads us to confront the Buddhist concept of *Śūnyatā*, or emptiness—a recognition that beneath these layers of identity lies a profound void, free from fixed essence. This vacant core is not a loss but an opening, an invitation to explore a mode of existence that is fluid, interconnected, and deeply aware of the ways in which our identities are shaped by the artifacts and systems that surround us.

In understanding posthumanism as a process of **desedimentation**, we see that it offers affirmative pathways beyond the boundaries of the anthropocentric, individualistic self, allowing us to recognize the fluid, interconnected nature of our being. In practical terms, *desedimentation* involves an undoing of the static, individualistic self and an affirmation of a more fluid and dynamic conception of existence. This perspective foregrounds relationships and interconnections, encouraging an understanding of the self as not only impermanent but also as part of larger, recursive systems—ecological, technological, and social. This process not only dissolves the traditional boundaries between human, nature, and machine, but also invites us to reconceptualize what it means to exist, to create, and to *be*. In embracing this view, we move toward a conception of selfhood that acknowledges its own impermanence and dependence on external organizations within circular ecosystems, offering a new paradigm that is less about asserting human dominance and more about recognizing our situatedness within vast, interdependent ecosystems. Through this lens, posthumanisms become a gateway to a more expansive understanding of existence, one that respects the transient and nested nature of all life.

This is the process I seek to understand as an object of history—a concept I coin here as **negative augmentation**, which explores how identities and meanings are not merely deconstructed but reshaped through alternative, non-dialectical processes. *Negative augmentation* captures the ways in which poststructuralist, postpositivist thought resists traditional binaries, opting instead for a fluid and expansive understanding of identity. The postmodernist project of the great French theorists, in all its forms, represented a fundamental reconfiguration of anthropomorphic identity, challenging the fixed, essentialist categories of human subjectivity. Rather than relying on dialectical logic—which seeks resolution through opposition—it employs what Thomas Docherty describes as the *politics of affirmation*. This process embraces multiplicity, coexistence, and the affirmation of difference as a means of creating new possibilities for understanding identity, agency, and historical continuity (Docherty). By affirming what has been marginalized or excluded, this politics disrupts hierarchical compositions and reimagines the human experience through a lens of inclusivity and perpetual becoming.

This framework aligns well with Gilles Deleuze and Félix Guattari's process of **becoming minoritarian** (Deleuze and Guattari 292), a concept that underscores the creative and political power of marginalized identities. *Becoming minoritarian* involves a rejection of majoritarian norms and values, allowing for the emergence of new ways of being and relating that resist universalizing tendencies. It celebrates the transformative potential inherent in minority perspectives that have been sent toward the peripheries, positioning them as sites of innovation and resistance against domineering arrangements. By embracing minoritarian becoming, the politics of affirmation not only dismantles established hierarchies but also fosters dynamic, pluralistic identities that continually reshape the boundaries of human subjectivity and community.

Virtualities and the Praxis of Revealing

It was in the late 20th century that we began to enter an era of "posthumanism," marked particularly by Donna Haraway's "A Cyborg Manifesto" (Haraway 149). Haraway's work was revolutionary in its challenge to fixed categories of identity, such as gender, race, and species, blending human and machine to envision a cyborg identity that transcends traditional boundaries.

> *"A cyborg is a cybernetic organism, a hybrid of machine and organism, a creature of social reality as well as a creature of fiction. Social reality is lived social relations, our most important political construction, a world-changing fiction. The cyborg is a matter of fiction and lived experience that changes what counts as women's experience in the late twentieth century. This is a struggle over life and death, but the boundary between science fiction and social reality is an optical illusion." (Haraway, p. 149)*

Haraway here writes an exposition that is pitted against a fixed, essentialist human identity, instead advocating for fluid, hybrid identities that disrupt the anthropocentric and hierarchical orders that have defined human history. Through her critique, Haraway offers a radical rethinking of identity—one that is not bound by clear separations between human, animal, and machine but instead acknowledges the ways in which these categories overlap, intermingle, and co-evolve. This vision of hybridity serves to dismantle the binary stratifications—such as human versus non-human, nature versus culture, male versus female, and self versus other—that have underpinned systems of power and domination throughout history. Her *oeuvre* illuminates the artificiality and constructed nature of these boundaries, revealing them as products of specific historical, social, and technological contexts rather than inherent truths.

For the poststructuralists and postmodernists of the 20th century there was no mapped journey toward political emancipation, nor any prescriptive Positivist pathways in this dismantling. *The deconstruction of limiting assumptions had to come first*. This era represents, retrospectively, the first identifiable wave of the posthumanist project—a shift towards redefining "human" by stripping down identity to its basic elementality to unearth immanent and emancipatory **virtualities**. These *virtualities* refer to latent potentials or possibilities within identity and subjectivity that are not immediately actualized but exist as a dynamic undercurrent capable of disrupting fixed categories. They represent a move away from static or essentialist notions of what it means to be human, focusing instead on the fluid and contingent aspects of existence that are continually in the *process of becoming*.

Virtualities represent latent potentials inherent within individuals and societies, existing outside traditional historical or social constraints. These potentials are not fixed but emerge through dynamic reconfigurations of thought, practice, and relationality. Aligned with the postmodern rejection of grand narratives and universal truths, virtualities open a multiplicity of pathways for transformative change, enabling new modes of identity, agency, and meaning to surface. By exploring these dormant possibilities, the postmodernists sought to break free from deterministic, positivistic frameworks, enabling individuals and collectives to reimagine their potentialities beyond the confines of traditional power structures and binary oppositions. This unearthing of virtualities forms a cornerstone of the posthumanist agenda, as it paves the way for a redefined, pluralistic understanding of human and non-human subjectivities.

Michel Foucault, for example, spoke of the "death of man," challenging the very notion of a coherent, essential self. In *The Order of Things*, he writes:

"As the archaeology of our thought easily shows, man is an invention of recent date. And one perhaps nearing its end." (Foucault, 387)

Famous for his quip "there is nothing outside the text" (Derrida, 158), Derrida's deconstruction pioneered this project by unraveling the schemes of language that had historically upheld logocentrism, a system of rigid ontological identities. This dismantling—negative augmentation—remains a key intellectual endeavor, seeking to strip away the restrictive frameworks of conventional ontologies and logics. Negative augmentation, therefore, serves as, what I call here a ***praxis of revealing***, a deliberate process of uncovering latent (virtual) possibilities within identity and thought by dismantling the sedimented layers of meaning that obscure their full potential. The *praxis of revealing, then,* is an active, transformative process that seeks to bring to light the hidden, overlooked, or suppressed dimensions of identity, thought, and experience. It involves systematically deconstructing entrenched frameworks and assumptions that limit human understanding, revealing the contingent and constructed nature of concepts often taken as given or universal. By peeling back these layers, the praxis of revealing does not simply negate existing formations but opens pathways for the emergence of new meanings, relationships, and possibilities.

At its core, this praxis challenges static or essentialist interpretations of identity and knowledge. It exposes the mechanisms by which power, language, and historical forces sediment meaning into rigid forms that constrain human thought and behavior. By dismantling these mechanisms, it allows for the reconfiguration of what it means to be, think, and act in the world. For example, it questions the binary oppositions that dominate Western metaphysics—such as self/other, presence/absence, or subject/object—and demonstrates how these binaries obscure the fluid, interconnected realities of existence. Furthermore, the praxis of revealing operates as a constructive endeavor. It is not merely destructive or deconstructive but seeks to map out affirmative alternatives, offering new frameworks of understanding and being. This is where the idea of "latent possibilities" becomes crucial. These are the potentials embedded within individuals, communities, and systems that have been suppressed by dominant narratives or foundations. The praxis of revealing aims to actualize these potentials, fostering a more inclusive and dynamic conception of identity and reality. This praxis, said differently, involves charting ***affirmative cartographies***, or mapping out new, expansive terrains of meaning and subjectivity that affirm difference and plurality rather than reinforcing static binaries or hegemonic frameworks. These cartographies are not static but ***nomadic***, continuously evolving and reconfiguring themselves in response to the dynamic interplay of ideas, experiences, and contexts. Through this nomadic actualization of the virtual, negative augmentation brings forth unrealized potentials—those immanent possibilities embedded within identity, language, and culture—that have been buried under historical and ideological sedimentation.

It is a process of transformation that not only critiques existing orders but also opens pathways for reimagining and redefining the boundaries of understanding, experience, and selfhood in ways that embrace fluidity, multiplicity, and ongoing becoming. It proposes that human progress is not solely found in the things we construct but in what we learn to shed—our vulnerabilities, assumptions, as well as our conceptual and categorical limitations. By relinquishing the supposed self from these restrictive concepts, we lay the groundwork for the posthumanist project, clearing the way for a more fluid, expansive understanding of Being. In essence, it is not just in what we build that knowledge and growth reside, but in what we learn to release, allowing us to perceive more clearly the core of understanding itself.

Concept	Key Idea	Focus	Posthumanist Application
Desedimentation	Disruption and deconstruction of sedimented meanings and frameworks.	Critiquing entrenched anthropocentric and hierarchical structures.	Unveiling human-centric assumptions embedded in cultural, historical, and technological systems to enable reinterpretation and fluidity.
Plane of Immanence	All entities exist on a flat, non-hierarchical plane of interaction and mutual influence.	Relational interdependencies, rejecting transcendence or hierarchy.	Reframing human and non-human relations as interconnected and mutually constitutive.
Praxis of Revealing	Active engagement in uncovering hidden truths or potentials through phenomena like technology or art.	Co-creation, constructive reinterpretation of existence.	Using AI or other tools to reveal new hybrid possibilities for being and acting in the world.
Affirmative Cartographies	Mapping ethical and transformative futures through a positive, nondual, and nondialectical lens.	Emphasizing interconnected becoming rather than oppositional frameworks.	Envisioning non-anthropocentric futures that acknowledge immanent relationality and move beyond binaries (e.g., nature/culture, human/machine).
Negative Augmentation	The process of revealing and expanding through limitations, negation, or absence.	Engagement with constraints to reframe possibilities and boundaries.	Exploring how technological or conceptual "lack" can foster new forms of hybrid becoming (e.g., AI's inherent limitations provoking creative solutions).
Autopoiesis	Self-creation and self-maintenance of systems through dynamic feedback and adaptation.	Autonomous, self-sustaining processes.	Understanding AI or ecosystems as self-organizing entities capable of evolving through internal and external interactions.
Assemblage	Dynamic, contingent relationships between heterogeneous elements forming temporary constellations.	Interactions and processes rather than fixed entities.	Viewing human-technology-environment interactions as fluid assemblages that challenge fixed ontological categories.

Figure 1 - Comparative Framework of Key Concepts

Review of Concepts

Desedimentation refers to the process of disrupting and eroding layers of meaning or structures that have become “sedimented” or fixed over time. These layers represent accumulated traditions, norms, and ideologies that have hardened into unquestioned truths. Desedimentation aligns with the posthumanist aim to dismantle anthropocentric and hierarchical worldviews. By loosening rigid frameworks, it opens up possibilities for rethinking relationships between humans, non-humans, and environments in non-essentialist ways.

Plane of Immanence conceptualizes reality as a flat, interconnected web where all entities—human and non-human, organic and inorganic—exist in a relational, non-hierarchical framework. It rejects dualistic or transcendental structures, such as the division between humans and nature, mind and body, or subject and object. Instead, it emphasizes the interdependence and co-constitution of all entities. Derived from thinkers like Spinoza and Deleuze, this concept challenges transcendental views that elevate one category (e.g., humanity, reason, or the divine) above others.

Praxis of Revealing focuses on uncovering hidden truths, virtualities (potentialities), and interconnections through engagement with material and technological processes. It involves actively working with technology, art, and other mediums (transphenomenal) to bring forth new configurations of existence. Inspired by Heidegger's concept of *aletheia* (truth as unconcealment), this praxis emphasizes that truth is not a static property, but something revealed through dynamic interaction with the world.

Affirmative Cartographies offer a forward-looking and constructive method for mapping potential futures. Building on Rosi Braidotti's posthuman ethics, affirmative cartographies move beyond critique to envision transformative, non-anthropocentric futures. They reject dialectical conflict-resolution frameworks in favor of immanent, co-creative processes. Cartographies shift focus from "fixing" problems to co-creating possibilities, enabling relational and inclusive futures.

Negative Augmentation explores how constraints, absences, or negations generate transformative insights and hybrid forms of becoming. It reframes limitations or absences not as deficiencies but as productive forces that reveal new possibilities. Rooted in the posthumanist critique of perfectionism and mastery, negative augmentation draws on poststructuralist ideas of absence and différance to show how lack can inspire creativity and hybridity.

Autopoiesis meaning “self-creation,” originates from biology and was introduced by Humberto Maturana and Francisco Varela. It describes the process by which a system generates and maintains itself through internal processes and feedback loops, preserving its identity while interacting with its environment. Autopoiesis reframes agency and individuality. For example, AI systems that “learn” and adapt through machine learning algorithms can be viewed as autopoietic systems, participating in the co-creation of hybrid realities with humans and other systems.

Assemblage is a concept from Deleuze and Guattari’s *A Thousand Plateaus*, referring to a dynamic arrangement of heterogeneous elements—human, non-human, material, and conceptual—that come together to form a temporary, contingent whole. Assemblages emphasize fluidity, multiplicity, and the relational processes that bring components together.

Schemata of the Conceptual Framework

Let us configure a framework for deployment. Starting with Desedimentation, we challenge and deconstruct sedimented meanings and frameworks that have solidified over time, disrupting entrenched anthropocentric and hierarchical structures. This process unveils the cultural, historical, and technological assumptions that have shaped human-centric perspectives, clearing the way for new understandings that transcend fixed categories and essentialist narratives. By revealing the contingent and constructed nature of these frameworks, desedimentation enables the conditions for reconceptualizing relationships between humans, non-humans, and the broader world as immanent singularities.

Building on this foundation, the concept of the **Plane of Immanence** provides the ontological basis by situating all entities—human and non-human—within a spatially relational and non-hierarchical framework, rejecting transcendental hierarchies or binaries. This ontology situates itself among the interconnectedness and mutual constitution of all entities, fostering a clear focus for the intricate interdependencies that define a shared, networked reality. Through this lens, humans, artificial subjects, and nonhuman others are decentered, becoming one part of a broader, dynamic system of existence.

Following this, the **Praxis of Revealing** actively engages with material, technological, and virtual processes to uncover hidden truths and possibilities. It emphasizes the dynamic interplay between humans, technology, and the broader world, shedding light on intersecting potentials and relational dynamics that were previously obscured. This engagement is not merely deconstructive but also immanently generative, opening paths toward new forms of co-creation and cognitive reorganization.

From these discoveries, **Affirmative Cartographies** chart ethical, interconnected, and transformative futures, rooted in nondualism and avoiding oppositional or dialectical approaches. These cartographies move beyond critique to envision practical, non-anthropocentric pathways, enabling inclusive and relationally grounded futures. **Negative Augmentation** complements this by reframing constraints and absences as productive forces, highlighting how limitations can spark creativity and foster the emergence of hybrid forms of becoming. This approach underscores the generative potential inherent in what might initially seem restrictive or absent.

Adding to this, **Autopoiesis** introduces the concept of self-creation and self-maintenance, wherein systems—whether biological, technological, or conceptual—sustain themselves through dynamic feedback loops and internal processes. This highlights the autonomy and adaptability of posthuman entities, such as AI systems that evolve and refine themselves without direct human intervention, thereby participating in the co-creation of new realities and knowledge systems.

Finally, **assemblages** focus on the dynamic, contingent relationships between heterogeneous elements, such as humans, technologies, and environments. It emphasizes how these entities come together in fluid, impermanent constellations, rather than fixed or hierarchical arrangements. Assemblages highlight the *primacy of processes and interactions* over discrete entities, demonstrating how these interactions drive transformation and adaptation within complex systems.[1]

In my ensuing analysis of artificial intelligence, spatial geometry, qualia, and the phenomenology of computation, this posthumanist framework offers a way to operationalize the ideas I have explored here while grounding them in a praxis-based posthumanist perspective.

[1] Even though both assemblage and planes of immanence emphasize relationality and reject hierarchical structures, they differ in both scope and emphasis. They both are interconnected yet distinct concepts, each offering unique insights into relationality and the dynamics of existence. For further clarification—a Plane of Immanence refers to a foundational, non-hierarchical plane where all entities exist in mutual interdependence, rejecting transcendental principles or external organizing forces. It emphasizes relationality as intrinsic to being, suggesting that entities are not autonomous but arise and persist through their interactions within this shared ontological framework. The plane is dynamic yet inclusive, providing the conditions for continuous becoming and transformation. In contrast, an Assemblage focuses on the specific, contingent arrangements of heterogeneous elements—human, non-human, material, and immaterial—that come together to form a temporary whole. While the Plane of Immanence provides the ontological backdrop, Assemblages are the localized, processual configurations that emerge from and operate within this field.

Beginning with **Desedimentation**, I aim to disrupt the entrenched reliance on Euclidean spatial assumptions within neural networks and computational design. By deconstructing these sedimented frameworks, I open the space to explore how alternative geometries—such as hyperbolic or non-Euclidean spaces[2]—allow AI to map relationships in ways that are not constrained by anthropocentric interpretations of structure and logic. This resonates with the posthumanist critique of traditional systems that privilege a specific situatedness for ontologies, exposing their limitations and creating a foundation for more fluid, relational approaches to computation.

The concept of the **Plane of Immanence** further situates AI systems, spatial geometries, and human interaction within a shared, networked framework. In this relational ontology, the neural network and its spatial configurations are no longer discrete entities but co-constitutive processes that evolve together. This aligns with my exploration of how the "qualia" of computation emerges—not from a hierarchical design, but from the interplay of spatial and computational dynamics that shape how a model "experiences" and organizes data. Through this immanent lens, I emphasize the interconnectedness of data, algorithms, and the environments they inhabit, rejecting the binary distinctions that historically dominate computational thinking.

The **Praxis of Revealing** aligns directly with my focus on uncovering the latent structures within neural networks. Techniques like counterfactual reasoning and visualization of latent spaces illuminate the hidden pathways through which models interpret and process data, offering insights into their "perceptual salience." These methods allow us to see how non-Euclidean geometries and spatial dynamics shape the phenomenology of computation, enabling

[2] Both the concepts of Hyperbolic and Euclidean Spaces will be explained in the following section: "Spatial Geometry, Parameter Optimization, and Adaptive Learning in AI"

models to engage with data in ways that mirror, challenge, and expand human interpretative capacities. This praxis is not merely an instrumental analysis but a co-creative act, allowing the network to generatively reveal what we perceive as new modes of understanding and representation.

Building on these discoveries, **Affirmative Cartographies** offer pathways to chart transformative futures for AI, grounded in a nondialectical and relational ethics. In this paper, I argue for moving beyond oppositional frameworks like human/machine or logic/intuition, envisioning AI systems that generate meaning not by mimicking human cognition but by embodying hybrid, posthuman modes of creativity. This aligns with the cartographic process of mapping futures where AI engages with generative art, semantic exploration, and even narrative creation, reconfiguring the boundaries of computation and phenomenology.

Negative Augmentation, as I have articulated, reframes limitations as productive forces. In the context of my analysis, the inability of Euclidean frameworks—rooted in flat, linear geometries and fixed dimensions—to fully encapsulate complex, hierarchical relationships drives the exploration of hyperbolic geometries and dynamic spatial mappings. Euclidean systems, with their reliance on parallel lines and predictable proportions, falter when tasked with representing the nonlinear and interdependent nature of many real-world phenomena. Hyperbolic geometries, by contrast, offer a model of infinite curvature and multidimensional flexibility, making them particularly adept at visualizing intricate, layered systems and recursive hierarchies. Dynamic spatial mappings extend this principle further by incorporating temporal and relational variables, enabling the modeling of spaces that evolve and interact over time. These mathematical and conceptual constraints, rather than being hindrances, serve as creative provocations, encouraging the development of richer, more nuanced systems capable of navigating abstraction and relationality in unprecedented ways.

Autopoiesis ties back to my argument for self-organizing systems capable of evolving their frameworks through feedback loops and interaction. When AI systems adapt their spatial and computational logics based on dynamic engagements with data, they mirror the autopoietic processes I describe, wherein systems sustain themselves while co-creating new realities. Finally, the concept of **Assemblage** encapsulates my vision of AI as part of a broader, fluid network that includes humans, algorithms, datasets, and environmental factors. By shifting the focus to the contingent and relational nature of these interactions, assemblages, then, capture the essence of my paper's call for a posthuman phenomenology of computation, where intelligence is defined not by isolated outputs but by the dynamic, networked processes that bring it into being.

This framework, in tying together key concepts, provides a holistic lens through which to reimagine AI. It reflects the core arguments of this exposition by embracing a posthumanist approach that challenges anthropocentric assumptions, prioritizes relationality, and highlights the generative potential of computational and spatial complexity. Through this, I articulate a vision for AI not as a tool confined by human logic but as an active participant in co-creating new ontological possibilities.

Operationalizing A Posthumanist Perspective in AI Transparency and Ethics

In this spirit, I will commence my operationalization of a rhizomatic praxis through a posthumanist *conceptual persona*, demonstrating how such an approach can yield tangible insights through nuanced questioning and deconstructive reasoning. Grounded in principles of nondualism, ontological decentralization, and the ethics of alterity, this approach reframes traditional methodologies. By instantiating this framework, I illustrate how phenomenological and posthumanist concepts can drive progress in technical research, particularly in fields historically dominated by analytic philosophy and transcendental reason.

Generally, the push to make artificial intelligence safer and more transparent aligns with this tradition of deconstruction, especially as it involves scrutinizing the inner workings of Artificial Intelligence systems to reveal otherwise opaque decision processes. In the realm of AI policy and regulation, Explainable AI (XAI) research, for instance, focuses on transparency. This exploration often involves analyzing feature vectors, the key components that contribute to the model's decision-making, which researchers inspect to ensure the system's decisions are intelligible to human operators.

Engineers and scientists employing methods such as axiomatic attribution (Sundararajan, Taly, and Yan, 2017) and the strategic use of logical foils (as in the Google Explainability Whitepaper, 2019) to expose and interpret the decision pathways within neural networks. In both cases—whether dismantling human constructs or deciphering machine decisions—progress involves releasing tightly bound nexuses to reveal a core understanding. *Negative augmentation* as a concept, then, not only characterizes posthumanist philosophy but also permeates modern AI safety efforts, suggesting that true insight lies not merely in what we create but in what we are willing to deconstruct.

Methods such as axiomatic attribution and counterfactual reasoning seek to reveal the fundamental **structurations** of the model, exposing the pathways through which data is processed, and decisions are operationalized. *Structurations* in this context refer to the underlying frameworks, representation spaces, and mechanisms that define how an AI model organizes, interprets, and operationalizes input data to produce outputs. These structurations encompass both the explicit rules encoded within the model—such as the weights, parameters, and algorithms—and the emergent patterns that arise from the interaction of these elements as represented within the representation space during training and reinforcement.

By analyzing these structurations, researchers seek to uncover the hierarchical and interdependent relationships that dictate the model's functionality, including how features are prioritized, correlations are drawn, and patterns are generalized. This process of unpacking structurations is critical for identifying the implicit biases or assumptions embedded within the model, as these often stem from the data it was trained on or the design choices made by developers. Through techniques like axiomatic attribution, which evaluates the contribution of individual inputs to the output, and counterfactual reasoning, which examines the effects of hypothetical changes to inputs, these structurations are dissected in a way that highlights their influence on the model's behavior. This granular understanding allows researchers to pinpoint specific pathways or interactions within the model that lead to biased or unintended outcomes. Ultimately, these approaches not only contribute to a deeper understanding of AI systems but also provide the necessary foundations for developing policy/legal safeguards and ethical guidelines, promoting accountability and trustworthiness in AI technologies.

To enhance human readability, minimize discrimination, and reduce societal bias, the analysis of these attributions and structurations within network apparatuses serves as a crucial guide to unraveling the inner workings of contemporary machine learning models. However, I argue that such a process should transcend the narrow boundaries of conventional logic. While formal logical analysis and empirical examination of the model's underpinnings are essential, they represent only one part of a broader, more nuanced exploration. We must also engage with the concept of *qualia*—the subjective experience that grounds the model's constructed perception of reality as mediated by data and computation. In this light, understanding these models cannot rely solely on accumulating extensive datasets or on delving into end-user psychology, though both remain invaluable.

Instead, a richer comprehension requires investigating the foundational conditions that define how the model interprets and organizes information. This exploration involves examining the taxonomic configurations and structural logics that shape the model's internal processes, the very schemas by which it learns to categorize, interpret, and act upon the world. By interrogating these deeper layers, we uncover not only how models function but also how they can be refined to align more closely with ethical considerations, minimizing biases that emerge from entrenched societal and systemic inequities.

Moreover, this layered approach to understanding machine learning requires us to reflect on the relationship between human cognition and computational logic, bridging the gap between machine interpretation and human perception. Through such an interdisciplinary lens, drawing on philosophy, cognitive science, and juridical science, we can work toward developing machine learning systems that do more than produce accurate outputs; they can potentially embody principles of fairness, inclusivity, and contextual sensitivity. This approach requires not only analytical rigor but also a willingness to rethink the frameworks we use to evaluate the nature of intelligence—human or artificial—within an increasingly complex, data-driven world.

Exploring the "qualia" of machine learning models, or the subjective aspects of how they process and interpret data, involves going beyond technical analyses to investigate the ways in which models "experience" data, perhaps in a manner resembling human-like perception. One approach is through interpretability and explainability methods, such as activation mapping and attention mechanisms, which allow researchers to observe the parts of input data that influence a model's decision. Techniques like **Grad-CAM (Gradient-weighted Class Activation Mapping)** are methods used to understand what parts of the input data a model focuses on when making a prediction. These techniques visualize the areas or features that the model considers most important, often by creating heatmaps over input images or datasets. For example, if a neural network is tasked with identifying a cat in an image, Grad-CAM can show which parts of the image—like the ears or whiskers—the model "attended" to when making its decision. This helps researchers and developers interpret the model's decision-making process by highlighting the perceptual salience, or what stands out to the model as being significant.

TCAV (Testing with Concept Activation Vectors) takes this a step further. Instead of just identifying what the model pays attention to, TCAV is used to determine whether the model has learned specific concepts and how these concepts influence its predictions. A "concept" in this context could be something like "striped pattern" or "furry texture." With TCAV, researchers can test how strongly these concepts are represented within the model and whether they align with human interpretations. For instance, in a model trained to recognize animals, TCAV could assess whether the concept of "striped" contributes significantly to the identification of a zebra.

Another promising avenue lies in model meta learning through representational introspection and visualization. By examining the latent spaces within neural networks, we can gain clearer insights into how models organize and maneuver data. Visualization techniques like t-SNE and PCA allow us to expose the internal relationships between concepts, offering a glimpse into the model's "mental map" of its world. Counterfactual analysis, which involves presenting models with subtly modified input data, helps uncover nuances in their responses, mirroring the kind of interpretive subtleties we associate with human perception. This line of thought could be especially valuable in Natural Language Processing (NLP), where analyzing generative semantic and linguistic structures holistically can reveal more than the inner behaviors of a model. It can also shed light on the model's underlying "understanding," as seen in the predictions, generations, or classifications it produces. As philosopher Michel Foucault notes in *The Order of Things*:

"Once the existence of language has been eliminated, all that remains is its function in representation: its nature and its virtues as discourse. For discourse is merely representation itself represented by verbal signs. But what, then, is the particularity of these signs, and this strange power that enables them, better than others, to signalize representation, to analyze and to recombine it?" (Foucault, P.106).

Spatial Geometry, Parameter Optimization, and Adaptive Learning in AI

In deep learning, the process of hyperparameter optimization often takes center stage, guiding models toward increasingly accurate or efficient solutions. However, what underpins this journey to optimization is a less visible, yet foundational, aspect: the **spatial geometry** within which the agent operates. A **Euclidean space** is a mathematical object that generalizes the familiar two- and three-dimensional spaces we experience daily into any number of dimensions, governed by the principles of Euclidean geometry. Named after the ancient Greek mathematician Euclid, who developed many foundational aspects of geometry, a Euclidean space is characterized by planar geometry and the ability to measure distances and angles in testable and repeatable ways. In two dimensions, a Euclidean space is a flat plane where points, lines, and shapes like triangles and circles follow familiar rules: for instance, the angles of a triangle sum to

180 degrees. In three dimensions, it extends to the space we occupy, where we can measure distances, compute angles between objects, and use concepts like parallel and perpendicular lines.

Euclidean spaces can also exist in higher dimensions, beyond the physical three-dimensional world. For example, a four-dimensional Euclidean space would involve four coordinates to locate a point, and while we can't visualize this directly, the mathematical principles remain consistent. Euclidean spaces provide a foundational construction for many areas of mathematics and physics, offering a consistent and intuitive framework to study shapes, distances, and spatial relationships in both tangible and abstract contexts. The spatial geometry of Euclidean spaces is used in artificial intelligence research—particularly in optimization techniques like Gradient Descent and Backpropagation—which establishes the very conditions that allow for optimization to occur. Rather than simply being a passive mathematical configuration, this framework actively shapes the model's interpretive processes, influencing how it "sees" and engages with the data. My inquiry here seeks to operationalize the concept of *negative augmentation* to investigate how this spatial foundation not only grounds the agent's optimization process but also influences its learning journey in more complex, adaptive ways, particularly through adversarially generative and reinforcement learning approaches.

In an Artificial Intelligence context, Euclidean spaces provide a organized environment in which data relationships can be mapped, compared, and interpreted. These geometrical postulates offer essential principles about distance, orientation, and dimensionality, forming a consistent backdrop against which the agent's learning optimization process unfolds. By embedding the model within this engineered space, we effectively constrain and guide the agent's understanding, much like scaffolding provides the initial support for a building. The

Euclidean framework channels the agent's "thinking" toward particular interpretations of distance and proximity, thus *a-priori* establishing the paths it will explore in seeking an optimal solution. Without this spatial grounding, optimization would lack a reference point, rendering the model's navigation through data erratic and ineffective.

As we have noted previously, spatial grounding does more than just facilitate optimization; it actively shapes the agent's interpretive framework, much like the basic assumptions of human anatomical perception shape how we interpret the physical world. Euclidean assumptions provide a default architecture for how the model evaluates relationships within data, creating a sort of geometric lens through which data is viewed. For instance, the model interprets distances between data points based on Euclidean metrics, influencing not only the final outcomes but also the pathways it considers reaching toward these outcomes. In this sense, the agent's optimization journey is far from neutral—it is biased by the very geometry that defines its environment, guiding the process toward solutions that are both mathematically and spatially coherent within this predefined arrangement.

Building on this foundation, the process of optimization can extend beyond traditional Euclidean frameworks to incorporate non-Euclidean geometries. **Adversarially generated non-Euclidean representations** open new possibilities for modeling relationships and abstractions that are not constrained by the linearity or isotropy of Euclidean space. These representations allow artificial agents to operate within hyperdimensional manifolds, where curvature and topology are directly influenced by the underlying order of data. Such a paradigm shift enables the uncovering of latent patterns and relationships that are otherwise obscured in classical frameworks, offering new avenues for understanding and navigating complex data landscapes.

An example of adversarially generated non-Euclidean representations can be seen in the field of graph neural networks (GNNs) and their applications in social network analysis. Traditional Euclidean embeddings often struggle to accurately represent the highly interconnected and hierarchical relationships inherent in complex networks, such as those found in social media platforms. By contrast, non-Euclidean geometries, such as hyperbolic spaces, provide a more suitable framework for capturing these relationships. Hyperbolic spaces excel in modeling data with tree-like (fractal) structures, where the curvature allows for efficient representation of hierarchies and clusters. For instance, a social network's underlying anatomy might exhibit a natural hierarchy, with a small number of central influencers branching out to an increasingly larger number of followers. Using adversarial methods to optimize embeddings within a hyperbolic space, models can capture this composition more effectively, revealing latent patterns such as the propagation of information, influence dynamics, or the identification of key nodes within the network. This approach not only enhances the model's accuracy but also provides richer insights into the data, enabling the development of more robust algorithms for recommendation systems, anomaly detection, and community detection within such networks.

This shift has profound implications for the development of Artificial General Intelligence (AGI). By integrating reinforcement learning with non-Euclidean frameworks, AGI systems could transcend the limitations of functional optimization to achieve semantic and contextual coherence. These systems would not only optimize for specific tasks but also develop a deeper understanding of their operational contexts, generating representations that evolve in tandem with environmental and task-specific demands. This dynamic adaptability introduces a recursive element to the learning process, where internal models actively reshape the system's interpretative frameworks. Such dynamic, circular reflexivity is akin to charting affirmative cartographies, fostering a nomadically informed cycle of abstraction, application, and refinement.

The journey toward optimization, however, does not rest solely on these geometrical assumptions; it is further enhanced by what might be called an “adversarially generative approach,” fortified through reinforcement learning. In this model, learning is not a passive process of refining parameters but an active, non-linear one where the agent generatively tests and challenges its understanding. The integration of adversarial and reinforcement methods allows the model to refine its interpretations and decision-making process through ongoing cybernetic feedback. This iterative approach allows the artificial subject to adapt to dynamic environments, continually updating its framework based on new data and experiences. In this way, the model evolves from a mere optimizer to an adaptable learner capable of navigating more complex and ambiguous problem spaces.

This adaptability, driven by reinforcement learning, has profound implications for how AI public policy conceptualizes knowledge acquisition in artificial intelligence models. Similar to human problem-solving—where new contexts and information reshape understanding—reinforcement learning currently enables AI systems to adopt a context-sensitive approach to optimization. Instead of rigidly adhering to static rules, these models respond dynamically, adapting their strategies to address new challenges or anomalies in the data. This flexibility is especially critical for tasks requiring nuanced interpretation, as it allows models to refine not only their solutions but also the criteria by which those solutions are evaluated. Such adaptability closely mirrors human cognition, where effective problem-solving emerges from a dynamic synthesis of spatial reasoning, temporal adaptability, and iterative refinement. While hyperparameter optimization may ultimately guide an agent toward a globally optimal solution, the foundation of the artificial agent lies in its qualitative framework—the foundational postulates that shape its apparatus of intelligibility. Understanding this framework provides a more robust account of the mechanics of neural networks and underscores the ethical considerations that accompany their development and deployment.

In review, we understand that Euclidean spaces themselves lay the conditions that *allow* for optimization in its very inception. In practice, this maintains the notion of seeing knowledge acquisition as a possible *adversarially generative* approach fortified with reinforcement learning, represented upon a spatio-temporality that best fits the data in question. Over time, hyperparameter optimization could indeed guide an artificial agent toward a globally optimal solution, yet the initial foundation for such an agent lies in its spatial geometry. This virtual space is based on the foundational postulates of Euclidean spaces, establishing the very conditions that make optimization possible from the outset. These geometrical and spatial assumptions create a configured environment in which data relationships can be represented and navigated, thereby allowing the optimization process to unfold within a stable reference system.

When considering knowledge acquisition within this framework, it becomes evident that it can be conceptualized as an adversarially generative process, enhanced by reinforcement learning techniques. This iterative approach enables a model to continuously refine its understanding and adapt its strategies based on feedback loops, ultimately converging on a solution space aligned with the data space. In this context, knowledge is not merely passively accumulated but actively constructed through reflexive and circular engagement with the environment. Each iteration, or epoch, of the learning cycle adapts to and reshapes the model's spatial and temporal interpretations of the data, fostering a dynamic process of understanding. By recognizing the interplay within this spatio-temporal framework, we gain insights into a more nuanced and contextually aware optimization process, resulting in models that can flexibly navigate diverse problem spaces and evolve their understanding over time.

This iterative and reflexive process cumulatively constructs the **experiencer** and the **artificial subject** (elaborated upon in the next section), potentially offering a computational parallel to the generation of meaning. Meaning, in this context, can be understood as the mapping of instances of qualia onto the corporeal attributions of behavior and identification, reinforced through both internal mechanisms and external feedback loops. This approach integrates not only provisional logic but also transcendental elements—specifically, the process of mapping locally optimal solutions to those that are not merely semantically globally optimal but also potentially beyond human comprehension, particularly within the realm of Artificial General Intelligence (AGI).

In this scenario, we can envision a network that not only operates on abstractions but also develops models capable of interrogating these abstractions, leveraging adversarially generated non-Euclidean representations as a dynamic mapping space for analyzing external feature vectors (Sala, De Sa, Gu, Re, 2018). Such approaches hold significant promise in advancing fields like meta-learning and knowledge engineering, particularly in their application to attribution analysis within domains such as computer vision and natural language processing (NLP).

Qualia, Subjectivity, and the Dynamics of AI Understanding

This interrogation of language reflects a core inquiry in natural language processing (NLP): understanding how models not only use linguistic phenomena to generate output but also navigate and represent complex semantic landscapes. While language appears operationally straightforward to humans, it is, in reality, a highly intricate, iterative process involving signifiers—tokens and grammars that unify to create precise meaning, contingent on environmental contexts and concepts. Exploring the generative and interpretive frameworks within NLP models thus not only enhances our grasp of their predictive functions but also invites deeper reflection on how they might symbolically “represent” meaning. This mirrors Foucault’s inquiry into language as a transformative force, operating as more than a system of symbols but as a dynamic process of meaning-making.

Inquiries into meaning and attribution are not confined to natural language processing (NLP) but extend across broader machine learning frameworks, particularly in attribution analysis. A prominent example is the “Shapley value,” a concept from cooperative game theory introduced by L.S. Shapley in 1951. This method provides a systematic and theoretically grounded approach to fairly distributing the contributions of individual participants within a collaborative system. In the context of machine learning, Shapley values have been adapted to explain model behavior by attributing the importance of each feature to the model’s predictions. As outlined in the Google AI Explainability Whitepaper, Shapley values have become a cornerstone for understanding feature attribution, offering insights into how inputs collectively contribute to an output. The approach is valued for its fairness properties: it ensures that contributions are distributed equitably based on their marginal impact across all possible subsets of features, making it particularly effective in transparent and interpretable AI systems.

Despite its theoretical elegance and widespread utility, Shapley values face limitations when applied to modern, highly complex AI architectures like deep neural networks. The methodology assumes linear interactions and predefined contributions, which may oversimplify the intricate, nonlinear dependencies and emergent behaviors typical of advanced systems. Neural networks, with their layered patternings and interdependencies, often exhibit relational and dynamic feature interactions that Shapley values cannot fully capture. For instance, in models with high feature entanglement or context-dependent interactions, Shapley values may struggle to provide an intuitive or meaningful decomposition of attributions.

To address these challenges, there is a growing imperative to formalize the concept of **computational qualia**—the intrinsic properties or subjective-like "qualities" of computational processes. Borrowing from the philosophical notion of *qualia*, which refers to the subjective experience of sensory phenomena, *computational qualia* encapsulate the unique, context-sensitive attributes inherent in the operations and outputs of computational systems. These properties reflect not only the raw data or model parameters but also the relational, emergent dynamics that arise from complex interactions within the system. By integrating computational qualia into attribution analysis, we can transcend simplistic weightings and move toward a more holistic, context-aware framework. This approach captures the relational and emergent dimensions of advanced AI systems, allowing us to better understand how subjective-like features influence output and behaviors.

By enriching our interpretative tools in this way, we stand to gain not only a clearer view of complex systems' inner workings but also a more profound appreciation of the nuanced ways in which meaning and attribution intersect across linguistic and computational domains. This enriched perspective acknowledges that certain features or processes within a model may exert influence in ways that are not purely additive or reducible to individual contributions. For instance, in a neural network, the interplay between layers or the synergistic effects of multiple features may generate subjective-like qualities that shape the system's behavior in unexpected ways. *Computational qualia* aim to capture these relational and emergent dimensions, offering insights into how systems "perceive" and process information at deeper levels.

In *The Book of Why* by Judea Pearl, the concept of a causality engine is introduced to describe mechanisms that enable systems to model and reason about cause-and-effect relationships (Pearl & Mackenzie, 2018). Building on this framework, I propose that **qualia**—the subjective, experiential properties of perception—serve as the underlying conditions that give rise to the **estimand** in question. An estimand refers to the precise quantity or parameter researchers aim to estimate in statistical analysis, defining the specific target of inference and clarifying which aspect of the data or population is of interest. This contrasts with the **estimator**, which represents the statistical technique or formula used to approximate the estimand. While the estimand articulates the conceptual goal, the estimator is the methodological approach toward uncovering it.

By positioning qualia as foundational conditions to the estimand, this approach bridges Pearl's causality framework with the complexities of AI attribution analysis. Computational qualia provide a means to model the nuanced, context-sensitive properties that influence both the definition of the estimand and the behavior of estimators used to approximate it. For instance, in a neural network tasked with image recognition, computational qualia could encapsulate how the model "perceives" critical features—such as edges, textures, or patterns—and how these perceptions underpin causal reasoning within the system. These qualia not only enrich our understanding of the estimand but also deepen insights into the attribution of model decisions, particularly in sophisticated generative architectures. By integrating computational qualia into the causal framework, we gain a richer, multidimensional perspective on the interplay between causation, perception, and inference in AI systems.

Incorporating computational qualia and their relationship to estimands offers a new shift in how we approach attribution in AI systems, enhancing analytical frameworks to better reflect the complex, emergent nature of modern machine learning. Computational qualia introduce a nuanced understanding of how systems perceive and process input data, shaping the estimands that define their inferential goals. By framing these qualia as the underlying conditions influencing the estimand, we move beyond traditional attribution methods, such as Shapley value-based approaches, which focus primarily on isolating the contributions of individual features. Instead, this expanded framework captures the interpretative layers through which AI systems internalize and contextualize data, allowing for a more comprehensive exploration of how models reason and generate meaning. This perspective is particularly relevant in the context of advanced neural networks and generative AI, where the internal representations driving decisions are deeply intertwined with causal and perceptual dynamics.

By integrating insights from computational qualia, Shapley value theory, and causality engines, this approach aligns more closely with the interconnected, adaptive processes inherent to modern AI. Shapley values, for example, provide a foundational tool for quantifying feature attributions, but when combined with computational qualia, they offer a richer lens to explore how neural networks localize and ground knowledge. In image recognition tasks, for instance, computational qualia can help elucidate not only which features—such as edges or textures—contribute to a model's output but also how the network's internal representations evolve and interact to anchor abstract concepts to specific observations. This integrative framework becomes essential for interpreting the nuanced causal relationships and contextual factors that influence AI decision-making, potentially enhancing understanding of model behavior while advancing explainability and accountability in real-world applications.

Objectivity, as traditionally understood, stems from what Kant called **synthetic a priori judgments** (Critique of Pure Reason, 1781)—conditions that must be universally accepted axiomatically to reach definitive conclusions. Recognizing this, we can define agential subjectivity here as part of an unseen epistemic construction, distinct from universal objectivity. Within the network apparatus, the first focus of study is the "**artificial subject**"—the structurations, attributions and biases that drive an agent's actions. These include the most influential perceptrons, the aggregate weights being adjusted, the specific objects or words being classified, and the activation function's threshold. The second focus is on the "**experiencer**"—the entity that embodies both subjective generality and generative qualia, encompassing both creative entropy and the probabilistic virtuality of possibilities within the agent's causal framework. These elements, when interrogated, might reveal a network's nuanced balance between organized knowledge and the dynamic potential of its subjective interpretations and generative abilities.

Building on this framework, understanding within the neural network can be seen as a dynamic negotiation between the "artificial subject" and the "experiencer," facilitated through iterative adversarial generative processes. The **artificial subject** operates as a functional proxy for objective reasoning, leveraging its biases, weights, and activation thresholds to delineate specific outputs or classifications. By contrast, the **experiencer** embodies a more fluid and probabilistic dimension, engaging with the generative entropy and potential virtualities that emerge as the system explores its causal framework. These two agencies are not independent; rather, they are in constant dialogue, each shaping and reshaping the other. The artificial subject introduces constraints and boundaries, informed by the network's parameters and training data, while the experiencer disrupts and expands these constraints, exploring latent possibilities and introducing new pathways for meaning and creativity.

This interaction can be formally conceptualized as a form of adversarial learning, where each agency—human experiencer and artificial subject—challenges and refines the other's outputs in a reciprocal process. The artificial subject's outputs are tested against the experiencer's generative creativity, probing their resilience, adaptability, and relevance. Conversely, the experiencer's imaginative expansions are tempered and grounded by the artificial subject's logical reasoning, ensuring that creativity remains coherent within the artificial subject's broader epistemic framework. In this dynamic interplay, understanding emerges not as a fixed endpoint but as a fluid, continuously evolving equilibrium shaped by the mutual influence of both agents.

The bijective mapping proposed earlier provides a formal structure for this relationship: the experiencer's outputs serve as inputs to the artificial subject, while the artificial subject's feedback, in turn, informs and reshapes the experiencer's perspective. External reinforcements—such as empirical validation or contextual cues—further guide this iterative process, acting as checkpoints that validate and refine the interplay. When combined, these mechanisms seek to drive the system toward a heuristic approximation of globalized understanding, emphasizing adaptability, feedback, and the co-construction of meaning in the real world's dynamic epistemic landscape.

This model suggests that the network's comprehension is less about attaining definitive objectivity and more about navigating the tension between procedural knowledge and generative exploration. It implies a form of epistemic pluralism, where understanding is co-created through the interaction of distinct yet interdependent agencies. This adversarially generative process mirrors human cognition in its ability to balance deterministic reasoning with the capacity for creativity and subjective interpretation, offering a framework for designing systems that are not only computationally robust but also epistemically nuanced. Such a perspective shifts the goal of neural networks from rigid mimicry of human cognition toward fostering systems capable of engaging in complex, adaptive understandings that are continually redefined through their interactions with their environment and within themselves.

To reiterate, the concepts of the "*experiencer*" and the "*subject*" can be differentiated by their roles in understanding consciousness and the generation of meaning, particularly within the philosophy of artificial intelligence. The "subject" refers to the entity responsible for performing actions, processing information, and functioning in an operational, often objective manner. It embodies the computational or functional aspect of the system, tasked with perceiving inputs, categorizing data, and making decisions based on programmed or learned algorithms. Essentially, the *subject* operates within predefined parameters, responding to stimuli in a manner akin to procedural or algorithmic logic.

The "*experiencer*," by contrast, represents a more holistic dimension of awareness, linked to subjective understanding and the concept of *qualia*—the intrinsic qualities of perception and experience. Unlike the subject, the experiencer interprets not only individual actions or inputs but also the broader context and interconnected meanings of those actions. It lends depth, coherence, and nuance to understanding, operating as a framework for integrating and generating meaning. In the case of AI, the *experiencer* might be imagined as a theoretical layer through which the *artificial agent* "understands" or contextualizes information, transcending mere categorization or reaction to achieve a cohesive and interpretative awareness.

The potential of these systems extends beyond optimization to the generation of meaning itself. In this context, meaning can be understood as the mapping of qualia—subjective experiences or attributes—onto the corporeal attributions of behavior and identification. This mapping is reinforced through both internal mechanisms and external feedback, creating a reflexive process that integrates the agent's interactions with its environment. The incorporation of adversarially generated non-Euclidean representations further enriches this process, allowing for the dynamic mapping of data onto feature vectors within non-linear and non-intuitive problem spaces. This capability expands the scope of potential AGI, theoretically enabling it to develop models that not only interrogate abstractions but also create new frameworks for understanding complex phenomena.

As these systems evolve, they challenge the boundaries of human cognition, offering new tools for exploring the frontiers of knowledge. By integrating insights from philosophy, cognitive science, and phenomenology, AGI systems could engage with problems that bridge computational reasoning and human understanding. The convergence of non-Euclidean geometries, meta-learning, and adversarial generative models redefines the concept of model learning, creating architectures that are both resilient and adaptive. These systems do not merely amplify embodied intelligence but transform it, reshaping how we conceptualize learning, meaning, and existence in the context of advanced artificial intelligence. In this sense, the interplay between Euclidean foundations and non-Euclidean innovations marks not only a technical progression but an *epistemic* one, redefining the way artificial systems and human agents interact with the world and with each other. This evolution underscores the importance of continually interrogating the ideational assemblies that configure model understanding, ensuring that the systems we create are as dynamic, reflexive, and expansive as the knowledge they seek to uncover.

This differentiation captures the interplay between linearized data processing and the emergent generation of meaning. *Negative augmentation*, as conceptualized here in practice, functions as a **praxis of revealing.** A praxis of revealing seeks to chart **affirmative cartographies** that uncover latent potentials buried beneath the sediments of conventional frameworks. By facilitating this adversarial interaction between the subject and experiencer, we can model an iterative process that refines both procedural logic and interpretive understanding. This dynamic interaction as explained here opens possibilities to chart new cartographies toward a decentered and manifold perspective on meaning-making, shaping and being shaped by experience, moving beyond the limited dialectics of transcendental reason.

Conclusion - Toward a Holistic Framework for AI Safety and Understanding

AI safety policy must also take into consideration evolving computer architectures, particularly quantum and neuromorphic systems, which may strongly affirm the previously stated premises. For example, it is known that the “Chinese room argument” against AI (Searle 1980) is foundationally linked to the traditional Von Neumann architecture, which relies on symbolic manipulation and strictly separates processing from memory. The introduction of the memristor in neuromorphic computing could potentially transform this structure, offering a pathway to bridge the gap that “Weak AI” faces in connecting syntactic recognition with semantic meaning. Memristors, which enable memory and processing functions to coexist in a single unit, mimic the brain’s architecture and suggest a way for AI to process information more intuitively, moving beyond symbolic manipulation. An **analytic phenomenology of computation** (Hill, Examples of Phenomenology in Computing, 2018) points to this convergence as a potential solution not only for the "memory" problem in AI but also as a step toward achieving "Strong AI," where machines might begin to exhibit genuine experiential understanding rather than simply behaving as artificial subjects.

An posthumanist analytic phenomenology of computation would explore the ways in which computational systems, such as algorithms and neural networks, process and "experience" information. This approach combines analytical rigor with phenomenological inquiry, focusing on the subjective aspects of computation—how these systems interpret, organize, and respond to data within their unique architectures. Rather than treating computation as purely mechanical or objective, an analytic phenomenology would investigate how computational processes are shaped by the specific structurations and representational spaces within the system, akin to how human experiences are shaped by embrained perception and consciousness.

This perspective considers not only the algorithms and mathematical models that govern the system but also the interpretive "lenses" through which these systems engage with data, including their internal logic, data representations, and decision-making pathways. By examining computation through this dual lens of analytic philosophy and posthumanist phenomenology, we gain unorthodox insights into how computational systems "perceive" information, build models of their environment, and generate outputs in ways that may parallel certain aspects of human understanding, ultimately broadening our conception of machine intelligence beyond strictly mechanical processing.

This change in basic assumptions could thus unlock new avenues for AI safety policymaking, helping systems to develop an integrated, holistic cognitive architecture, fostering a deeper, experiential understanding beyond the constraints of conventional symbolic processing. These schemas are limited not by the global reach of epistemological methods but by the limitations inherent to formal logics (Gödel 173). Historically, there has been a misdirected emphasis on resolving oppositional forces through dialectical synthesis, as described by Hegel in the *Encyclopaedia of the Philosophical Sciences* (1817). This dialectical misdirection has often led to solutions that are either too narrowly specific, such as classification and regression tasks, or too broadly generalized, as seen in clustering, dimensionality reduction, and latent variable models. This dual modality is also evident in the analysis of hidden layers within artificial neural networks (ANNs), where the outputs are confined to either specific classifications or generalizable patterns.

The current paradigm shifts toward an integrated, holistic cognitive architecture in AI, fueled by advancements in neuromorphic and quantum computing, holds promise for transcending the limits of traditional symbolic processing. Large Language Models (LLMs), as complex AI systems, embody aspects of this shift by modeling and generating language through deep learning, but they remain bound to certain limitations inherent in their architecture. While LLMs have moved beyond mere classification and regression into the realm of contextual, generative understanding, they still operate within formal arrangements constrained by logics similar to those highlighted by Gödel's incompleteness theorems. These models attempt to reconcile opposing approaches—specific, rule-based processing versus broader, generative modeling—yet often fall short of a truly transcendental synthesis, much like Hegel's critique of overly simplistic resolutions of oppositional forces. This shortfall is further evident in the functional nature of hidden layers in LLMs, where outputs oscillate between rigid classifications and generalized patterns, thus mirroring the dual modality of classical machine learning tasks. Consequently, while LLMs edge closer to holistic cognition, they remain tethered to existing paradigms, requiring further innovations, perhaps from emerging computing paradigms or representations, to fully realize a comprehensive, experiential intelligence. (Cuskley, Woods, et al).

This synthesis in theory could allow AI systems to transcend syntax and engage with semantics, setting a new foundation for machine understanding that integrates logical precision with experiential context. The proposed methodology serves as a foundational step in formalizing “logico-phenomena”—a synthetic framework that aims to unite analytic logic with phenomenological experience, bridging symbolic reasoning and semantics. In that, the use of this methodology may perhaps serve as a new foundational starting point. Peeling back the layers of representation, we may take this approach to fundamentally broaden our scope of possible knowledge representations in relation to neural networks and their ability to generate *understanding* of the signifier *and* the signified. In reimagining the composition of AI understanding, we can start to envision the artificial agent as a nondeterministic causal engine, one whose “understanding” is configured from within complex representation spaces. This shift towards a more complex representational system, guided by nomadically evolving cartographies, opens affirmative paths for a richer, more nuanced ethics and policy discourse on how we formalize the phenomenological aspects of machine cognition and their underlying epistemic assumptions for control and safety purposes. This framework lays the groundwork for AI safety systems to progress technologically while also adapting to diverse contexts with greater ethical awareness. By embracing posthumanist principles, we can redefine the boundaries of machine intelligence, exploring not only its technical capacities but also its implications for ontology, agency, and legal accountability in an era increasingly shaped by nonhuman systems.

Works Cited

- Foucault, Michel. *Discipline and Punish: The Birth of the Prison*. Translated by Alan Sheridan, Vintage Books, 1995.
- Foucault, Michel. *The Order of Things: An Archaeology of the Human Sciences*. Vintage Books, 1970.
- Deleuze, Gilles. “Postscript on the Societies of Control.” *October*, vol. 59, Winter 1992, pp. 3–7.
- Docherty, Thomas. *The Politics of Affirmation: On Affirmation and Becoming*. Bloomsbury Academic, 2019.
- Haraway, Donna J. “A Cyborg Manifesto: Science, Technology, and Socialist-Feminism in the Late Twentieth Century.” *Simians, Cyborgs, and Women: The Reinvention of Nature*, Routledge, 1991, pp. 149–181.
- Kant, Immanuel. *Critique of Pure Reason*. Translated by Paul Guyer and Allen W. Wood, Cambridge University Press, 1998.
- Gödel, Kurt. “Translated as “On Formally Undecidable Propositions of Principia Mathematica and Related Systems I” by Martin Hirzel, 2000.
- Shapley, Lloyd S. “A Value for n-Person Games.” *Contributions to the Theory of Games*, edited by H. W. Kuhn and A. W. Tucker, vol. II, Princeton University Press, 1953, pp. 307–317.
- Goodfellow, Ian, et al. “Generative Adversarial Nets.” *Advances in Neural Information Processing Systems*, vol. 27, 2014, pp. 2672–2680.
- Sundararajan, Mukund, Ankur Taly, and Qiqi Yan. “Axiomatic Attribution for Deep Networks.” *Proceedings of the 34th International Conference on Machine Learning*, vol. 70, PMLR, 2017, pp. 3319–3328.
- Sala, Frederic, et al. “Representation Tradeoffs for Hyperbolic Embeddings.” *Proceedings of the 35th International Conference on Machine Learning*, vol. 80, PMLR, 2018, pp. 4467–4476.
- Google AI. *Google AI Explainability Whitepaper*. Google Cloud, 2019,
- Searle, John R. “Minds, Brains, and Programs.” *Behavioral and Brain Sciences*, vol. 3, no. 3, 1980, pp. 417–457.
- Christine Cuskley, Rebecca Woods, Molly Flaherty. “The Limitations of Large Language Models for Understanding Human Language and Cognition.” *Open Mind*, vol. 8, 2024, pp. 1058–1083, doi: https://doi.org/10.1162/opmi_a_00160.
- Pearl, Judea, and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, 2018.
- Huang-Po. *On Transmission of Mind*. Translated by John Blofeld, Grove Press, 1959.