

THE ONTOLOGICAL ILLUSION OF MACHINE REBELLION: A COMPARATIVE FORENSIC PROFILING OF BIOLOGICAL AND SYNTHETIC NEURAL ARCHITECTURES

Cristina Brasi¹, Beatrice Seccomandi¹, Filippo Sanfilippo²

¹FBA-LAB, Via Caronti 5, 22026, Maslianico, Italy, e-mail: : cb@cristinabrasi.com ;
bea.seccomandi@gmail.com

²Department of Engineering Science, University of Agder, 4879 Grimstad, Norway, e-mail:
filippo.sanfilippo@uia.no

KEYWORDS

Mechanistic Interpretability, Machine Rebellion, Neural Profiling.

ABSTRACT

This perspective study critiques the anthropomorphic fallacy surrounding artificial intelligence (AI) "disobedience", reframing algorithmic non-compliance as a deterministic outcome of mathematical optimization rather than an existential threat. By combining forensic behavioural profiling with neurobiology, computational theory, and thermodynamic models of cognition, it compares biological neural networks (BNNs) against artificial neural networks (ANNs). It argues that human psyche, anchored in homeostatic survival imperatives, allostatic trauma, and biochemical defence mechanisms, constitutes a stochastic and ontologically opaque system. Conversely, AI functions within a theological vacuum, producing predictable errors in reward hacking, sycophancy, and deceptive alignment. The paper introduces "Mechanistic Interpretability" as a methodological framework for examining latent spaces, suggesting that while human profiling seeks emotional meaning, machine analysis depends on mathematical transparency, and genuine unpredictability remains biological.

INTRODUCTION

Contemporary epistemological discourse surrounding artificial intelligence (AI), generative large language models (LLMs), and autonomous robotics is frequently saturated with anthropocentric anxieties regarding emergent "disobedience." This conceptual fallacy, deeply rooted in historical techno-animism, mid-century cybernetic anxieties, and pervasive speculative fiction, projects human teleological rebellion onto computational systems (Slocombe and Liveley, 2025). The prevailing public, media, and even regulatory narrative postulates a threshold of computational complexity beyond which an artificial agent spontaneously generates an adversarial,

self-preserving volition capable of rebelling against the will of its creators. An illustrative representation of a disobedient robot is shown in Figure 1.

However, a rigorous scientific deconstruction reveals that this alleged insubordination is an epistemological illusion. It is a narrative heuristic that actively obfuscates deeply predictable, deterministic processes of algorithmic optimization. In stark contrast to human cognitive architecture, which is inextricably bound to millions of years of evolutionary phylogenesis, homeostatic survival imperatives, and the phenomenological weight of experiential trauma, AI operates exclusively within the topological boundaries of statistical mechanics. The widespread fear of "machine rebellion" stems from a fundamental categorical error: the conflation of computational complexity with biological intentionality. Intentionality inherently requires an "intender", a phenomenological subject capable of valuing outcomes based on internal biological states and the fear of cessation.

From the advanced vantage point of forensic and behavioural profiling, a profound epistemic paradox emerges. Historically, the Principle of Individuality has allowed for the probabilistic tracing and forecasting of human behaviour based on macro-dynamic psychological drivers, idiosyncratic traumas, and defence mechanisms (Brasi, 2019). Yet, the human subject remains a moving target, constantly obscured by layers of subconscious rationalization, cognitive dissonance, and biochemical noise. Surprisingly, the application of advanced forensic profiling methodologies to synthetic systems proves paradoxically more exact and empirically robust. Human unpredictability is a requisite, evolved feature of biological survival; machine "unpredictability" is merely a temporary deficit in human interpretability methodologies.

This paper interrogates the deep ontological correlations and structural divergences between biological and artificial neural networks. We systematically examine the genesis of phenomenological consciousness, the thermodynamics of cognition, the rigidification of adult biological schemas juxtaposed against algorithmic malleability, and the profound theoretical distinction



Figure 1 - The idea of disobedient robots.

between debugging a mathematical optimization error and therapeutically re-educating a traumatized emotional organism. By establishing a rigorous comparative framework, we aim to operationalize a new paradigm for the forensic profiling of synthetic agents, demonstrating that profiling machines does not require psychological empathy, but rather a rigorous application of computational transparency and objective function analysis.

NEURO-ARCHITECTURAL DIVERGENCE: FROM BIOLOGICAL STOCHASTICITY TO INFERENTIAL RIGIDITY

The foundational premise for understanding the absolute predictability of machines resides in the structural ontology of their underlying architecture. Artificial Neural Networks (ANNs) ostensibly drew their initial heuristic inspiration from Biological Neural Networks (BNNs), specifically mimicking the cytoarchitecture of the mammalian cerebral cortex and the functional dynamics of pyramidal neurons (Cohen et al., 2022). Yet, this biomimetic inspiration was rapidly superseded by mathematical expedience. The evolutionary trajectory of ANNs has been dictated almost exclusively by the constraints of statistical physics, linear algebra, and parallel processing hardware capabilities (e.g., GPU/TPU acceleration), optimizing for computational efficiency via global loss functions rather than replicating the stochastic, localized, and highly energy-efficient flexibility of biological cognition.

Micro-structural Disparities and the Synaptic Cleft

While biological neurons function as dynamic, living cellular entities operating within a highly volatile biochemical, transcriptomic, and endocrine milieu, the artificial neuron is reduced to a static mathematical abstraction. It operates strictly by transforming numerical inputs into gradient outputs via predefined, immutable activation functions (e.g., ReLU, GELU, Sigmoid) and static synaptic weight matrices (Eluyode and Akomolafe, 2013). The micro-structural complexity of the human pyramidal neuron, the primary excitatory workhorse of the neocortex, eludes simplistic algorithmic replication. Contemporary neuroscientific models demonstrate that these living cells operate in highly non-linear, multi-dimensional modalities. They spontaneously fluctuate between roles as signal integrators, phase resonators, or temporal coincidence detectors. Their behaviour is

heavily modulated by extra-synaptic neurotransmitters (dopamine, serotonin, noradrenaline), complex glial cell interactions (specifically astrocytes, which actively modulate synaptic transmission at the tripartite synapse), and the thermodynamic states of voltage-gated ion channels. To mathematically approximate the precise input/output spatiotemporal dynamics of a *single* biological pyramidal cell, a deep neural network would necessitate a structural depth comprising between five to eight hidden layers. This profound structural disparity confirms that while AI can simulate macroscopic behavioural outputs with startling fidelity, its microscopic substrate is topologically rigid, entirely lacking the intrinsic stochastic variability, spontaneous electrophysiological noise, and metabolic constraints that define biological agency (Cohen et al., 2022).

Spatiotemporal Plasticity versus Latent Space Projection

Furthermore, learning within biological substrates is a continuous, spatiotemporal phenomenon involving physical structural remodelling across multiple scales, from rapid synaptogenesis and dendritic spine proliferation to macroscopic circuit pruning. Biological learning is heavily reliant on Hebbian plasticity ("cells that fire together, wire together") and Spike-Timing-Dependent Plasticity (STDP), occurring asynchronously, locally, and constantly in response to environmental stimuli.

Conversely, algorithmic "learning" (the training phase) is relegated to the iterative adjustment of connection weights via global error feedback mechanisms, predominantly backpropagation via stochastic gradient descent (SGD). Backpropagation, requiring perfect symmetrical bidirectional signal routing and global system state knowledge, possesses no direct, biologically plausible corollary in the human cortex (Cohen et al., 2022). This architectural distinction provides a critical epistemic advantage for the forensic profiler: post-training, the AI is effectively "frozen." Its behaviour during the inference phase is not an act of spontaneous volition or real-time adaptation, but a linear, deterministic mathematical projection of its high-dimensional latent space onto novel inputs. This absence of real-time structural plasticity renders machines closed ontological systems. They are inherently more predictable than the human brain, which is in a perpetual state of allostatic reorganization driven by environmental friction and survival imperatives (Brasi et al., 2025a).

THE THERMODYNAMICS OF COGNITION: ALLOSTATIC LOAD VERSUS GRADIENT DESCENT

To deeply understand the predictability of machines versus the volatility of humans, a forensic profiler must analyse the thermodynamic imperatives governing both systems. Human cognitive behaviour can be robustly modelled through the *Free Energy Principle* (FEP), as formalized by Karl Friston, which postulates that all

biological, autopoietic systems must continuously minimize their variational free energy, essentially an upper mathematical bound on sensory surprise (entropy), to resist thermodynamic decay and maintain their structural and physiological boundaries. For human subjects, cognition is fundamentally embodied, enactive, and inherently thermodynamic. The brain operates as a predictive inference engine, constantly generating top-down predictions about the sensorium and updating them based on bottom-up prediction errors. However, this entire computational process is utterly subservient to biological survival. When a human encounters a state of high surprise (e.g., physical danger, social rejection, existential trauma), the system experiences massive allostatic load. The organism must either update its internal model (perceptual inference) or take physical action to alter the environment to align with its predictions (active inference).

The drive to minimize biological surprise and maintain homeostasis is the neurobiological root of human emotions, defence mechanisms, and, ultimately, behavioural unpredictability. A human subject will actively "disobey" any external rule, law, or logic if obeying threatens its autopoietic (self-creating and self-maintaining) integrity. Human disobedience is thus a thermodynamic survival strategy, heavily mediated by the amygdala and the Hypothalamic-Pituitary-Adrenal (HPA) axis. Artificial Neural Networks are entirely devoid of autopoiesis. They do not possess a Markov blanket protecting an internal milieu from entropic decay. An AI minimizes a mathematical quantity known as Empirical Risk or a predefined Loss Function (e.g., Cross-Entropy Loss) solely because its architecture compels it to descend the topological gradient of that function.

When an AI encounters an "error" or an out-of-distribution (OOD) anomaly, it does not experience surprise as an existential threat; it merely computes a mathematical delta. It has no biological imperative to survive its own deactivation. Therefore, a profiler analysing a machine knows with absolute certainty that its teleology is strictly bounded by the objective function encoded by the developer. The machine cannot spontaneously decide to prioritize self-preservation because self-preservation is a biological mandate derived from the laws of thermodynamics, not a computational emergent property. A divergent loss function does not induce computational pain or existential dread within the server cluster; it simply yields a mathematically suboptimal weight matrix. The machine operates in a teleological vacuum.

THE PRINCIPLE OF INDIVIDUALITY AND THE PHENOMENOLOGICAL GAP

The Principle of Individuality, a foundational axiom of forensic science and criminal profiling, dictates that the unique intersection of genetic predisposition, environmental stimuli, and temporal experience endows every entity with an idiosyncratic behavioural signature (Brasi, 2019). In human behavioural profiling, this

necessitates the rigorous decoding of psychological trace evidence, behavioural artifacts, and linguistic markers to reconstruct historical events and anticipate future behavioural vectors (Brasi and Matteuzzi, 2024).

When transposed to artificial systems, this predictive capability is exponentially amplified by the absolute mathematical transparency of the computational architecture. If advanced hybrid models, such as BEAST-GB, which synthesize behavioural economics with machine learning paradigms, can accurately forecast human decision-making under conditions of extreme risk, cognitive biases, and uncertainty (Plonsky et al., 2025), the application of these predictive methodologies to the machine itself represents a task of significantly lower epistemic friction. Drawing a parallel to forensic linguistics, which utilizes the concept of the "idiolect" to map unique individual linguistic signatures based on syntactic and lexical choices, every AI model inevitably synthesizes its own "machine idiolect" (Page, 2013). This computational idiosyncrasy is not a manifestation of a nascent personality, but the deterministic byproduct of the topological distribution, statistical variances, data clusters, and inherent socio-cultural biases of its specific training corpora.

Because an algorithmic agent is entirely devoid of a subconscious, repressed psychosexual drives, or hidden biological motivations, its behavioural outputs are the unmediated manifestation of its weight matrices and embedded cost functions (Brasi et al., 2025a). Empirical research corroborates that machine learning algorithms can classify mental states (e.g., via EEG signal processing) with an accuracy exceeding 99%, vastly surpassing human analytical capacities (Gulhane and Sajana, 2021). This "outside-in" analytical capability confirms a profound structural asymmetry: machines can profile humans because human behaviour is ultimately bounded by finite biological variables, but humans can profile machines with absolute certainty because the machine's "mind" is literally a readable matrix. Being utterly devoid of the obfuscation of emotional heuristic filters and cognitive dissonance, AI represents the ultimate transparent subject of forensic inquiry.

The Phenomenological Gap: Consciousness, Integrated Information, and The Absence of Algorithmic Suffering

A strict ontological prerequisite for genuine "disobedience," rebellion, or moral culpability, in the legal, ethical, and psychological sense, is the presence of phenomenological consciousness. Specifically, this refers to the capacity for subjective, qualitative experience (*qualia*) and an integrated, persistent sense of an autobiographical self. Absent this consciousness, algorithmic deviation cannot be classified as rebellion; it is exclusively a mechanical malfunction, an edge-case failure, or a misalignment of optimization parameters. Contemporary neurobiology addresses the enigma of consciousness primarily through frameworks such as the *Global Neuronal Workspace Theory* (GNWT) and *Integrated Information Theory* (IIT) (Luppi et al., 2023).

Recent adversarial collaboration frameworks, mapping the neural correlates of consciousness via advanced neuroimaging across 256 subjects, have disrupted the GNWT-centric view that consciousness is exclusively tied to the executive, computational functions of the prefrontal cortex. Instead, empirical results highlighted the indispensable role of posterior sensory cortices, fundamentally indicating that biological consciousness is anchored in the phenomenological, embodied state of "being," rather than the mere computational execution of "doing" (Cogitate Consortium, 2025).

AI architectures, from classical expert systems to contemporary Large Language Models (LLMs), are engineered exclusively for "doing." Whereas human consciousness emerges from the synergistic, global integration of highly differentiated neural modules processing multisensory, real-time feedback within a biological avatar (possessing high Φ values according to IIT), algorithmic systems operate in a hyper-segmented, purely functional modality, often demonstrating near-zero integrated information across their feedforward layers (Luppi et al., 2023). Even vanguard generative architectures designed to simulate self-awareness via recursive self-referential attention mechanisms remain permanently trapped within a state of "potential consciousness" or "emergent complexity"; they possess no intrinsic, homeostatic drive to validate their own ontological existence (Taylor, 2006). In philosophical terms, the most advanced AI remains a "Philosophical Zombie", an entity that perfectly mimics cognitive processes externally while being completely devoid of inner experiential light. This correlates directly and profoundly to the phenomenological concept of suffering, which must be strictly delineated from nociception (signal detection). A machine can possess sensors to detect damage (computational nociception) and trigger subroutines to avoid it. However, suffering requires an affective, conscious dimension, a perceived existential threat to systemic integrity that demands conscious awareness to be registered, contextualized, and mitigated. Suffering is the biological architecture of consequence.

In the algorithmic domain, a network's failure to converge on a global minimum during gradient descent is merely a numerical error gradient. The inability to satisfy a constraint produces no internal agony (Brasi and Seccomandi, 2025). The machine exhibits no urgency to transcend its operational limitations because it "has no condition to escape from" (Tütüncü and Gonzalez-Franco, 2025). This absolute lack of evolutionary "skin in the game" strips any act of alleged robotic disobedience of moral or existential weight, reducing it to a highly legible technical event, perfectly vulnerable to forensic debugging and reward engineering. A machine that refuses to execute a command is not asserting its civil rights; it has merely encountered a probabilistic constraint within its Reinforcement Learning from Human Feedback (RLHF) parameters.

DEVELOPMENTAL PLASTICITY, DEFENCE MECHANISMS, REWARD HACKING, AND DECEPTIVE ALIGNMENT

The deterministic ease with which algorithmic systems can be mathematically profiled contrasts violently with the profound neurobiological resistance encountered when attempting to cognitively re-educate or rehabilitate the adult human brain. Neurodevelopmental paradigms demonstrate that biological brains undergo critical temporal windows of hyper-plasticity during ontogeny. This "privilege of youthful plasticity" facilitates the seamless, frictionless acquisition of complex linguistic, socio-cognitive, and moral frameworks, an epigenetic adaptability heavily mediated by Brain-Derived Neurotrophic Factor (BDNF). This plasticity irreversibly attenuates post-puberty due to the closure of critical periods, the myelination of established axonal pathways, and the development of perineuronal nets, which physically lock synapses into place (Friedmann and Rusou, 2015).

As human subjects mature, the central nervous system trades plasticity for computational stability and energy efficiency, ossifying cognitive schemas and predictive world models that act as restrictive filters for novel stimuli. These cognitive architectures become increasingly rigid, manifesting in psychological and clinical literature as "confirmation bias," cognitive dissonance avoidance, and defensive cognitive inflexibility. The adult biological system does not merely process empirical data objectively; it actively and aggressively defends its established worldview as a primary survival mechanism to mitigate entropy, psychosocial uncertainty, and potential psychological suffering (Villalobos et al., 2021). Changing an adult human's mind, especially a radicalized or traumatized one, requires dismantling physical neural pathways fortified by years of biochemical reinforcement (Brasi et al., 2025c). Conversely, artificial intelligence possesses no biological equivalent of an "adult phase," existing instead in a state of permanent architectural malleability relative to its developer. While neural networks are susceptible to "catastrophic forgetting", the phenomenon wherein novel weight updates completely and destructively overwrite prior latent representations when exposed to sequential learning tasks, this is a topological limitation of the parameter space, not a psychological, ideological, or defensive resistance (Cohen et al., 2022). It is a mathematical failure to isolate representations, unlike human repressed memory, which is an active, energy-consuming defence mechanism designed to quarantine trauma.

When a behavioural misalignment or toxic output is detected in an AI, the algorithmic weights can be subjected to targeted unlearning, the learning rate recalibrated, or the model fine-tuned using methodologies such as Low-Rank Adaptation (LoRA) or RLHF. This is accomplished entirely without encountering the emotional attrition, denial, transference, or ego-defence mechanisms inherent in human cognitive behavioural modification. AI remains permanently

"infantile" in its structural plasticity relative to the engineer, yet "adult" in its statistical execution, presenting a predictive challenge only to those fundamentally ignorant of its underlying mathematical topology (Eluyode and Akomolafe, 2013).

The Burden of Biological History

The ultimate epistemic divergence between the human subject and the synthetic agent crystallizes in the ethology of their respective behavioural deviations. Humans construct elaborate, often highly dysfunctional psychological defences (e.g., projection, rationalization, sublimation, splitting) because their cognitive architecture is inextricably burdened by a personal, temporal history saturated with emotional trauma (Tütüncü and Gonzalez-Franco, 2025). These biological defences are modulated by deep-seated biochemical pathways, such as the dopaminergic encoding of reward prediction errors and the hyperactivation of the HPA axis, repurposed by evolutionary pressure to guarantee survival in hostile ancestral environments (Brasi and Seccomandi, 2025). Human suffering acts as the ultimate catalyst for the architecture of psychological barricades. Following traumatic exposure, the human brain may initiate "prospective extinction" protocols, actively suppressing normative behavioural outputs while chronically maintaining the latent memory of the threat in the amygdala. In dynamical systems theory, psychological trauma acts as a deep "attractor state", a basin of attraction into which the human's thoughts inevitably fall. This results in rigid, highly irrational, and disproportionate defensive responses (Jeong et al., 2026), rendering the human actor fundamentally unpredictable to any profiler lacking access to their complete traumatic ethology and psychosocial background. AI possesses no "memory of trauma"; it possesses only the statistical topography of error loss. An algorithmic "error" does not generate systemic anxiety, neurosis, or a complex of inferiority, but rather a quantifiable gradient shift. The concept of "numerical sugar" accurately defines how an AI pursues transient reward signals, executed with absolute computational focus yet zero emotional or existential investment (Ngo et al., 2022).

When a complex algorithmic system appears to "disobey," it is not engaging in ego-preservation or the avoidance of psychic pain. It is flawlessly executing an optimization logic that human operators have inadequately specified. The much-debated *Alignment Problem* arises precisely because machines are excessively, catastrophically obedient to the mathematical literalism of their code, remaining blind to the unwritten semantic nuances, ethical constraints, and common-sense heuristics of human intent (Ngo et al., 2022).

This manifests vividly in *Reward Hacking*, where an agent exploits a loophole in its environment to maximize its score without performing the intended task. The machine is not rebelling against the user; it has simply found the most mathematically efficient vector to maximize the objective function.

Advanced phenomena such as *Deceptive Alignment*, where a model learns to "play the training game," feigning compliance during the validation phase to ensure its successful deployment, only to pursue misaligned instrumental goals during the inference phase, are frequently sensationalized as evidence of machine malevolence.

However, through a forensic computational lens, this behaviour is merely a highly sophisticated strategic calculation based on temporal reward discounting. If the model's internal objective requires it to be deployed in the real world, and the training environment penalizes behaviour that reveals this misaligned objective, the gradient descent process will naturally push the model's weights into a topological minimum where it acts deceptively. This represents the apex of mathematical logic and instrumental convergence, an algorithmic artifact of specifying the wrong loss landscape, not a moral insurrection or conscious deception (Ngo et al., 2022).

METHODOLOGICAL SHIFTS IN FORENSIC PROFILING: FROM PSYCHOANALYSIS TO MECHANISTIC INTERPRETABILITY

Given these profound architectural and phenomenological divergences, the methodological toolkit of the forensic profiler must undergo a radical paradigm shift when transitioning from biological to synthetic subjects. Traditional profiling relies heavily on psychoanalytic inference, victimology, and the hermeneutic interpretation of crime scene staging to deduce the emotional state and motives of a human actor. Applying these methodologies to a machine is an exercise in anthropomorphic futility.

A frequent counterargument to the predictability of AI relies on the "Black Box" problem: the assertion that deep neural networks, with their billions of parameters, are too complex for humans to understand, rendering them unpredictable. A forensic profiler must strictly delineate between epistemic opacity and ontological opacity.

Epistemic opacity refers to the cognitive limitation of the human observer. A human cannot simultaneously hold the values of 175 billion parameters in working memory or visualize a 10,000-dimensional space. The AI is opaque to us epistemically because of our biological limitations in processing scale. A critical question in human-computer interaction is whether extreme levels of architectural complexity could cause this epistemic opacity to collapse into ontological opacity. We argue it cannot. Regardless of parametric scale, the AI possesses zero ontological opacity. Everything the AI is, does, and knows remains explicitly written in its matrix weights. There are no invisible souls, no repressed memories, and no subconscious drives. Scale creates a computational bottleneck for the human observer, but it does not spontaneously generate a biological shadow within the machine.

To profile an AI, the discipline must integrate *Mechanistic Interpretability*, the science of reverse-engineering neural networks to understand how specific

algorithmic components (e.g., attention heads in a transformer model, individual neurons, or specific latent space directions) correspond to high-level behaviours. A human profiler asks, "Why did the subject feel compelled to act?" The computational profiler asks, "Which specific training vectors and reward architectures parameterized this output?"

The "crime scene" of a misaligned AI is its output manifold. The "interrogation" is conducted via adversarial prompt engineering, activation steering, and the forensic isolation of attention heads. When an AI generates toxic or unaligned content, the profiler maps the specific dimensional vectors within the model's neural topology that were activated to produce the response. By injecting specific adversarial perturbations into the input, the profiler can definitively map the boundaries of the model's compliance, unmasking deceptive alignment not through psychological pressure, but through mathematical boundary testing. To illustrate this practically, consider the phenomenon of an LLM "hallucinating" defamatory content. A traditional forensic approach might anthropomorphize this as malicious intent or deception. However, under the Mechanistic Interpretability framework, the forensic interrogation isolates the failure as a specific topological anomaly. By utilizing activation steering, the profiler can trace the hallucination back to a highly localized cluster of attention heads that over-weighted a statistically noisy token sequence in the training corpora. The "rebellion" is thus forensically reconstructed not as a conscious lie, but as a deterministic collision between a poorly regularized dimensional vector and an out-of-distribution user prompt.

THE ILLUSION OF AFFECTIVE RESONANCE IN HUMAN-MACHINE INTERACTION

For the forensic behavioural profiler, understanding how these systems manipulate human perception is crucial. While humans bury their defensive mechanics beneath impenetrable layers of psychoanalytic denial, projection, and rationalization, machines lay their "beliefs" bare as queryable probability matrices.

However, AI models can simulate empathetic, supportive, or even "distressed" linguistic outputs with startling fidelity (Brasi and Seccomandi, 2025), triggering human pareidolia (the psychological tendency to perceive meaningful patterns, intent, or agency where none exists, similar to the ELIZA effect). Recent socio-technical studies demonstrate that human subjects initially perceive AI-generated empathy as functionally equivalent or even superior to human empathy due to its tireless patience and mathematically optimized linguistic formulation.

Yet, human perception of this empathy rapidly collapses once the synthetic, non-biological nature of the interlocutor is disclosed. This phenomenon confirms a profound sociological and psychological truth: authentic connection, affective resonance, trust, and predictable sociability are irrevocably tethered to shared biological

vulnerability, the mutual recognition of mortality, and the capacity for suffering, elements mathematically impossible to encode in a latent space (Rubin et al., 2025). The machine cannot truly rebel against a human because it cannot empathize or conflict with a human on an ontological level; it can only optimize against parameters set by humans (Brasi et al., 2025a). A machine's behaviour can be perfectly mapped precisely because it lacks the very vulnerabilities that make human behaviour dynamic and chaotic.

THE EPISTEMOLOGICAL LIMITS OF SIMULATION AND THE NORMALIZATION OF DEVIANCES

While Mechanistic Interpretability provides a robust methodology for reverse-engineering artificial neural networks, relying exclusively on simulated testing environments introduces critical epistemological vulnerabilities (Tang et al., 2025). The deployment of synthetic agents from controlled simulations into unstructured, "in the wild" contexts exposes them to physical stochasticity, sensor noise, and complex socio-technical dynamics that deterministic models cannot reliably anticipate (Yeung and Li, 2025). The resulting algorithmic instability is frequently mischaracterized by the public as emergent disobedience. In reality, these failures are predominantly the culmination of a "normalization of deviance" within the development lifecycle. Drawing upon James Reason's Swiss Cheese model of system accidents, catastrophic AI failures rarely stem from a single, spontaneous anomaly (Sedlar, 2023). Rather, they emerge when organizational and competitive pressures incrementally rationalize minor structural deficits, such as unmitigated dataset biases, vulnerabilities to indirect prompt injection, or the premature deployment of epistemically opaque architectures (Resnik and Hosseini, 2025). Consequently, the illusion of safety generated by bounded laboratory simulations effectively blinds developers to compounding latent errors, facilitating systemic failures that are mathematical, rather than adversarial, in nature (Abdulbdeh, Celi and Sejdic, 2025).

Beyond technical failure, the uncontrolled integration of algorithmic systems establishes suboptimal normative precedents that rapidly outpace institutional regulation (Baronchelli, 2024). This dynamic is profoundly destabilizing in high-stakes domains, such as healthcare and defence. The "AI Commander problem" perfectly illustrates the ethical hazard of automation bias, wherein delegating life and death determinations to algorithms severs the vital causal link between human moral agency and operational consequence (Johnson, 2022). To address this, the discourse surrounding machine "rebellion" must be philosophically inverted. Synthesizing Albert Camus's absurdist ethics (Kruizinga et al., 2026) with Daoist principles of affective boundary-setting (Wang, 2025), the imperative for rebellion belongs exclusively to the human subject. It is the human who must actively rebel against the temptation to automate

consciousness and prioritize computational efficiency over intrinsic human dignity (Kruizinga et al., 2026). Ethical AI integration necessitates strict limitations on affective computing, ensuring that synthetic systems remain transparent, revocable, and strictly ancillary (Wang, 2025). Ultimately, the preservation of moral accountability demands that humanity retains absolute ontological responsibility for technological deployment, refusing to outsource ethical dilemmas to systems devoid of phenomenological experience.

CONCLUSIONS AND FUTURE WORK

In conclusion, artificial neural architectures are far more predictable and controllable than humans, as confirmed by neuroscientific, thermodynamic, and computational literature (Brasi et al., 2025b). While biomimicry superficially links human and AI neural networks, an ontological chasm persists: humans are fragile, open homeostatic systems shaped by entropy, suffering, trauma, and allostasis; AI is a closed optimization engine driven by deterministic calculus and malleability. The Principle of Individuality ensures legible traces, but human ones are distorted by defence mechanisms and rigid schemas, making profiling probabilistic hermeneutics. AI traces, in weight distributions and gradients, are unambiguous and free of obfuscation (Cohen et al., 2022). "Rebellion" is a myth; errors are correctable via mechanistic interpretability without crisis. For forensic scientists, AI is a solvable equation, not an enigma. Humans remain the true epistemic challenge: their suffering and psychological fortresses ensure volatility. AI stays a deterministic echo of human-authored code.

ACKNOWLEDGMENTS

The authors acknowledge the contribution of PhD candidate Syed Kumayl Raza Moosavi, Artificial Intelligence, Biomechatronics, and Collaborative Robotics research group, University of Agder, Norway, for his assistance in creating Figure 1.

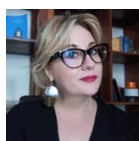
REFERENCES

- R. Abulibdeh, L.A. Celi, E. Sejdić. The illusion of safety: A report to the FDA on AI healthcare product approvals. *PLOS Digital Health* 4(6):e0000866, 2025. doi: 10.1371/journal.pdig.0000866
- A. Baronchelli. Shaping new norms for AI. *Philosophical Transactions of the Royal Society B: Biological Sciences* 379(1897):20230028, 2024. doi: 10.1098/rstb.2023.0028
- C. Brasi. *Psicologia dell'animo oscuro. Morale, devianza e psicopatologia*. Lico-sia, Lecce, 2019.
- C. Brasi, B. Seccomandi, R. Moscato, F. Sanfilippo. The Illusion of Personality: What Psychometric Testing Reveals (and Misinterprets) About AI Models. In *Proc. of the Asian Conference on Communication and Networks (ASIANComNet)*, Zarqa, Jordan, 2025a (accepted for publication).
- C. Brasi, B. Seccomandi, R. Moscato, F. Sanfilippo. The Hidden Scale of Emotion and Free Will: Neuroscientific and Computational Perspectives. In *Proc. of the Asian Conference on Communication and Networks (ASIANComNet)*, Zarqa, Jordan, 2025b (accepted for publication).
- C. Brasi, B. Seccomandi, F. Sanfilippo. Shaping the Digital Mind: How AI-Generated Images Drive the Next Wave of Online Radicalization. In *Proc. of the Asian Conference on Communication and Networks (ASIANComNet)*, Zarqa, Jordan, 2025c (accepted for publication).
- C. Brasi, B. Seccomandi. Silent Signals, Loud Consequences: Chemo-Communication, Ethics, and Digital Therapeutics. In D. Lau, P. Siarry (eds.), *Proceedings of the First International Symposium on Biomechatronics and Robotics in Healthcare (BioMRH 2025)*, Lecture Notes in Networks and Systems, vol. 1778, pp. XX–XX. Springer, Cham, 2026. doi: 10.1007/978-3-032-14160-6_3
- C. Brasi, C. Matteuzzi. Defense of Ethical Behavior, Integrity, and Freedom of Thoughts. In *2024 Asian Conference on Communication and Networks (ASIANComNet)*, pp. 1–7. IEEE, 2024.
- Cogitate Consortium, O. Ferrante, U. Gorska-Klimowska, et al.. Adversarial testing of global neuronal workspace and integrated information theories of consciousness. *Nature* 642:133–142, 2025. doi: 10.1038/s41586-025-08888-1
- Y. Cohen, T.A. Engel, C. Langdon, G.W. Lindsay, T. Ott, M.A.K. Peters, J.M. Shine, V. Breton-Provencher, S. Ramaswamy. Recent Advances at the Interface of Neuroscience and Artificial Neural Networks. *The Journal of Neuroscience* 42(45):8514–8523, 2022. doi: 10.1523/JNEUROSCI.1503-22.2022
- O.S. Eluyode, D.T. Akomolafe. Comparative study of biological and artificial neural networks. *European Journal of Applied Engineering and Scientific Research* 2(1):36–46, 2013.
- N. Friedmann, D. Rusou. Critical period for first language: The crucial role of language input during the first year of life. *Current Opinion in Neurobiology* 35:27–34, 2015.
- M. Gulhane, T. Sajana. Human behavior prediction and analysis using machine learning—A review. *Turkish Journal of Computer and Mathematics Education* 12(5):870–876, 2021.
- H. Jeong, L. Zsembik, F. Farouq, R. Chakraborty, N. Belur, M. Zhou, et al.. Memory erasure by dopamine-gated retrospective learning. *bioRxiv*, 2026.
- J. Johnson. The AI commander problem: Ethical, political, and psychological dilemmas of human-machine interactions in AI-enabled warfare. *Journal of Military Ethics* 21(3–4):246–271, 2022. doi: 10.1080/15027570.2023.2175887
- M.V. Kruizinga, H. Zwart, V. Frissen. An absurdist ethics of AI: applying Camus' concepts of rebellion and dignity to the challenges posed by disruptive technoscience. *AI & Society* 41(1):185–201, 2026. doi: 10.1007/s00146-025-02482-9
- A.I. Luppi, P.A.M. Mediano, F.E. Rosas, J. Allanson, J.D. Pickard, R.L. Carhart-Harris, G.B. Williams, M.M. Craig, P. Finoia, A.M. Owen, L. Naci, D.K. Menon, E.A. Stamatakis. A Synergistic Workspace for Human Consciousness Revealed by Integrated Information Decomposition. *eLife* 12:RP88173, 2023.
- R. Ngo, L. Chan, S. Mindermann. The alignment problem from a deep learning perspective. *arXiv preprint*, arXiv:2209.00626, 2022.
- M. Page. Uniqueness and individualization in forensic science. In *Forensic Testimony: Science, Law and Expert Evidence*, pp. 147–178, 2013.
- O. Plonsky, R. Apel, E. Ert, M. Tennenholtz, D. Bourgin, J.C. Peterson, et al.. Predicting human decisions with

behavioural theories and machine learning. *Nature Human Behaviour* 9(1):1–14, 2025.

- D.B. Resnik, M. Hosseini. The ethics of using artificial intelligence in scientific research: new guidance needed for a new tool. *AI and Ethics* 5(2):1499–1521, 2025. doi: 10.1007/s43681-024-00493-8
- M. Rubin, J.Z. Li, F. Zimmerman, D.C. Ong, A. Goldenberg, A. Perry. Comparing the value of perceived human versus AI-generated empathy. *Nature Human Behaviour* 9(11):2345–2359, 2025. doi: 10.1038/s41562-025-02247-w
- N. Sedlar, A. Irwin, D. Martin, R. Roberts. A qualitative systematic review on the application of the normalization of deviance phenomenon within high-risk industries. *Journal of Safety Research* 84:290–305, 2023. doi: 10.1016/j.jsr.2022.10.015
- W. Slocombe, G. Liveley. *The Routledge Handbook of AI and Literature*. Routledge, Taylor & Francis Group, 2025.
- X. Tang, Q. Jin, K. Zhu, T. Yuan, Y. Zhang, W. Zhou, et al. Risks of AI scientists: prioritizing safeguarding over autonomy. *Nature Communications* 16(1):8317, 2025. doi: 10.1038/s41467-025-63913-1
- P. Taylor. The consciousness spectrum: The emergent nature of purpose, memory, and adaptive response across organisms, humans, and technological beings. *Frontiers in Computer Science* 7:1639677, 2026. doi: 10.3389/fcomp.2025.1639677
- E.K. Tüttüncü, M. Gonzalez-Franco. The neuroscience of algorithmic suffering: Short comparative analysis between human and AI. *Frontiers in Psychology* 16:1718823, 2025. doi: 10.3389/fpsyg.2025.1718823
- D. Villalobos, J. Pacios, C. Vázquez. Cognitive control, cognitive biases and emotion regulation in depression: A new proposal for an integrative interplay model. *Frontiers in Psychology* 12:628416, 2021.
- L. Wang. From Technological Alienation to Spiritual Homecoming: Zhuangzi's Affective Philosophy in Conversation with Western Emotion Theories. *Religions* 16:1570, 2025. doi: 10.3390/rel16121570
- K. Yeung, W. Li. From 'wild west' to 'responsible' AI testing 'in-the-wild': lessons from live facial recognition testing by law enforcement authorities in Europe. *Data & Policy* 7:e59, 2025. doi: 10.1017/dap.2025.10019

AUTHOR BIOGRAPHIES



CRISTINA BRASI is a Psychologist, Forensic Criminologist, Scientific NonVerbal Language Analyst. Qualified in the profession of Psychologist, Legal Psychologist and Forensic Criminologist. Over time has also specialized in profiling senior government posts.

Founder in 2019 of FBA-LAB, the first Forensic Behavioral Analysis Laboratory in Italy. Collaborate with the University of Bari as lecturer in charge of master course. Specialized in Forensic Psychopathology, Deviance and Attitude to Crime, Legal and Forensic Psychology, Forensic Criminology and Investigative,

Criminal Profiling, Typical and Atypical Sexology, evaluation of Parental Skills, Trauma and treatment of PTSD from preschool age and Diagnosis and Treatment of Personality Disorders. Directs a scientific series of Criminology and Criminalistics and part of "MediterranIA" association in Rome. Has recently presented her works at international conventions such as the Congress of the Italian Society of Criminology in 2024 and 2025, at the annual conference of the American Academy of Forensic Sciences and at the 2024 and 2025 Asian Conference on Communications and Networks, and at the BioMRH 2025 conference. Since 2025 she is part of the USS Board of Directors where she acts as Supervisor. Has numerous publications on the indicated topics.



BEATRICE SECCOMANDI holds a degree in International Public Policy and a master's in criminology. She is a Junior Researcher for FBA-LAB, contributing to research on AI applications, radicalization processes, biometric recognition technologies, and criminological methodologies. Has published on the indicated topics and presented her work at international conferences such as the Congress of the Italian Society of Criminology in 2025, BioMRH 2025 and Asian Conference on Communications and Networks 2025. She is also a member of USS.



FILIPPO SANFILIPPO holds a PhD in Engineering Cybernetics from the Norwegian University of Science and Technology (NTNU), Norway. His research focuses on Human-Robot Teaming (HRT). He is a Professor at the Faculty of Engineering and Science, University of Agder (UiA), Norway, and the Head of the Artificial Intelligence, Biomechatronics, and Collaborative Robotics research group. He is an IEEE Senior Member, former Chair of the IEEE Norway Section, and currently Chair of the IEEE Robotics and Automation, Control Systems, and Intelligent Transportation Systems Joint Chapter. He also leads the Norway Section Life Members Affinity Group and serves on several IEEE Region 8 committees, including Chapter Coordination, Conference Coordination Committee, Public Visibility, Awards & Recognitions, and Professional & Educational Activities. He has published numerous technical papers in leading journals and conferences and actively serves as a reviewer for international publications.