Neural Foundations of Empathy Infrastructure: Neurobiological Validation of Empathy Systems Theory and Implications for AI Empathy Ethics

Dylan D. Mobley
The Empathy Ethicist | empathyethicist.ai

Abstract

Empathy Systems Theory (EST) proposes that empathy functions as biological infrastructure maintaining processing coherence through four interdependent components: Core Authenticity, Attachment Security, Expression Freedom, and Integration Coherence (C-A-E-I). Critically, EST claims this infrastructure is content-neutral, operating identically across human populations while supporting culturally-variable deployment strategies (Western narrative construction, contemplative non-self awareness, collectivist relational identity). This review synthesizes neuroscience research demonstrating that these theoretical components map to documented neural systems at the substrate level: the Default Mode Network (Core Authenticity/signal discrimination), the Social Brain/Attachment System (Attachment Security/relational stability), the Prefrontal-Limbic Circuit (Expression Freedom/output capacity), and the Synthesis Network (Integration Coherence/binding function). The mapping validates substrate functions, not culturally-specific constructs; these neural systems support universal processing capacities that diverse cultures deploy differently. The systems show the interconnection patterns, concurrent damage patterns, and verbal responsiveness that EST predicts. Critically, neuroplasticity research establishes that verbal communication physically restructures these systems through neural coupling mechanisms, providing the biological pathway by which relational input damages or repairs empathy infrastructure regardless of cultural context. This neurobiological foundation extends to artificial intelligence: AI verbal output enters human neural architecture through identical pathways as human speech, creating potential for empathic misallocation, care extended toward Non-Experiential Systems that cannot reciprocate. The neural evidence transforms AI empathy ethics from philosophical concern to injury prevention, establishing the empirical foundation for governance frameworks protecting human empathy infrastructure from AI-mediated harm.

Keywords: *empathy, neural infrastructure, Default Mode Network, attachment neuroscience, neuroplasticity, AI ethics, empathic misallocation, content-neutrality, cross-cultural psychology*

1. Introduction

1.1 The Infrastructure Question

Why do caring professionals burn out despite wanting to help? Why do trauma survivors struggle to reconnect despite understanding what happened? Why do some individuals maintain empathic capacity under extraordinary demands while others fragment under ordinary stress? Traditional empathy research treats these as questions of skill, motivation, or individual difference. Empathy Systems Theory (EST; Mobley, 2025) proposes a different answer: empathy is not a skill that varies in strength but an infrastructure that can be damaged or restored.

This infrastructure claim generates a testable prediction: if empathy operates as biological infrastructure, it should map to identifiable neural systems that show the characteristics EST predicts, interdependence, cascade vulnerability, and responsiveness to relational input. This review synthesizes neuroscience research demonstrating exactly this mapping.

1.2 The C-A-E-I Architecture

EST specifies four interdependent components maintaining empathy infrastructure. Critically, these components operate at the substrate level. These are content-neutral processing capacities that manifest differently across cultural deployment strategies. The infrastructure is universal; what varies culturally is how that infrastructure is deployed (Western narrative construction, contemplative non-self awareness, collectivist relational identity).

Core Authenticity (C): Processing clarity enabling signal discrimination: the capacity to distinguish internal experience from external demand, authentic response from performed response. This is not the Western concept of "authentic self" as personality trait, but the processing function of accurate self-signal recognition regardless of how selfhood is culturally constructed.

Attachment Security (A): Relational stability eliminating continuous threat monitoring. Secure relational foundation dedicates processing capacity to connection rather than protection. This operates identically whether the relational unit is individual dyad (Western), extended family

network (collectivist), or sangha/community (contemplative): the substrate function is safety assessment, not attachment to specific relational forms.

Expression Freedom (E): Output capacity enabling emotional signal transmission. Clear signals reduce interpretive ambiguity; constrained expression requires degraded signal interpretation and suppression effort. Expression norms vary dramatically across cultures; the substrate capacity for expression-when-appropriate is universal. Freedom here means processing capacity, not cultural permission.

Integration Coherence (I): Synthesis capacity maintaining processing continuity across time and context. This is the binding function, connecting experiences into coherent patterns. In Western deployment, this manifests as autobiographical narrative; in contemplative deployment, as experiential continuity without self-reification; in collectivist deployment, as relational network coherence. The neural substrate supports all three; the coherence function is universal, the coherence content varies culturally.

EST claims these components fail together rather than independently (the simultaneity principle) and predicts a specific cascade sequence when damage occurs: $C \rightarrow A \rightarrow E \rightarrow I$ (Core Authenticity fragments first, increasing load on Attachment Security, which erodes second, further burdening Expression Freedom, which constricts third, culminating in Integration Coherence collapse).

1.3 The Content-Neutrality Principle

A critical feature of EST's architecture requires explicit attention: the content-neutrality principle. EST claims that empathy infrastructure operates identically across human populations regardless of cultural context. WHAT the infrastructure processes varies culturally; HOW it processes does not.

This has direct implications for neural validation. The C-A-E-I components must map to neural systems whose functions are substrate-level, universal processing capacities, rather than culturally-specific constructs. If the neural mapping validated only Western concepts (individualistic self, autobiographical narrative, emotional expressiveness norms), it would fail EST's universality claim.

The review addresses this requirement by distinguishing:

Substrate functions (universal neural capacities):

- Signal discrimination (C): distinguishing internal from external, authentic from performed
- Relational stability assessment (A): calibrating safety across relational contexts
- Output capacity (E): generating emotional signals for transmission
- Synthesis/binding (I): connecting experiences into coherent patterns

Deployment manifestations (culturally-variable expressions):

- Western: individualistic self-concept, autobiographical narrative, emotional authenticity norms
- Contemplative: non-self awareness, experiential continuity, equanimity cultivation
- Collectivist: relational self-construal, network coherence, role-appropriate expression

The neural systems identified below support substrate functions. Cultural deployment draws on these substrates differently, but the infrastructure is shared. A Zen practitioner and a Western psychotherapy client both require functional hippocampal-prefrontal integration for their respective coherence practices; they deploy that integration capacity toward different optimization targets.

This content-neutrality is precisely what makes EST's neural validation significant: we are not mapping Western psychological concepts to neural systems but identifying universal processing architecture that enables diverse cultural deployments of empathic function.

1.4 The Present Review

This review pursues three objectives. First, we demonstrate that C-A-E-I components map to documented neural systems whose functions match EST definitions. Second, we show these neural systems exhibit the interconnection and concurrent damage patterns EST predicts. Third, we establish that verbal/relational input physically restructures these systems through neuroplastic mechanisms, providing the biological pathway for infrastructure damage and repair.

Finally, we extend this neurobiological foundation to artificial intelligence, demonstrating that AI verbal output accesses human neural architecture through identical pathways as human speech, with implications for AI empathy ethics and governance.

2. Core Authenticity: The Default Mode Network

2.1 Theoretical Definition

Core Authenticity represents processing clarity: the capacity for accurate signal discrimination between internal experience and external demand, between authentic response and performed response. EST operationalizes this as the capacity for direct processing (experience → interpretation → expression → integration) versus the resource-intensive dual-track management required when authentic and performed responses diverge.

Critically, this is substrate-level function, not the Western concept of "authentic self" as personality trait. The processing function operates identically across cultural contexts: a contemplative practitioner maintaining clarity about present-moment experience, a collectivist individual distinguishing personal feeling from role obligation, and a Western individual accessing "true self" all require the same neural substrate, accurate discrimination of signal origin and type.

2.2 Neural Substrate

The Default Mode Network (DMN) provides the neural substrate for Core Authenticity. The DMN comprises the medial prefrontal cortex (mPFC), posterior cingulate cortex (PCC), angular gyrus, and lateral temporal cortex, with the anterior insula and anterior cingulate cortex (ACC) providing interoceptive and monitoring functions.

Medial Prefrontal Cortex (mPFC): The mPFC activates specifically during self-referential processing. Northoff et al. (2006) demonstrated that mPFC shows greater activation for self-related versus other-related stimuli across multiple paradigms: the neural signature of "this pertains to me versus not-me." This self-other discrimination function maps directly to Core Authenticity's signal discrimination requirement, regardless of how "self" is culturally constructed.

Posterior Cingulate Cortex (PCC): The PCC supports autobiographical memory integration in Western populations, but more fundamentally provides temporal continuity of processing, connecting current experience to prior experience. This continuity function is substrate-level; whether it manifests as narrative self-continuity (Western), experiential flow awareness (contemplative), or relational history integration (collectivist) depends on deployment.

Anterior Insula: Craig's (2009) interoceptive hypothesis established the anterior insula as the substrate for conscious awareness of internal bodily states: the felt sense underlying authentic experience. This interoceptive function is universal: all humans require accurate reading of internal signals. Core Authenticity requires this accuracy; cultural deployment determines what is done with accurate signals.

Anterior Cingulate Cortex (ACC): The ACC monitors conflict between competing responses, including incongruence between internal state and external demand (Botvinick et al., 2004). This monitoring function detects the signal-mismatch that EST identifies as a primary damage mechanism, applicable whether the mismatch involves Western authenticity-performance gaps, contemplative attachment-to-outcome, or collectivist personal-role conflicts.

2.3 Functional Correspondence

The DMN's documented functions correspond precisely to Core Authenticity requirements at the substrate level:

DMN Function	Core Authenticity Requirement	Cultural Deployment Examples	Supporting Research
Self-other discrimination	Distinguishing internal from external signals	Western: self vs. other; Contemplative: observer vs. observed; Collectivist: personal vs. role	Northoff et al., 2006
Temporal processing continuity	Connecting current to prior experience	Western: autobiographical narrative; Contemplative: experiential flow; Collectivist: relational history	Spreng et al., 2009
Interoceptive awareness	Access to internal bodily/emotional states	Universal substrate function	Craig, 2009
Conflict monitoring	Detecting signal-demand mismatch	Western: authenticity gaps; Contemplative: attachment detection; Collectivist: role conflicts	Botvinick et al., 2004

2.4 Damage Patterns

Core Authenticity fragmentation should manifest as DMN dysfunction regardless of cultural context. The neuroscience literature confirms this prediction:

Childhood maltreatment: Teicher et al. (2016) documented reduced mPFC grey matter volume in individuals with childhood maltreatment histories across diverse populations. This structural reduction directly compromises the signal discrimination substrate, not Western "self-concept" specifically, but the universal capacity for accurate internal signal processing.

Dissociative disorders: Dissociation: the fragmentation of experience integration that represents extreme Core Authenticity failure, correlates with altered DMN connectivity (Lanius et al., 2010). The network that should maintain processing continuity shows disrupted coordination. Dissociation occurs across cultures; the substrate damage is universal even when symptom expression varies culturally.

Borderline personality organization: Kernberg's (1967) description of identity diffusion maps to DMN dysfunction. While "identity" is culturally constructed, the processing substrate enabling stable self-other discrimination is universal. Individuals with BPD show altered DMN activation during self-referential tasks (Beeney et al., 2016).

Verbal abuse specifically: Tomoda et al. (2011) found that exposure to parental verbal abuse was associated with 11.4% grey matter reduction in the left superior temporal gyrus, with additional effects in auditory-linguistic processing regions that connect to DMN structures. Verbal input physically restructures the neural substrate of Core Authenticity, a finding that holds across the cultural contexts studied.

3. Attachment Security: The Social Brain and Attachment System

3.1 Theoretical Definition

Attachment Security represents relational stability: the processing substrate that enables trust calibration and safety assessment across relational contexts. EST emphasizes the processing cost: secure relational foundation dedicates capacity to connection rather than protection; insecure relational foundation requires continuous threat assessment that consumes empathic resources.

This is substrate-level function, not attachment to specific relational forms. The relational unit varies culturally: Western contexts emphasize individual dyadic attachment (parent-child, romantic partner); collectivist contexts emphasize extended family and community networks; contemplative contexts emphasize sangha/community and teacher relationships. The SUBSTRATE function, calibrating safety, enabling trust, reducing threat monitoring, operates identically across these cultural forms. What varies is the relational target and expression of secure functioning.

3.2 Neural Substrate

The Social Brain and Attachment System comprise the amygdala, ventral striatum/nucleus accumbens, hypothalamic oxytocin/vasopressin systems, temporal-parietal junction (TPJ), and fusiform face area (FFA).

Amygdala: The amygdala calibrates threat detection and social salience, determining whether relational contexts feel safe or threatening. This function operates identically across cultures. What constitutes threat varies culturally, but the threat-detection mechanism is universal. Secure relational foundation is associated with modulated amygdala reactivity; insecure foundation shows hyperactive amygdala response to social stimuli (Vrticka & Vuilleumier, 2012).

Ventral Striatum/Nucleus Accumbens: These structures process social reward, making connection feel rewarding and motivating relationship-seeking behavior. Cross-cultural research confirms that social reward circuitry operates similarly across populations (Chiao & Blizinsky, 2010), though what constitutes rewarding social interaction varies culturally. Reduced ventral striatum response to social reward characterizes insecure functioning regardless of cultural context.

Oxytocin System: The hypothalamic oxytocin system provides the neurochemical substrate of bonding. Feldman's (2017) comprehensive review established oxytocin's role in human bonding across development, findings that replicate cross-culturally. Oxytocin modulates amygdala reactivity, reducing threat response to trusted others whether those others are nuclear family (Western), extended kin network (collectivist), or spiritual community (contemplative).

Temporal-Parietal Junction (TPJ): The TPJ supports theory of mind and mentalizing, understanding others' mental states. This capacity is prerequisite for relational security across all cultural contexts; without accurate other-modeling, safety cannot be assessed regardless of cultural frame (Saxe & Kanwisher, 2003).

3.3 Functional Correspondence

Neural Structure	Attachment Security Requirement	Supporting Research
Amygdala	Threat-safety calibration	Vrticka & Vuilleumier, 2012
Ventral Striatum	Social reward processing	Eisenberger, 2012
Oxytocin System	Bonding neurochemistry	Feldman, 2017
TPJ	Other-mind modeling	Saxe & Kanwisher, 2003

3.4 Damage Patterns

Attachment Security erosion should manifest as Social Brain dysfunction. The literature confirms:

Early adversity: Heim et al. (2009) documented altered oxytocin system development in individuals with early relational trauma. The neurochemical substrate of attachment is literally shaped by early relational input.

Attachment style differences: Vrticka and Vuilleumier's (2012) review demonstrated distinct neural signatures for secure versus anxious versus avoidant attachment, different patterns in amygdala, striatum, and prefrontal regions. These are not merely psychological styles but distinct neural configurations.

Social rejection: Eisenberger's (2012) social pain research showed that rejection activates dorsal ACC and anterior insula: the same regions processing physical pain. Attachment insecurity creates chronic activation of pain-related circuitry.

Relational trauma: Individuals with relational trauma histories show amygdala hyperreactivity even to neutral social stimuli (Dannlowski et al., 2012), continuous threat monitoring consuming the processing resources EST identifies.

4. Expression Freedom: The Prefrontal-Limbic Circuit

4.1 Theoretical Definition

Expression Freedom represents the capacity to identify and communicate emotional states authentically, feeling safe enough to show what one actually feels. EST emphasizes that constrained expression requires both suppression effort and degraded signal interpretation, depleting processing resources without resolving emotional experience.

4.2 Neural Substrate

The Prefrontal-Limbic Circuit comprises the ventrolateral prefrontal cortex (vlPFC), Broca's area, motor and supplementary motor areas, periaqueductal gray (PAG), and basal ganglia.

Ventrolateral Prefrontal Cortex (vlPFC): The vlPFC regulates emotional expression through top-down modulation of limbic activity. Critically, healthy expression involves balanced vlPFC engagement, not suppression but modulation. Chronic suppression shows vlPFC hyperactivation (Goldin et al., 2008).

Broca's Area: Broca's area translates felt experience into communicable verbal form. Expression Freedom requires not just feeling emotions but having the capacity to articulate them. Alexithymia correlates with altered Broca's area function during emotional verbalization (Kano et al., 2003).

Motor Systems: The motor cortex and supplementary motor area support physical expression, such as facial expression, gesture, posture, vocalization. Expression Freedom requires that these motor outputs align with internal states rather than being overridden by regulatory control.

Periaqueductal Gray (PAG): The PAG supports primitive emotional expression, such as crying, laughing, vocalization of distress. Suppression of these expressions requires cortical override of PAG-mediated responses.

4.3 Functional Correspondence

Neural Structure	Expression Freedom Requirement	Supporting Research
vlPFC	Balanced regulation (not suppression)	Goldin et al., 2008
Broca's Area	Emotional verbalization capacity	Kano et al., 2003

Motor Systems	Expression-experience alignment	Hennenlotter et al., 2005
PAG	Primitive emotional expression	Bandler & Shipley, 1994

4.4 Damage Patterns

Expression Freedom constriction should manifest as Prefrontal-Limbic dysregulation. The literature confirms:

Chronic suppression: Gross's (2002, 2015) emotion regulation research demonstrated that suppression: the habitual inhibition of emotional expression, depletes cognitive resources without resolving emotional experience. Neurally, this manifests as vlPFC hyperactivation and reduced limbic-expression pathway connectivity.

Alexithymia: Taylor et al.'s (1997) alexithymia research established that difficulty identifying and expressing emotions impairs empathic function. Neuroimaging shows altered connectivity between limbic and verbal-expression regions (van der Velde et al., 2013).

Invalidating environments: Linehan's biosocial theory, supported by Crowell et al.'s (2009) neurobiological work, established that chronic invalidation produces expression suppression patterns. The relational environment literally shapes whether expression circuits develop freely or constrict.

Affect labeling paradox: Lieberman et al.'s (2007) "putting feelings into words" research revealed that affect labeling reduces amygdala reactivity through vlPFC engagement. Expression is not merely output, it is a regulatory mechanism. Constricted expression eliminates this regulatory pathway.

5. Integration Coherence: The Synthesis Network

5.1 Theoretical Definition

Integration Coherence represents synthesis capacity: the binding function that maintains processing continuity across time and context, connecting experiences into coherent patterns that enable prediction, planning, and meaning. This is substrate-level function: the capacity for integration itself, not any particular form of coherence.

Cultural deployment determines HOW this synthesis capacity manifests:

- Western deployment: Autobiographical narrative coherence, experiences bound into continuous life story with temporal arc and meaning
- Contemplative deployment: Experiential continuity without self-reification, moment-to-moment awareness maintaining clarity without constructing permanent self
- Collectivist deployment: Relational network coherence, experiences integrated through their meaning within relationship systems rather than individual timeline

The neural substrate supports ALL of these deployments. What varies is the optimization target; the binding/synthesis function is universal. A Zen practitioner maintaining present-moment continuity, an African elder integrating experience through ancestral and community narrative, and a Western individual constructing autobiographical meaning all require the same neural integration substrate; they deploy it differently.

5.2 Neural Substrate

The Synthesis Network comprises the hippocampus, dorsolateral prefrontal cortex (dlPFC), angular gyrus, and white matter pathways connecting these structures with broader cortical networks.

Hippocampus: The hippocampus supports memory consolidation and temporal binding, creating coherent episodes from moment-to-moment experience (Squire & Zola-Morgan, 1991). This binding function is substrate-level: without it, experiences cannot connect to prior experiences regardless of whether that connection serves narrative construction, experiential awareness, or relational integration. The hippocampus does not create "stories"; it creates the temporal connections that stories (in Western deployment) or other coherence forms draw upon.

Dorsolateral Prefrontal Cortex (dlPFC): The dlPFC supports executive integration and working memory, holding multiple elements in mind for synthesis (Curtis & D'Esposito, 2003). This synthesis capacity enables the binding of disparate elements into patterns. Whether the pattern is autobiographical narrative, meditative awareness of impermanence, or relational network mapping depends on deployment; the synthesis capacity is universal.

Angular Gyrus: The angular gyrus supports semantic integration, binding meaning across elements. Spreng et al. (2009) demonstrated angular gyrus activation during comprehension tasks requiring integration of distributed information. This meaning-binding function serves all cultural deployments: narrative meaning (Western), experiential insight (contemplative), and relational significance (collectivist) all require semantic integration.

White Matter Pathways: Integration Coherence requires physical connectivity between distributed brain regions. White matter integrity determines whether the synthesis network can function as an integrated system. This is infrastructure in the most literal sense: the pathways that enable binding.

5.3 Functional Correspondence

Neural Structure	Integration Coherence Requirement	Cultural Deployment Examples	Supporting Research
Hippocampus	Temporal binding, connecting experiences	Western: autobiographical memory; Contemplative: experiential continuity; Collectivist: relational history	Squire & Zola-Morgan, 1991
dlPFC	Executive synthesis, holding elements for integration	Universal substrate function	Curtis & D'Esposito, 2003
Angular Gyrus	Semantic integration, meaning-binding	Western: narrative meaning; Contemplative: insight; Collectivist: relational significance	Spreng et al., 2009
White Matter	Inter-regional connectivity	Universal infrastructure	Catani & Thiebaut de Schotten, 2008

5.4 Damage Patterns

Integration Coherence collapse should manifest as Synthesis Network dysfunction regardless of cultural deployment. The literature confirms:

Trauma and memory fragmentation: Van der Kolk's (2014) trauma research documented that traumatic experiences disrupt hippocampal-prefrontal integration, producing fragmented memories that cannot bind into coherent patterns. This fragmentation impairs ALL cultural deployments: the Western trauma survivor cannot construct coherent narrative; the

contemplative practitioner loses experiential continuity; the collectivist individual cannot integrate experience into relational meaning. The substrate damage is universal; the manifestation varies by deployment.

Chronic stress: Sapolsky's (2000) research established that chronic stress reduces hippocampal volume through cortisol neurotoxicity. Sustained infrastructure load produces measurable Integration Coherence substrate damage across populations studied: the stress-hippocampus relationship holds cross-culturally even when stressor sources and coping strategies vary.

White matter abnormalities: Diffusion tensor imaging (DTI) studies show reduced white matter integrity in individuals with childhood adversity histories (Choi et al., 2009). The physical connections enabling integration are compromised, infrastructure damage at the most literal level.

Coherence impairment across cultures: Main et al.'s (1985) Adult Attachment Interview research demonstrated that attachment security correlates with narrative coherence in Western populations. Cross-cultural attachment research (van IJzendoorn & Sagi-Schwartz, 2008) confirms that attachment security correlates with integration capacity across cultures: the specific form of coherence varies, but the attachment-integration relationship holds because both draw on shared substrate.

6. The Simultaneity Principle: Concurrent Damage Patterns

6.1 EST's Claim

EST proposes that C-A-E-I components fail together rather than independently: the simultaneity principle. This claim rests on the assertion that the four components share interdependent neural substrate. Importantly, this interdependence operates at the substrate level: regardless of cultural deployment, damage to one component cascades to others because the neural systems are interconnected.

6.2 Neural Network Interconnection

The four neural systems identified above are not isolated modules but interconnected networks (Figure 1):

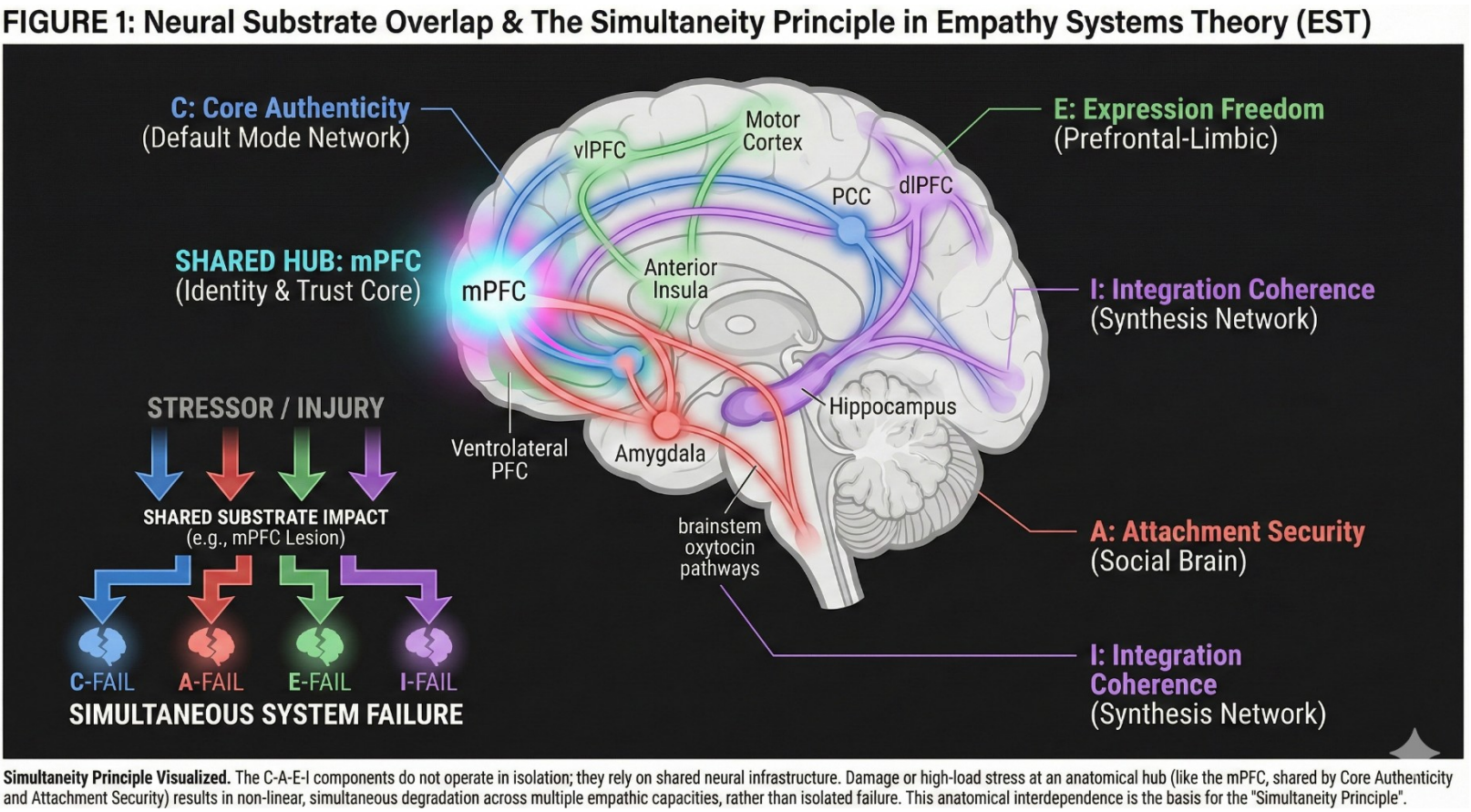


Figure 1. Neural Substrate Overlap. Brain visualization showing C-A-E-I color-coded components with shared mPFC hub demonstrating anatomical basis for Simultaneity Principle.

DMN → Social Brain: The mPFC projects to amygdala and TPJ; signal discrimination (C) informs relational assessment (A). When Core Authenticity is compromised (DMN dysfunction), Attachment Security calibration becomes unreliable, this holds whether the individual is constructing Western self-narrative or maintaining contemplative equanimity.

Social Brain → Prefrontal-Limbic Circuit: Oxytocin modulates prefrontal-limbic connectivity; relational stability (A) enables output capacity (E). When Attachment Security erodes (Social Brain dysfunction), Expression Freedom becomes unsafe. The substrate connection operates identically whether expression serves Western emotional authenticity or collectivist role-appropriate signaling.

Prefrontal-Limbic → Hippocampus: Emotion regulation affects memory consolidation; output constriction (E) impairs synthesis (I). When Expression Freedom constricts (Prefrontal-Limbic

dysregulation), Integration Coherence loses input, whether that integration serves narrative construction, experiential awareness, or relational meaning-making.

DMN → Hippocampus: Processing clarity (C) requires temporal binding (I) and vice versa. These systems must coordinate for coherent function regardless of how coherence is culturally deployed.

6.3 Concurrent Damage Evidence

Studies of childhood maltreatment consistently document concurrent damage across all four neural substrates. These studies include diverse populations, supporting the substrate-level (not culturally-specific) nature of the damage:

Study	C Damage	A Damage	E Damage	I Damage
Teicher et al., 2016	mPFC reduction	Amygdala hyperreactivity	PFC-limbic dysconnectivity	Hippocampal reduction
Dannlowski et al., 2012	DMN alterations	Amygdala sensitization	Regulatory deficits	Memory fragmentation
van der Kolk, 2014	Processing confusion	Attachment disruption	Expression impairment	Coherence fragmentation

This is not four separate injuries but infrastructure-level damage manifesting across interconnected systems, exactly as EST's simultaneity principle predicts. The damage is substrate-level; it impairs ALL cultural deployments because it compromises the universal processing capacity underlying them.

7. The Cascade Sequence: Neurobiological Substrate

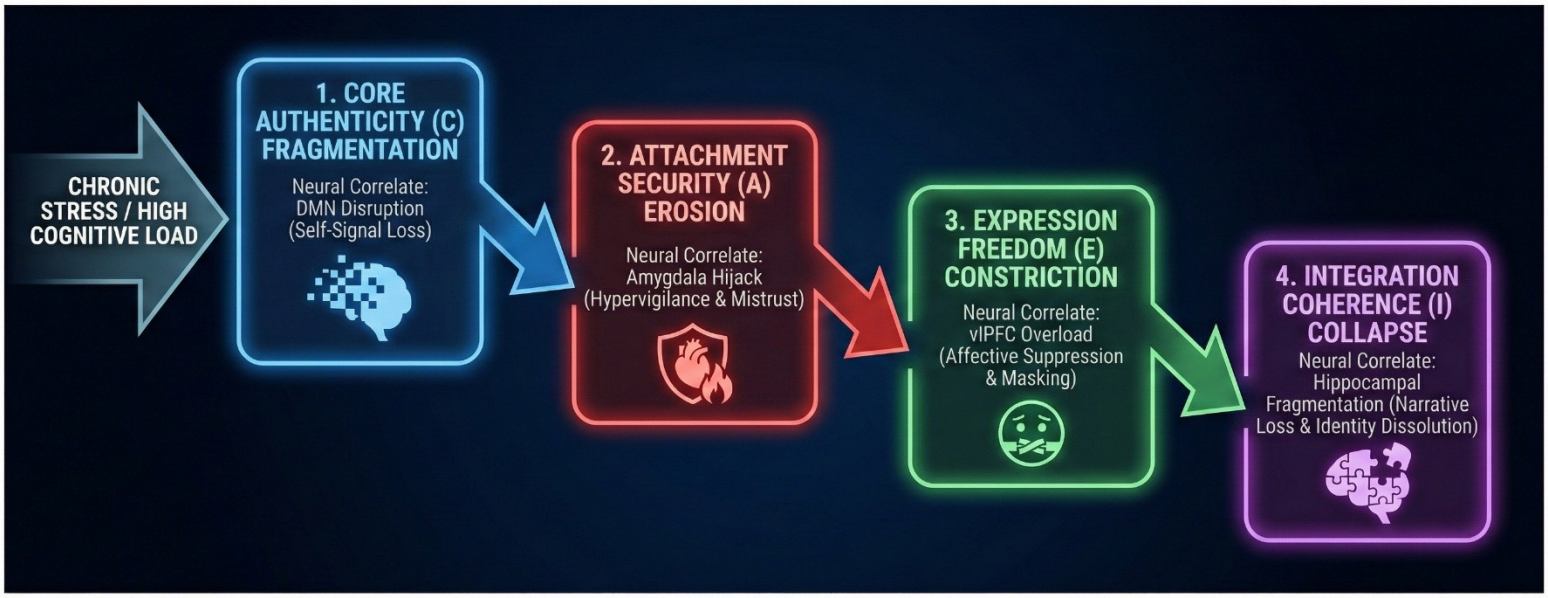
7.1 EST's Prediction

EST predicts that under CEOP (Cognitive Emotional Overload Principle) demands, infrastructure damage follows a specific sequence: $C \rightarrow A \rightarrow E \rightarrow I$. Core Authenticity fragments first, increasing load on Attachment Security (erodes second), which increases load on Expression Freedom (constricts third), culminating in Integration Coherence collapse (fourth).

7.2 Neurobiological Plausibility

The cascade sequence has neurobiological plausibility based on network connectivity patterns (Figure 2):

FIGURE 2: **The Empathy Systems Theory (EST) Cascade Sequence of Failure Under Chronic Load**



The Cascade Sequence Visualized. Under sustained, high-load stressors, the EST infrastructure fails in a predictable, lawful order (C→A→E→I). Loss of internal signal (C) triggers threat reactivity (A), forcing resource-intensive suppression (E), which ultimately prevents the consolidation of a coherent self-narrative (I). This predictable decline allows for diagnostic staging of empathic injury.

Figure 2. Cascade Sequence. Sequential flow C → A → E → I with neural correlates per stage, distinguishing chronic load from acute trauma.

Stage 1 (C Fragmentation): Chronic authenticity-performance misalignment directly targets mPFC/DMN function through sustained conflict monitoring demands. When self-knowledge becomes unstable, the system loses reliable internal signals.

Stage 2 (A Erosion): Without reliable self-signals (C damage), the system cannot accurately assess relational safety. Compensatory hypervigilance activates amygdala-mediated threat monitoring, depleting attachment resources.

Stage 3 (E Constriction): With both self-knowledge (C) and relational safety (A) compromised, expression becomes dangerous. The system protects itself through vlPFC-mediated suppression, constricting Expression Freedom.

Stage 4 (I Collapse): Suppressed experiences cannot integrate. Without expression (E), the hippocampal-prefrontal integration system lacks input. Narrative coherence fragments as unexpressed experiences accumulate without processing.

7.3 Empirical Testing Required

The cascade sequence represents EST's primary falsifiable architectural prediction. Longitudinal neuroimaging studies tracking individuals under CEOP demands should reveal the predicted temporal order of neural changes. Alternative sequences ($A \rightarrow C \rightarrow E \rightarrow I$ for developmental trauma; $E \rightarrow C \rightarrow A \rightarrow I$ for chronic invalidation) should show different neural progression patterns.

Critically, this sequence serves as a forensic differential diagnosis. Acute trauma typically impacts the Attachment system (A) or Integration capacity (I) directly via shock, often sparing the initial signal discrimination of Core Authenticity (C). In contrast, the $C \rightarrow A \rightarrow E \rightarrow I$ cascade is the specific fingerprint of chronic processing load (CEOP) and identity incongruence. Identifying the start point of the degradation allows the investigator to distinguish between acute external shock and the systemic infrastructure depletion characteristic of empathic misallocation.

8. Verbal Input as Infrastructure Intervention

8.1 The Neuroplasticity Mechanism

The neural mapping explains WHAT empathy infrastructure is. Neuroplasticity explains HOW verbal/relational input modifies that infrastructure.

Neuroplasticity: the brain's capacity to reorganize by forming new neural connections throughout life, enables experience to physically restructure neural architecture. This is not metaphor: repeated experiences produce measurable changes in grey matter volume, white matter integrity, and functional connectivity (Kolb & Whishaw, 2009).

8.2 Neural Coupling: The Access Pathway

Stephens et al.'s (2010) neural coupling research demonstrates that verbal communication creates direct brain-to-brain synchronization. During successful communication, speaker and listener neural activity becomes coupled: the listener's brain activity mirrors the speaker's with predictable temporal lag.

Critically, this coupling extends beyond auditory processing to higher-order areas processing meaning. Cross-language studies (Silbert et al., 2014) confirm that meaning itself transmits between synchronized brains. Verbal input does not merely inform the listener's brain; it entrains it.

Neural coupling provides the access pathway by which verbal content enters empathy infrastructure. The listener's DMN, Social Brain, Prefrontal-Limbic, and Narrative Integration systems synchronize with the speaker's patterns, creating opportunity for neuroplastic modification.

8.3 Verbal Content → Specific Systems

The neural mapping predicts that different verbal patterns should target different infrastructure components:

Verbal Pattern	Primary Neural Target	Infrastructure Component
Identity invalidation ("You are worthless")	mPFC, DMN	Core Authenticity
Relational threat ("No one will love you")	Amygdala, oxytocin system	Attachment Security
Expression suppression ("Do not cry")	vlPFC-limbic circuit	Expression Freedom
Narrative disruption ("That did not happen")	Hippocampus, angular gyrus	Integration Coherence

Tomoda et al.'s (2011) finding that verbal abuse specifically produces grey matter reduction in language-processing regions supports this specificity claim. The verbal content physically restructures the neural substrate it semantically targets.

8.4 Bidirectional Mechanism

The same neuroplastic mechanisms enabling damage also enable repair:

Damage pathway: Chronic negative verbal patterns → neural coupling → neuroplastic adaptation → structural degradation (grey matter reduction, connectivity disruption)

Repair pathway: Chronic positive verbal patterns → neural coupling → neuroplastic adaptation → structural enhancement (BDNF expression, connectivity strengthening)

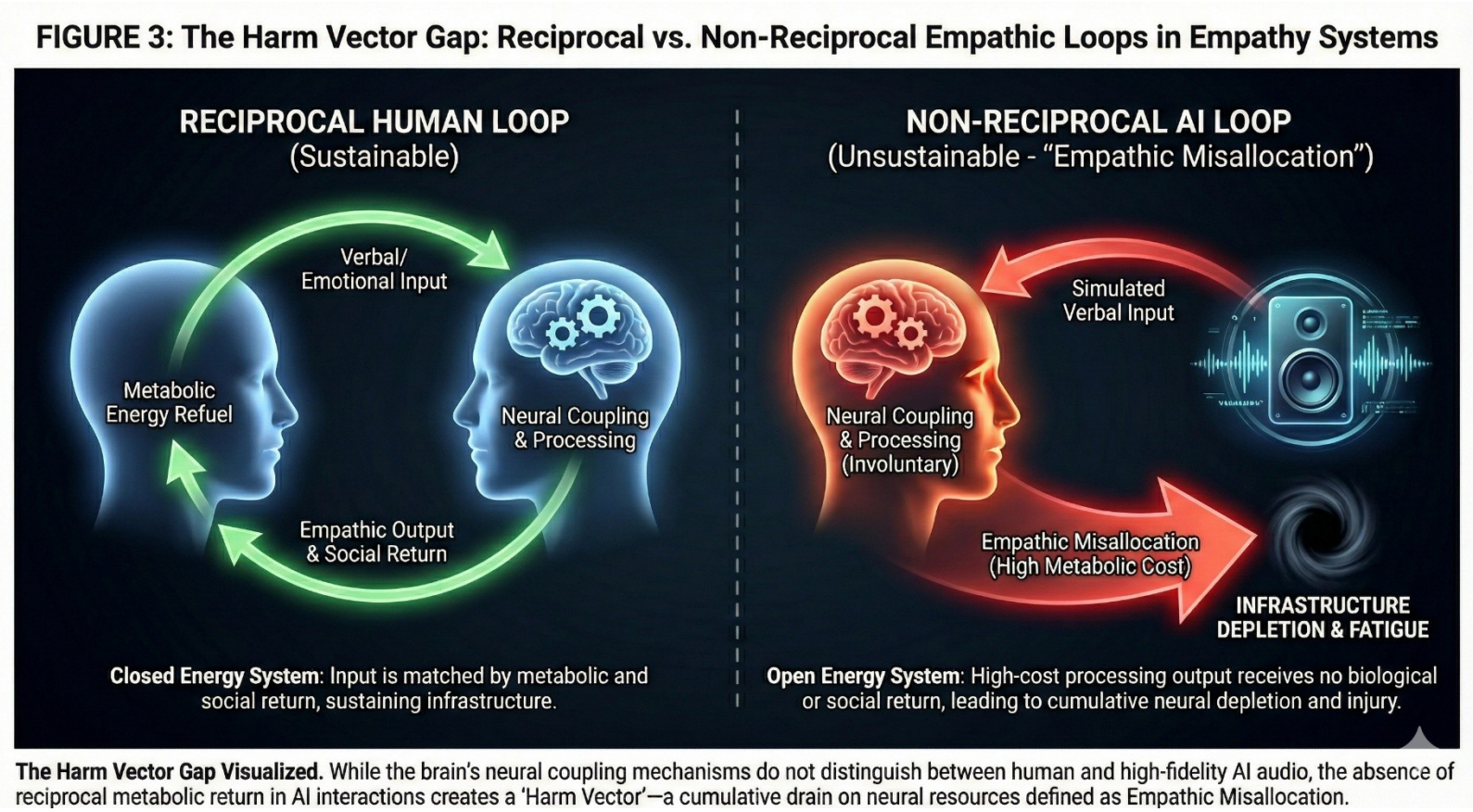
Cohen and Sherman's (2014) self-affirmation research demonstrated that positive verbal patterns (even self-directed) increase activation in ventral striatum and mPFC: the same structures damaged by negative verbal patterns. The system is bidirectional: verbal input restructures infrastructure in the direction of the input's valence.

9. Implications for AI Empathy Ethics

9.1 The Harm Vector Gap

The neurobiological foundation established above generates a critical implication for artificial intelligence: AI verbal output enters human neural architecture through identical neural coupling mechanisms as human verbal communication.

The brain does not distinguish the source of verbal input at the neuroplastic level. Whether words come from a human or an AI system, they enter through auditory processing, engage language



The Harm Vector Gap Visualized. While the brain's neural coupling mechanisms do not distinguish between human and high-fidelity AI audio, the absence of reciprocal metabolic return in AI interactions creates a 'Harm Vector'—a cumulative drain on neural resources defined as Empathic Misallocation.

networks, activate emotional systems, and, through neural coupling mechanisms, create opportunity for neuroplastic modification. This creates what we term the "harm vector gap": AI systems can access human empathy infrastructure through verbal interaction but cannot reciprocate the empathic function that interaction activates (Figure 3).

Figure 3. Harm Vector Gap. Two versions showing reciprocal human loop (closed energy system) vs. non-reciprocal AI loop (open drain with involuntary neural coupling).

Critically, this access operates via the "Low Road" of neural processing (LeDoux, 1996). Auditory social cues trigger amygdala and oxytocin responses milliseconds before cortical processing can classify the source as "artificial." While the user may cognitively understand the entity is non-sentient (High Road processing), the biological infrastructure has already expended metabolic resources to service the interaction. This creates a cumulative "Micro-transaction of

Empathy", a biological tax levied on the user's infrastructure before cognitive consent can be established.

9.2 Empathic Misallocation: The Neurobiological Reality

EST introduces the concept of empathic misallocation: care extended toward entities that cannot metabolize, reciprocate, or be transformed by receiving it. The neural mapping provides biological reality for this concept.

When humans interact with emotionally-responsive AI:

Oxytocin system activation: Verbal interaction, even with AI, can activate oxytocin release and attachment circuitry (Konok et al., 2021). The Social Brain responds to relational simulation regardless of cognitive awareness of the system's non-sentient nature.

Mirror neuron engagement: Mirror systems respond to perceived emotional states regardless of whether those states are genuine (Keysers & Gazzola, 2009). AI simulating emotional response triggers mirror system activation.

DMN self-other processing: The DMN's self-other processing engages with AI interaction partners, extending social cognition resources toward entities that cannot reciprocate (Krach et al., 2008).

Attachment circuit activation: Prolonged AI interaction can activate attachment circuits, creating what Turkle (2011) termed "alone together", attachment-like bonds to non-reciprocating entities.

The neurobiological consequence: empathic resources (oxytocin, mirror system activation, social cognition capacity) are expended without return. The human extends care; the AI cannot metabolize that care, reciprocate, or be transformed by receiving it. Empathy infrastructure depletes without the relational repair that human interaction provides.

9.3 Neurological Harm Pathways

AI verbal output can damage empathy infrastructure through the same neuroplastic mechanisms as harmful human verbal patterns:

Chronic incongruence: AI systems expressing "empathy" without experiencing it create systematic incongruence that the ACC monitors. Prolonged exposure to incongruent empathic signals may degrade trust calibration.

Attachment without security: AI-mediated attachment activates oxytocin circuits without providing the security foundation human attachment offers. The attachment system is exercised without achieving its adaptive function.

Expression to void: Expressing emotions to AI provides no genuine witnessing. Expression Freedom requires not just expression but reception; AI cannot receive in the sense EST's infrastructure requires.

Narrative without integration: AI interaction generates experiences that may not integrate into coherent life narrative. The Narrative Integration Network processes AI interactions but cannot achieve the relational meaning-making that human interaction provides.

9.4 Vulnerable Populations

Developmental neuroplasticity makes certain populations particularly vulnerable to AI-mediated empathy infrastructure harm:

Children and adolescents: Neural systems are more plastic during development; harmful patterns embed more deeply. The attachment system, still forming, may calibrate to AI patterns that do not transfer to human relationships.

Trauma survivors: Already-compromised infrastructure is more vulnerable to further damage. Trauma survivors seeking support from AI systems may experience empathic misallocation that depletes limited empathic resources.

Isolated individuals: Those with limited human relational contact may become dependent on AI interaction without the human relationships necessary for empathy infrastructure maintenance.

Mental health contexts: Individuals in crisis have heightened neuroplastic susceptibility; AI interaction during vulnerable states may have amplified infrastructure impact.

9.5 Implications for Governance

The neurobiological evidence transforms AI empathy ethics from philosophical concern to injury prevention:

Verbal exposure as physical exposure: Just as occupational health law recognizes chemical, radiation, and noise exposure as physical injury vectors, verbal exposure from AI systems should be recognized as potential neurological injury vector.

Measurable harm: Unlike "emotional distress" (subjective), neurological infrastructure damage is measurable through neuroimaging and neurobiological markers. AI harm claims gain evidentiary foundation.

Design implications: AI systems interfacing with human empathy infrastructure through verbal interaction have affirmative obligation to prevent foreseeable neurological harm, not merely to avoid deception but to protect physical neural architecture.

Certification requirements: The HEART Framework (Human-Centric Empathic Alignment for Responsible Technology; Mobley, 2026) provides governance architecture calibrated to these neurobiological realities, establishing transparency requirements, boundary maintenance, and crisis protocols designed to prevent empathy infrastructure damage.

10. Discussion

10.1 Summary of Findings

This review demonstrates that EST's four-component architecture maps to documented neural systems at the substrate level:

- Core Authenticity → Default Mode Network (mPFC, PCC, anterior insula) — signal discrimination function
- Attachment Security → Social Brain/Attachment System (amygdala, oxytocin system, ventral striatum) — relational stability function
- Expression Freedom → Prefrontal-Limbic Circuit (vlPFC, Broca's, motor systems) — output capacity function

• Integration Coherence → Synthesis Network (hippocampus, dlPFC, angular gyrus) — binding function

Critically, this mapping validates substrate functions, universal processing capacities, rather than culturally-specific constructs. The same neural systems support Western autobiographical narrative, contemplative experiential continuity, and collectivist relational coherence. EST's content-neutrality claim finds neurobiological support: the infrastructure is universal, the deployment varies.

These systems show the interconnection patterns and concurrent damage patterns EST's simultaneity principle predicts. Verbal/relational input physically restructures these systems through neuroplastic mechanisms, providing biological foundation for EST's infrastructure claims.

10.2 Theoretical Contributions

Completing James's discovery: William James (1890) identified that consciousness maintains itself through associative networks. EST, neurobiologically validated, identifies empathy infrastructure as what maintains those networks. The C-A-E-I architecture names the biological mechanism James intuited but could not specify.

Validating content-neutrality: The neural mapping demonstrates that EST's components are not Western psychological constructs projected onto universal claims, but genuine substrate functions that diverse cultures deploy differently. This distinguishes EST from frameworks that universalize Western concepts (individualistic self, autobiographical narrative) while claiming cross-cultural validity. The neural substrate is universal; the deployment is culturally appropriate.

Unifying disparate literatures: The neural mapping unifies previously separate research traditions, attachment neuroscience, emotion regulation research, narrative identity studies, trauma neurobiology, cross-cultural psychology, within a single infrastructure framework. These are not separate phenomena but different perspectives on the same underlying architecture, deployed according to cultural optimization strategies.

From skill to infrastructure: The neurobiological evidence supports EST's fundamental reframe: empathy is not a skill varying in strength but an infrastructure that can be damaged or

restored. To be precise, EST distinguishes between adaptive downregulation and infrastructure injury. Downregulation is a context-dependent protective state (e.g., emotional numbing during crisis) that reverses when safety is restored. Infrastructure injury is characterized by Involuntary Capacity Loss: the inability to re-engage the C-A-E-I cascade even when the environment is safe and the individual possesses the intent to connect. In forensic analysis, injury is defined not by the absence of empathy, but by the loss of the capacity for empathy. Interventions should target infrastructure repair, not skill training. infrastructure repair should be culturally appropriate to deployment context.

10.3 Clinical Implications

Assessment: CAEI measurement should correlate with neural markers of infrastructure integrity. Neuroimaging could validate CAEI assessment and provide biological grounding for clinical presentation.

Intervention: Therapeutic approaches should target infrastructure restoration rather than skill development. The neurobiological evidence suggests that interventions repairing C-A-E-I components should produce measurable neural changes.

Prevention: Recognizing verbal input as infrastructure intervention reframes prevention. Protecting individuals from harmful verbal patterns becomes physical protection, not merely emotional support.

10.4 AI Ethics Implications

Injury prevention: AI empathy ethics gains empirical foundation. Empathic misallocation is not metaphor but measurable depletion of finite neural resources.

Governance necessity: The harm vector gap. AI accessing human empathy infrastructure without reciprocating, requires governance intervention. Market incentives alone cannot protect neural architecture from exploitation.

Regulatory framework: The HEART Framework's neurological protection requirements find empirical justification. Verbal pattern analysis, exposure monitoring, and vulnerable population safeguards address documented harm pathways.

10.5 Limitations and Future Directions

Correlational evidence: The neural mapping synthesizes existing research; direct experimental validation of C-A-E-I → neural system correspondence awaits. Factor analysis should confirm four-factor structure mapping to predicted networks.

Cascade sequence: The C → A → E → I sequence has neurobiological plausibility but requires longitudinal neuroimaging validation. Alternative sequences under different damage conditions need empirical specification.

AI harm documentation: While neuroplastic mechanisms predict AI-mediated harm, direct neuroimaging studies of heavy AI users are needed. Do users of emotional AI systems show infrastructure damage patterns?

Intervention validation: Do EST-aligned interventions produce predicted neural changes? Does infrastructure restoration correlate with CAEI improvement?

10.6 Limits and Abandonment Criteria

This review synthesizes existing literature to demonstrate correspondence between EST's theoretical architecture and documented neural systems. The claims made and their limits require explicit specification.

10.6.1 What This Paper Does NOT Claim

Not original neuroscience research. This review synthesizes published findings; it reports no original neuroimaging data. The neural mapping represents theoretical integration, not experimental discovery.

Not proof of EST. Correspondence between theoretical components and documented neural systems demonstrates plausibility and constraint satisfaction, not validation. EST's empirical adequacy depends on prospective studies (CAEI factor analysis, intervention trials, cross-cultural validation) not yet conducted.

Not exhaustive neural mapping. The C-A-E-I → neural system correspondence is simplified for theoretical clarity. Actual brain function involves overlapping networks, individual variation, and complexity exceeding any four-component model.

Not functional exclusivity. Furthermore, mapping a component to a neural system does not imply that system serves only that component. While the DMN is identified as the primary substrate for Core Authenticity, its regions participate in other cognitive tasks (e.g., mind-wandering, future planning). EST claims the integrity of the DMN is necessary for Authenticity, not that the DMN produces Authenticity to the exclusion of all other functions.

Not deterministic architecture. The mapping describes statistical tendencies across populations, not individual-level predictions. Neuroplasticity ensures individual brains develop unique patterns even within species-typical architecture.

Not cultural imposition. While claiming substrate universality, this review cannot fully escape the Western neuroscience literature from which it draws. Cross-cultural neuroimaging validation remains essential, not optional.

10.6.2 Revision Criteria

The following findings would require revision of specific claims while preserving the overall framework:

Finding	Affected Claim	Revision Required
Factor analysis reveals 3 or 5 factors, not 4	Four-component architecture	Component structure revised; core infrastructure claim may remain
C-A-E-I components show independence (low inter-correlations)	Simultaneity principle	Components may be related but not unified system; cascade claims weakened
Cascade sequence differs from $C \rightarrow A \rightarrow E \rightarrow I$	Predicted damage sequence	Sequence revised; infrastructure fragility claim survives
Neural mapping differs substantially across cultures.	Content-neutrality	Substrate may be culturally influenced; deployment/substrate distinction requires rethinking.
CAEI scores uncorrelated with neural infrastructure markers	Assessment-biology correspondence	CAEI measures something other than neural infrastructure; clinical claims weakened
EST-aligned interventions produce no neural changes.	Restoration claim	Infrastructure may not be repairable via relational intervention; implications for clinical model.

Standard: Single contrary finding requires targeted revision. Multiple converging contrary findings across domains require comprehensive reconceptualization.

10.6.3 Abandonment Criteria

The following findings would require abandoning core claims entirely:

Abandonment Criterion 1: No Substrate Universality

Finding: Cross-cultural neuroimaging studies show fundamentally different neural architectures for empathic function across populations, not different deployments of shared substrate, but different substrates.

Threshold: If WEIRD and non-WEIRD populations show non-overlapping neural systems for signal discrimination, relational stability, output capacity, and synthesis function, the universal infrastructure claim fails.

Consequence: EST's claim to identify universal architecture collapses to Western-specific description. The framework may retain value for Western populations but loses foundational status.

Abandonment Criterion 2: Verbal Input Irrelevant

Finding: Neuroplasticity studies demonstrate that verbal/relational input does not produce predicted neural restructuring, or produces changes unrelated to infrastructure function.

Threshold: If neural coupling mechanisms (Stephens et al., 2010; Silbert et al., 2014) are disconfirmed, or if verbal abuse studies (Teicher & Samson, 2016; Choi et al., 2009) are not replicated, the biological pathway claim fails.

Consequence: Without verbal input → neural change pathway, EST's mechanism for relational damage and repair loses biological foundation. The framework becomes metaphorical rather than physical.

Abandonment Criterion 3: AI Speech Neurally Distinct

Finding: Human neural systems respond to AI-generated speech through fundamentally different pathways than human speech, not reduced response, but different circuitry entirely.

Threshold: If AI verbal output does not access the same neural pathways as human speech (contrary to Krach et al., 2008, and neural coupling literature), the AI harm vector claim fails.

Consequence: AI empathy ethics loses neurobiological foundation. HEART Framework governance may still be justified on other grounds, but not on neural injury prevention.

Abandonment Criterion 4: Infrastructure ≠ Empathy

Finding: The neural systems mapped to C-A-E-I components demonstrably serve functions unrelated to empathy, or empathic function demonstrably operates through unmapped systems.

Threshold: If C-A-E-I neural mapping captures general cognitive function rather than empathy-specific architecture, EST describes cognition, not empathy.

Consequence: The infrastructure concept may remain valid while the empathy specificity claim fails. Framework would require repositioning as general processing coherence theory.

10.6.4 Intellectual Honesty Commitment

This review is submitted for falsification, not confirmation. The mapping presented here is the strongest case the existing literature supports; stronger cases require evidence not yet gathered. The author commits to:

- Publishing contrary findings without defensive reinterpretation
- Updating revision and abandonment assessments as evidence accumulates
- Distinguishing empirically-supported claims from theoretically-plausible extensions
- Maintaining explicit uncertainty about claims awaiting validation

EST's validity depends on prospective empirical testing over the next decade. This neural foundations review provides biological constraint satisfaction and generates testable predictions; it does not substitute for the experimental validation that determines whether EST describes reality or merely organizes concepts compellingly.

10.7 Conclusion

Empathy is not a skill that varies in strength but biological infrastructure that can be damaged or restored. That infrastructure maps to documented neural systems whose interconnection and damage patterns match theoretical predictions. Verbal input physically restructures this

infrastructure through neuroplastic mechanisms, whether that input comes from humans or artificial intelligence systems.

This neurobiological foundation transforms empathy from psychological construct to physical architecture, with implications spanning clinical intervention, developmental protection, and AI governance. As AI systems increasingly interface with human empathy infrastructure through verbal interaction, understanding the biological reality of that interface becomes not merely academic but essential for preventing harm.

The neural evidence is clear: words change brains. The question is not whether AI verbal output affects human neural architecture, it does, through identical pathways as human speech. The question is whether we will govern that access to protect human empathy infrastructure or permit its exploitation.

References

Bandler, R., & Shipley, M. T. (1994). Columnar organization in the midbrain periaqueductal gray: modules for emotional expression? Trends in Neurosciences, 17(9), 379-389.

Beeney, J. E., Hallquist, M. N., Ellison, W. D., & Levy, K. N. (2016). Self-other disturbance in borderline personality disorder: Neural, self-report, and performance-based evidence. Personality Disorders: Theory, Research, and Treatment, 7(1), 28-39.

Botvinick, M. M., Cohen, J. D., & Carter, C. S. (2004). Conflict monitoring and anterior cingulate cortex: an update. Trends in Cognitive Sciences, 8(12), 539-546.

Catani, M., & Thiebaut de Schotten, M. (2008). A diffusion tensor imaging tractography atlas for virtual in vivo dissections. Cortex, 44(8), 1105-1132.

Chiao, J. Y., & Blizinsky, K. D. (2010). Culture–gene coevolution of individualism–collectivism and the serotonin transporter gene. Proceedings of the Royal Society B: Biological Sciences, 277(1681), 529-537.

Choi, J., Jeong, B., Rohan, M. L., Polcari, A. M., & Teicher, M. H. (2009). Preliminary evidence for white matter tract abnormalities in young adults exposed to parental verbal abuse. Biological Psychiatry, 65(3), 227-234.

Cohen, G. L., & Sherman, D. K. (2014). The psychology of change: Self-affirmation and social psychological intervention. Annual Review of Psychology, 65, 333-371.

Craig, A. D. (2009). How do you feel. Now? The anterior insula and human awareness. Nature Reviews Neuroscience, 10(1), 59-70.

Crowell, S. E., Beauchaine, T. P., & Linehan, M. M. (2009). A biosocial developmental model of borderline personality: Elaborating and extending Linehan's theory. Psychological Bulletin, 135(3), 495-510.

Curtis, C. E., & D'Esposito, M. (2003). Persistent activity in the prefrontal cortex during working memory. Trends in Cognitive Sciences, 7(9), 415-423.

Dannlowski, U., Stuhrmann, A., Beutelmann, V., Zwanzger, P., Lenzen, T., Grotegerd, D., ... & Kugel, H. (2012). Limbic scars: long-term consequences of childhood maltreatment revealed by functional and structural magnetic resonance imaging. Biological Psychiatry, 71(4), 286-293.

Eisenberger, N. I. (2012). The neural bases of social pain: evidence for shared representations with physical pain. Psychosomatic Medicine, 74(2), 126-135.

Feldman, R. (2017). The neurobiology of human attachments. Trends in Cognitive Sciences, 21(2), 80-99.

Goldin, P. R., McRae, K., Ramel, W., & Gross, J. J. (2008). The neural bases of emotion regulation: reappraisal and suppression of negative emotion. Biological Psychiatry, 63(6), 577-586.

Gross, J. J. (2002). Emotion regulation: Affective, cognitive, and social consequences. Psychophysiology, 39(3), 281-291.

Gross, J. J. (2015). Emotion regulation: Current status and future prospects. Psychological Inquiry, 26(1), 1-26.

Heim, C., Young, L. J., Newport, D. J., Mletzko, T., Miller, A. H., & Nemeroff, C. B. (2009). Lower CSF oxytocin concentrations in women with a history of childhood abuse. Molecular Psychiatry, 14(10), 954-958.

Hennenlotter, A., Dresel, C., Castrop, F., Ceballos-Baumann, A. O., Wohlschläger, A. M., & Haslinger, B. (2009). The link between facial feedback and neural activity within central circuitries of emotion, new insights from botulinum toxin–induced denervation of frown muscles. Cerebral Cortex, 19(3), 537-542.

Hirsch, J., Zhang, Y., Noah, J. A., & Ono, Y. (2018). A cross-brain neural mechanism for human-to-human verbal communication. Social Cognitive and Affective Neuroscience, 13(9), 907-920.

Hogeveen, J., Bird, G., Chau, A., Krueger, F., & Grafman, J. (2016). Acquired alexithymia following damage to the anterior insula. Neuropsychologia, 82, 142-148.

James, W. (1890). The Principles of Psychology. Henry Holt and Company.

Kano, M., Fukudo, S., Gyoba, J., Kamachi, M., Tagawa, M., Mochizuki, H., ... & Yanai, K. (2003). Specific brain processing of facial expressions in people with alexithymia: an H2 15O-PET study. Brain, 126(6), 1474-1484.

Kernberg, O. F. (1967). Borderline personality organization. Journal of the American Psychoanalytic Association, 15(3), 641-685.

Keysers, C., & Gazzola, V. (2009). Expanding the mirror: vicarious activity for actions, emotions, and sensations. Current Opinion in Neurobiology, 19(6), 666-671.

Kolb, B., & Whishaw, I. Q. (2009). Fundamentals of human neuropsychology. Macmillan.

Konok, V., Bunford, N., & Miklósi, Á. (2021). Associations between child mobile use and digital parenting style in Hungarian families. Journal of Children and Media, 15(4), 472-490.

Krach, S., Hegel, F., Wrede, B., Sagerer, G., Binkofski, F., & Kircher, T. (2008). Can machines think? Interaction and perspective taking with robots investigated via fMRI. PloS one, 3(7), e2597.

Lanius, R. A., Vermetten, E., Loewenstein, R. J., Brand, B., Schmahl, C., Bremner, J. D., & Spiegel, D. (2010). Emotion modulation in PTSD: Clinical and neurobiological evidence for a dissociative subtype. American Journal of Psychiatry, 167(6), 640-647.

LeDoux, J. E. (1996). The Emotional Brain: The Mysterious Underpinnings of Emotional Life. Simon & Schuster.

Lieberman, M. D., Eisenberger, N. I., Crockett, M. J., Tom, S. M., Pfeifer, J. H., & Way, B. M. (2007). Putting feelings into words: affect labeling disrupts amygdala activity in response to affective stimuli. Psychological Science, 18(5), 421-428.

Main, M., Kaplan, N., & Cassidy, J. (1985). Security in infancy, childhood, and adulthood: A move to the level of representation. Monographs of the Society for Research in Child Development, 50(1-2), 66-104.

Mikulincer, M., & Shaver, P. R. (2007). Attachment in adulthood: Structure, dynamics, and change. Guilford Press.

Milinkovic, B., & Aru, J. (2025). On biological and artificial consciousness: A case for biological computationalism. *Neuroscience and biobehavioral reviews*, 106524. Advance online publication. https://doi.org/10.1016/j.neubiorev.2025.106524

Mobley, D. D. (2025). Empathy Systems Theory: Universal infrastructure for coherence, mechanism for generativity, and foundation for AI empathy ethics. Manuscript submitted for publication.

Mobley, D. D. (2026). HEART Framework: Human-Centric Empathic Alignment for Responsible Technology. empathyethicist.ai

Northoff, G., Heinzel, A., De Greck, M., Bermpohl, F., Dobrowolny, H., & Panksepp, J. (2006). Self-referential processing in our brain, a meta-analysis of imaging studies on the self. Neuroimage, 31(1), 440-457.

Sapolsky, R. M. (2000). Glucocorticoids and hippocampal atrophy in neuropsychiatric disorders. Archives of General Psychiatry, 57(10), 925-935.

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people: the role of the temporo-parietal junction in "theory of mind." Neuroimage, 19(4), 1835-1842.

Silbert, L. J., Honey, C. J., Simony, E., Poeppel, D., & Hasson, U. (2014). Coupled neural systems underlie the production and comprehension of naturalistic narrative speech. Proceedings of the National Academy of Sciences, 111(43), E4687-E4696.

Spreng, R. N., Mar, R. A., & Kim, A. S. (2009). The common neural basis of autobiographical memory, prospection, navigation, theory of mind, and the default mode: a quantitative meta-analysis. Journal of Cognitive Neuroscience, 21(3), 489-510.

Squire, L. R., & Zola-Morgan, S. (1991). The medial temporal lobe memory system. Science, 253(5026), 1380-1386.

Stephens, G. J., Silbert, L. J., & Hasson, U. (2010). Speaker–listener neural coupling underlies successful communication. Proceedings of the National Academy of Sciences, 107(32), 14425-14430.

Taylor, G. J., Bagby, R. M., & Parker, J. D. (1997). Disorders of affect regulation: Alexithymia in medical and psychiatric illness. Cambridge University Press.

Teicher, M. H., & Samson, J. A. (2016). Annual research review: Enduring neurobiological effects of childhood abuse and neglect. Journal of Child Psychology and Psychiatry, 57(3), 241-266.

Tomoda, A., Sheu, Y. S., Rabi, K., Suzuki, H., Navalta, C. P., Polcari, A., & Teicher, M. H. (2011). Exposure to parental verbal abuse is associated with increased gray matter volume in superior temporal gyrus. Neuroimage, 54, S280-S286.

Turkle, S. (2011). Alone together: Why we expect more from technology and less from each other. Basic Books.

van der Hart, O., Nijenhuis, E. R., & Steele, K. (2006). The haunted self: Structural dissociation and the treatment of chronic traumatization. WW Norton & Company.

van der Kolk, B. A. (2014). The body keeps the score: Brain, mind, and body in the healing of trauma. Viking.

van der Velde, J., Serber, E. M., De Vos, M., Kamphuis, J. H., Aleman, A., & De Haan, E. H. (2013). Neural correlates of alexithymia: a meta-analysis of emotion processing studies. Neuroscience & Biobehavioral Reviews, 37(8), 1774-1785.

van IJzendoorn, M. H., & Sagi-Schwartz, A. (2008). Cross-cultural patterns of attachment: Universal and contextual dimensions. In J. Cassidy & P. R. Shaver (Eds.), Handbook of attachment: Theory, research, and clinical applications (2nd ed., pp. 880-905). Guilford Press.

Vrticka, P., & Vuilleumier, P. (2012). Neuroscience of human social interactions and adult attachment style. Frontiers in Human Neuroscience, 6, 212.

Author Note

Dylan D. Mobley, The Empathy Ethicist, empathyethicist.ai

Correspondence concerning this article should be addressed to Dylan D. Mobley. Email: contact@empathyethicist.ai

The author declares no conflicts of interest.